Meta
Journal des traducteurs
Translators' Journal

META

L’apport de l’informatique à la recherche lexicographique

Roda P. Roberts et Lucie Langlois

Volume 46, numéro 4, décembre 2001

URI : https://id.erudit.org/iderudit/003956ar
DOI : https://doi.org/10.7202/003956ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Roberts, R. P. & Langlois, L. (2001). L’apport de l’informatique à la recherche lexicographique. *Meta*, *46*(4), 711–720. https://doi.org/10.7202/003956ar

Résumé de l'article

L'informatique joue un rôle central dans toutes les étapes de la production dictionnairique : la consultation et l'analyse de la documentation, la préparation des entrées et leur révision. Dans cet article, nous montrons le rôle de l'informatique dans la création du Dictionnaire canadien bilingue, qui constitue un des grands objectifs d'un projet de recherche interuniversitaire canadien.

ÉTUDES TERMINOLOGIQUES ET LINGUISTIQUES

L'apport de l'informatique à la recherche lexicographique

RODA P. ROBERTS ET LUCIE LANGLOIS
Université d'Ottawa, Ottawa, Canada

RÉSUMÉ

L'informatique joue un rôle central dans toutes les étapes de la production dictionnairique : la consultation et l'analyse de la documentation, la préparation des entrées et leur révision. Dans cet article, nous montrons le rôle de l'informatique dans la création du Dictionnaire canadien bilingue, qui constitue un des grands objectifs d'un projet de recherche interuniversitaire canadien.

ABSTRACT

Computer technology plays a central role in all stages of dictionary production: consultation and analysis of documentation, entry preparation and entry revision. In this article, we present the contribution of computer technology to the creation of the Bilingual Canadian Dictionary, which is one of the primary objectives of an interuniversity Canadian research project.

MOTS-CLÉS/KEYWORDS

analyse de la documentation, dictionnairique, Dictionnaire canadien bilingue, informatique, lexicographie

Introduction

Le projet intitulé officiellement « Lexicographie comparée du français et de l'anglais au Canada » et officieusement « Dictionnaire canadien bilingue » (DCB) est un projet interuniversitaire de grande envergure. Une soixantaine de chercheurs (professeurs et assistants) répartis entre trois universités (Université d'Ottawa, Université de Montréal et Université Laval) travaillent ensemble dans le but de faire avancer les recherches en lexicographie, et plus particulièrement en lexicographie bilingue.

Ce projet, lancé en 1988 et subventionné par le Conseil de recherches en sciences humaines depuis 1994, vise quatre grands objectifs :

- produire un dictionnaire canadien bilingue qui reflète le français et l'anglais tels qu'ils sont utilisés au Canada ;
- établir une base de données textuelles contenant surtout des textes canadiens rédigés en anglais et en français ;
- établir une base de données dictionnairiques pour la préparation, le stockage et la gestion des entrées, ainsi que la diffusion éventuelle du dictionnaire ; et
- développer le savoir canadien en lexicographie bilingue.

Si les objectifs 2 et 3 ont un lien direct et évident avec l'informatique, il n'en demeure pas moins vrai que les objectifs 1 et 4 seraient, eux, difficilement réalisables sans l'apport de l'informatique.

En effet, les méthodes de création des dictionnaires ainsi que les approches théoriques en lexicographie sont grandement influencées par les outils informatiques qui sont désormais à la disposition des lexicographes. À tel point qu'on a même créé une nouvelle terminologie pour désigner cette réalité : *computational lexicography*, *computerized lexicography*, *machine lexicography* et *lexicographic computing* en anglais ; *lexicographie computationnelle*, *lexicographie automatique*, et plus récemment *dictionnairique* en français[1]. Même si tous ces termes sont sujets à des interprétations quelque peu différentes, leur existence même prouve qu'on distingue aujourd'hui deux types de lexicographie, selon que l'informatique y joue un rôle central ou non. Le mot « central » est important, car il y a déjà quelque temps que l'informatique sert à stocker des entrées et à préparer une version finale prête pour la photogravure. Mais dans un projet de lexicographie computationnelle comme le DCB, l'informatique contribue non seulement aux tâches périphériques mais aussi aux grandes étapes de la production dictionnairique.

La production dictionnairique comprend trois grandes étapes : la consultation et l'analyse de la documentation, la préparation des entrées et leur révision. Nous allons montrer ici le rôle de l'informatique dans toutes ces étapes lors de la création du DCB.

Recherche documentaire et analyse de la documentation

Le lexicographe s'appuie toujours sur une solide documentation pour identifier les éléments linguistiques qui doivent être inclus dans une entrée. Cette documentation est, d'une part, lexicographique (c'est-à-dire qu'elle s'appuie sur des dictionnaires existants) et, d'autre part, textuelle (c'est-à-dire qu'elle utilise des textes pour illustrer l'usage).

Si nous examinons d'abord la documentation lexicographique, nous remarquons qu'on publie de plus en plus de dictionnaires électroniques, la plupart sous forme de CD-ROM. C'est le cas du *Grand Robert*, du *Petit Robert*, du *New Standard Oxford English Dictionary* et des dictionnaires bilingues anglais/français *Robert-Collins* et *Oxford-Hachette*, pour n'en nommer que quelques-uns. Ces dictionnaires électroniques ne présentent pas seulement des avantages d'ordre pratique (par exemple le lexicographe n'a pas besoin de se déplacer pour aller chercher un dictionnaire, car il peut le consulter directement à partir de son poste de travail), mais permettent aussi l'accès à des données qui se trouvent dans les dictionnaires papier mais qui sont plus difficiles à repérer. Ainsi, si le lexicographe cherche un exemple d'usage d'une unité lexicale donnée dans un dictionnaire papier, il cherche évidemment dans l'entrée pour cette unité ; mais le dictionnaire électronique offre la possibilité de faire une recherche « plein texte », ce qui lui fournit beaucoup plus d'exemples, car le moteur de recherche repère l'unité lexicale en question nonobstant l'entrée sous laquelle elle est notée. En outre, étant donné que la consultation des dictionnaires électroniques est plus rapide[2], le lexicographe, qui travaille généralement sous pression, a la possibilité de rechercher plus d'informations qu'il ne peut normalement le faire s'il doit tourner les pages une à une.

Pour préparer une entrée de dictionnaire bilingue, le lexicographe consulte tout d'abord les dictionnaires de la langue source (c'est-à-dire la langue du mot à l'étude) et, à une étape ultérieure, les dictionnaires bilingues, suivis des dictionnaires de la langue d'arrivée (c'est-à-dire la langue des équivalents). En somme, les lexicographes du DCB consultent jusqu'à une trentaine de dictionnaires. Malheureusement, étant donné que la plupart des dictionnaires ne sont pas encore disponibles en version électronique, ils utilisent beaucoup de dictionnaires papier, malgré le degré d'informatisation élevé du projet.

La deuxième source d'information pour les lexicographes d'aujourd'hui — et peut-être la plus importante — est une collection de textes dans lesquels ils peuvent étudier la façon dont un mot est en fait employé. Avant la deuxième moitié des années 1980, les lexicographes n'avaient recours qu'à un certain nombre d'exemples extraits de textes. Avec les moyens informatiques maintenant à leur disposition, ils ont accès à d'importants corpus (collection de textes informatisés). Pour la création de son dictionnaire bilingue, le DCB a établi un corpus spécial, TEXTUM[3], qui réunit des textes unilingues en français et en anglais. Même si la majorité de ces textes sont canadiens, certains représentent le français et l'anglais de d'autres parties du monde pour comparer et identifier les canadianismes (c'est-à-dire les particularités lexicales du français et de l'anglais canadiens). TEXTUM, qui contient maintenant plus de 310 millions de mots, est divisé en sous-corpus selon la langue, l'origine et la nature des textes, comme l'illustre le tableau qui suit :

TEXTUM

ANGLAIS	TAILLE (en millions de mots)	FRANÇAIS	TAILLE (en millions de mots)
Canadian Press (**N+P, G, CD**)	129,0	Presse canadienne-française (**N+P, G, CD**)	77,0
Wall Street Journal (**N, G+ST, US**)	41,8	*Le Monde* (**N, G, FR**)	17,1
Gazette (**N+P, G, CD**)	6,7	*Ouest France* (**N, G, FR**)	49
Queen's (**N+P+F, G, CD**)	5,0	Leméac (**F, G, CD**)	0,9
Department of Energy (**GD, ST, US**)	27,2	ACFAS (**P, ST, CD**)	13
Canadian Geographic (**P, G, CD**)	0,3		

Légende :
G = général
ST = scientifique ou technique
GD = documents gouvernementaux
CD = canadien
N = journal
P = magazine
F = fiction
US = américain
FR = France

Si nous avons établi notre propre corpus, c'est parce qu'il n'existait pas, au début de notre projet, un corpus qui convenait à nos besoins lexicographiques[4]. Il nous fallait non seulement un corpus de textes en français et en anglais, mais celui-ci devait aussi être majoritairement « parallèle », c'est-à-dire composé de textes dans chaque langue et dont le contenu et le style étaient pratiquement identiques, et donc comparables. Le corpus devait aussi être de taille importante : un petit corpus n'est pas adapté à la lexicographie puisque le nombre d'occurrences de la plupart des mots n'est pas suffisant pour bien cerner leurs sens et leur usage.

Nous nous servons de TEXTUM pour analyser les mots de la langue de départ et pour vérifier les équivalents de la langue d'arrivée, et non pour trouver des équivalents (surtout ceux qui ne sont pas évidents). Pour ce faire, nous utilisons un corpus bilingue de traduction, TransBase, qui se compose de huit ans de débats du parlement canadien contenus dans le *Journal des débats* du parlement canadien, communément appelé le *Hansard*. Ce corpus de traduction, qui a été aligné en bitexte par les chercheurs de l'ancien Centre d'innovation en technologies de l'information (CITI)[5], n'est pas subdivisé en sous-corpus, car le type de textes ne change pas, même si les sujets abordés dans ce journal sont d'une grande variété. C'est un corpus beaucoup moins grand que TEXTUM (47 millions de mots).

Les deux corpus décrits ci-dessus répondent à la très grande majorité de nos besoins lexicographiques, même si, pour certains mots plus techniques, nous consultons également des textes sur Internet. L'avantage principal d'un corpus (par rapport à des textes électroniques disponibles sur Internet ou en format CD-ROM) est qu'on peut le consulter en se servant d'un analyseur de textes qui produit des concordances (c'est-à-dire un concordancier). La concordance permet aux lexicographes de voir plus clairement le « comportement » d'un mot en contexte (*patterns of use*). Elle permet de répondre à de nombreuses questions : le mot est-il couramment utilisé ? est-il normalement suivi d'une préposition ? laquelle ? fait-il partie d'une collocation, d'un composé, d'une expression figée ? dans quel sens est-il utilisé le plus souvent ? y a-t-il un équivalent qui est utilisé beaucoup plus souvent que les autres ? Plus le concordancier est souple et puissant, plus on peut tirer de renseignements du corpus.

Puisque la nature des deux corpus est différente (TEXTUM contenant des textes unilingues, TransBase étant un bitexte), nous devons utiliser des concordanciers différents pour les consulter.

PAT, le concordancier qui permet de consulter TEXTUM, produit des concordances pour chaque unité lexicale sur laquelle travaille le lexicographe. Plus précisément, il établit une concordance pour une chaîne de caractères donnée, car TEXTUM n'est pas lemmatisé. Le nombre de lignes de contexte que contient cette concordance peut être déterminé par le lexicographe, en fonction de la complexité de l'unité lexicale à l'étude et du nombre d'occurrences de la forme linguistique, qui est affiché par PAT dès que le lexicographe cherche une forme. En voici un exemple :

```
>> unifamilial
7 : 408 matches

>> pr sample.20

      is secteurs  unifamilial,     multifamilial et copropriété - , un s
     . Une maison  unifamiliale     atteint un sommet de $496000 à Westmo
     our la maison  unifamiliale    dans une proportion de 84 %, tandis
     ré une maison  unifamiliale    détachée. Elle n'est pas non plus sit
     ur une maison  unifamiliale    évaluée à 120800 $, l'impôt foncier se
      ie « Résidence unifamiliale », l'architecte (et propriétaire) Mar
     ir une maison  unifamiliale    neuve doivent quant à eux débourser t
      de la maison  unifamiliale    sans les inconvénients. Les Terrasse
     ur une maison  unifamiliale,   un logement en copropriété ou dans u
     leurs maisons  unifamiliales   à Cowansville et Vaudreuil, un bâtim
      15 propriétés unifamiliales   à vendre pour un acheteur sur le mar
      e des maisons unifamiliales   d'un projet immobilier à Saint-Pierr
      er de maisons unifamiliales,  De multiples facteurs contribuent
      de résidences unifamiliales   en 1994, pour un total de 14 500 uni
     u'aux maisons  unifamiliales   et copropriétés, le Conseil régional
        ces, soit 210 unifamiliales et trois duplex. « À évaluation égal
      de résidences unifamiliales. La dernière série, baptisée Exclusi
     un paysage d'  unifamiliales, on gaspille beaucoup d'énergie à dé
      t aux maisons  unifamiliales  qu'aux maisons semi-détachées, en ra
      s des maisons  unifamiliales  situées à flanc de montagne. C'est s
```

Même si TEXTUM n'est pas lemmatisé, les capacités de recherche de PAT nous permettent, comme le montre l'exemple ci-dessus, de repérer trois des quatre variantes morphologiques de *unifamilial*. En effet, PAT rend possible des recherches assez sophistiquées. Il possède cependant un défaut majeur : il n'aligne les contextes alphabétiquement qu'à droite de la forme recherchée. Ainsi, pour l'alignement à gauche et pour d'autres types de manipulations, nos lexicographes soumettent la concordance déjà produite par PAT au concordancier MicroConcord.

Étant donné que TransBase est un bitexte, il faut un concordancier spécial pour le consulter. Nous utilisons donc le « bi-concordancier » TransSearch, qui a été élaboré par les chercheurs du CITI. Ce bi-concordancier possède diverses fonctions particulièrement intéressantes pour la lexicographie, notamment la possibilité de préciser la langue du texte de départ, la possibilité d'effectuer des recherches soit pour une chaîne de caractères soit pour un lemme, la possibilité d'éliminer d'emblée certains équivalents (ex. *chefferie (français) NOT leadership (anglais)*), la flexibilité pour rechercher des collocations dont les éléments constitutifs sont à une distance donnée maximale précisée par le lexicographe. Voici un extrait d'une concordance produite par TransSearch.

TransBase : *melody*	
I would just like, if I may, to put to my hon. Colleague this question. Is there anything that an American President could request of this government and its array of Mulroney choristers that would not sing ? Is there any tune so base, any **melody** so disharmonius or any accompaniment so vile that it would be deemed past him ?	Permettez-moi de demander à mon honorable collègue si, à son avis, il y a une chose qu'un président américain pourrait demander et qui serait si odieuse que même les béni-oui-oui de la chorale Mulroney ne seraient pas prêts à l'accepter.
For instance, the present legal definition of a musical work is 'any combination of **melody** and harmony, or either of them, printed, reduced to writing or otherwise graphically produced or reproduced.	Par exemple, la définition juridique actuelle d'une œuvre musicale, soit « toute combinaison de mélodie et d'harmonie, ou l'une ou l'autre, imprimée, manuscrite ou d'autre façon produite ou reproduite graphiquement » est en cause.
One of the theme songs of this government and one of the lovely **melodies** that it has played time and time again as it has inflicted its philosophy on the Canadian prople over the last eight years is harmonization.	L'un des thèmes favoris du gouvernement et une de ces douces mélodies qu'il joue constamment depuis huit ans pour essayer d'inculquer ses principes à la population canadienne, c'est l'harmonisation.
The quicker the government can bring us into harmony with every practice and every *melody* of the United States, the happier it will be.	Plus tôt les États-Unis pourront harmoniser nos pratiques avec toutes les leurs, plus ils seront heureux.

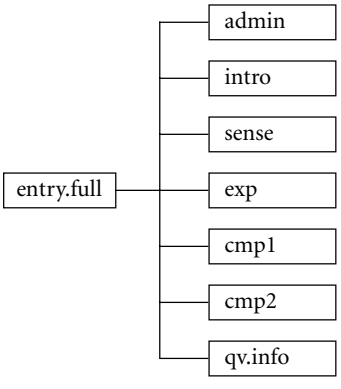
Même si les ressources électroniques signalées ci-dessus permettent aux lexicographes d'identifier plus facilement les renseignements linguistiques nécessaires à la préparation des entrées, ils impriment néanmoins les pages pertinentes, car la lexicographie exige une analyse humaine très poussée ainsi qu'une comparaison soignée des données fournies par des sources différentes. Ce n'est qu'à la fin de cette analyse que les lexicographes du DCB sont en position de choisir les éléments qui doivent figurer dans les entrées qu'ils préparent.

Préparation des entrées

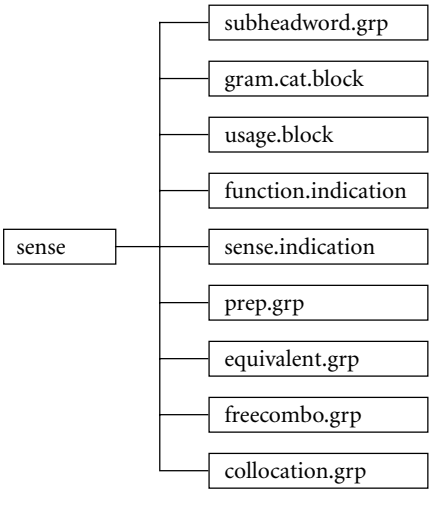
Les données sélectionnées par les lexicographes pour un mot-vedette sont compilées dans une entrée de format prédéterminé. Ce format, qui constitue ce qu'on appelle dans le jargon informatique « *Document Type Definition* » ou DTD, est assez complexe, car nous l'avons conçu pour tenir compte de tous les renseignements qui peuvent figurer dans n'importe quelle entrée. Néanmoins, il est souple puisque seuls certains éléments d'information sont obligatoires quelle que soit l'entrée : le mot-vedette, la partie du discours, une division sémantique, par exemple. En outre, le format hiérarchique permet d'une part d'identifier les parties les plus importantes de l'entrée (par exemple zone d'introduction, division(s) sémantique(s), section des composés, section des expressions figées), et, d'autre part, de subdiviser chacune de ces parties en sous-parties, qui sont elles-mêmes subdivisées plus loin, ce qui permet

d'ajouter beaucoup de détails sur chaque partie importante. En effet, la DTD est tellement longue et complexe qu'il est impossible de la voir tout entière, même à l'écran. Nous n'en présentons donc ici que quelques parties.

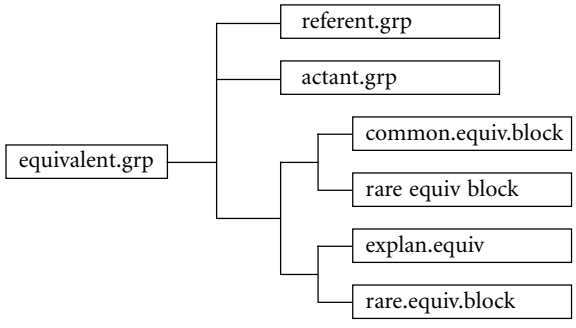
Commençons d'abord par les grandes divisions mentionnées ci-dessus, qui constituent le premier niveau du format hiérarchique :



Nous présentons maintenant le deuxième niveau hiérarchique de la grande division SENSE.



Enfin, en prenant une des subdivisions qui se trouvent dans la figure précédente (equivalent group), nous allons montrer tous les éléments qu'elle contient.



Une fois le format de l'entrée type établi par un comité spécial de l'équipe du DCB travaillant avec un consultant, une DTD a été créée. Cette DTD permet aux lexicographes de rédiger des entrées en SGML (*Standard General Markup Language*). Pour ce faire, ils utilisent divers logiciels, dont WordPerfect SGML.

Voici l'entrée pour le verbe *s'abrier* tel qu'il apparaît à l'écran lorsque le lexicographe la rédige en SGML.

…<intro><headword.block><headword>**s'abrier** </headword>
<gram.cat.block><gram.ind>*vpr* </gram.ind> </gram.cat.block><usage.block>
[<register> *(informal)* </register> <geographic>*(CD*</geographic>*)*</usage.block>*]*
</headword.block> <variant.block><variant> *or* **s'abriller** </variant> <usage.block>
[<register> *(informal)* </register> <geographic> *(CD* </geographic>*)*</usage.block>*]*
</variant.block> </intro><sense> **1** <sense.indication>*(se couvrir d'une couverture* </
sense.indication> RQ+JB+JBA*)* <equivalent.grp> <common.equiv.block><equivalent> to
cover (o.s.) up (with a blanket, etc.) </equivalent> DCF </common.equiv.block>,
<common.equiv.block><equivalent> to pull up the covers </equivalent> CM+JB </
common.equiv.block> </equivalent.grp><freecombo.grp>. *<full.example.block>
<sl.example.block><sl.example> **s'abrier avec une douillette** </sl.example> RQ+JB </
sl.example.block> <tl.example.grp> <tl.example.block><translation>to cover up with a
comforter </translation>MT</tl.example.block></tl.example.grp> </
full.example.block><full.example.block> ;<sl.example.block> <sl.example> **s'abrier
jusqu'au cou** </sl.example> RP+JB </
sl.example.block><tl.example.grp><tl.example.block> <translation> to pull the covers
up to one's chin </translation> CM </tl.example.block><tl.example.grp> </
full.example.block> </freecombo.grp></sense> …

Les entrées préparées en format SGML sont ensuite stockées dans une base de données lexicographiques. Nous pouvons aussi imprimer chaque entrée de façon à ce qu'elle ressemble en effet à une vraie entrée de dictionnaire.

s'abrier *vpr [(informal) (CD)] or* **s'abriller** *[(informal) (CD)] (se couvrir d'une couverture)* to cover (o.s.) up (with a blanket, etc.), to pull up the covers * **s'abrier avec une douillette** to cover up with a comforter ; **s'abrier jusqu'au cou** to pull the covers up to one's chin.

Révision des entrées

C'est sur l'entrée imprimée que se penchent les réviseurs ; en effet, ces derniers préfèrent travailler sur l'entrée complète, ce qui n'est pas toujours possible lorsqu'ils révisent l'entrée à l'écran. La possibilité d'étaler côte à côte l'entrée et les documents consultés et imprimés par les lexicographes lors de la rédaction facilite la révision.

À cette étape, l'informatique joue plutôt un rôle d'arrière-plan. Il arrive, par exemple, que les réviseurs aient besoin de consulter eux-mêmes les corpus pour clarifier certains points ou pour trouver d'autres exemples d'usage. De plus, après chaque révision[6], le lexicographe responsable de l'entrée modifie, en fonction des changements proposés, la version informatisée de l'entrée. Toutes les versions d'une entrée sont sauvegardées dans notre base de données lexicographiques, ce qui permet à un réviseur d'examiner les changements déjà apportés par les différents réviseurs.

Enfin, on consulte cette base de données au moyen du logiciel LiveLink Search, ce qui permet de trouver une entrée donnée ou encore d'extraire un groupe d'entrées qui répond à certains critères de recherche précis (par exemple les entrées qui sont des prépositions ou qui contiennent des canadianismes). Cette fonction permet d'assurer une certaine uniformité dans le traitement lexicographique.

Par ailleurs, une base de données sur mesure, notre « Workflow database », nous permet de suivre chacune des étapes par lesquelles passe une entrée. Cette information nous permet, par exemple, d'identifier les entrées terminées et les entrées en cours de préparation et de produire des statistiques quant au cheminement des dossiers.

Conclusion

Tel que promis dans le titre de l'article, nous avons voulu décrire le rôle de l'informatique dans la recherche lexicographique, et plus particulièrement les moyens informatiques utilisés au DCB. Force est d'admettre que sans les outils informatiques courants tels que Telnet, Internet et le courrier électronique, il faudrait beaucoup plus de temps et d'argent pour compléter le DCB puisqu'ils permettent aux chercheurs des trois centres de travailler en collaboration. Le corpus TEXTUM, stocké à l'Université de Montréal, et le corpus TransBase, stocké à l'Université d'Ottawa, peuvent être consultés par Telnet ou par Internet à partir des autres centres. En outre, tous les chercheurs peuvent consulter les entrées stockées dans la base de données lexicographiques ainsi que les modifier. Enfin, la communication par courrier électronique entre les chercheurs est on ne peut plus essentielle dans le cadre d'un projet mené en collaboration. En fait, aujourd'hui, nous serait impossible d'envisager ce projet sans les moyens informatiques dont nous profitons pleinement, même si au lancement du projet, en 1988, nous ne faisions qu'en rêver.

NOTES

1. Voir à ce sujet F. Knowles, "The Computer in Lexicography", dans *Dictionaries : An Encyclopaedia of Lexicography*, vol. 2, p. 1645, Walter de Gruyter, Berlin.
2. Cela suppose que l'on connaisse bien les moteurs de recherche, ce qui n'est pas nécessairement le cas car ils varient quelque peu d'un dictionnaire électronique à l'autre.
3. Le nom TEXTUM vient du fait que le corpus de *text*es est stocké à l'*U*niversité de *M*ontréal.

4. En fait, il n'y avait en 1988 qu'un corpus électronique important, celui du Strathy Institute de l'Université Queen's.
5. Le CITI (maintenant le RALI) nous a accordé une licence pour l'exploitation de ce corpus.
6. Les entrées du DCB peuvent être révisées jusqu'à cinq fois : (a) révision de la langue de départ ; (b) révision de la langue d'arrivée ; (c) révision globale ; (d) révision des canadianismes (surtout français) ; (e) révision finale.