La BDTS-concordances : un outil d'enrichissement de la pratique lexicographique

Chantal-Édith Masson, Hélène Cajolet-Laganière et Pierre Martel

Université de Sherbrooke – Sherbrooke –Québec – Canada
chantal-edith.masson@usherbrooke.ca, helene.cajolet.laganiere@usherbrooke.ca

Abstract

BDTS-concordances: an important technological tool for the enhancement of the lexicographic practice. The present paper lies within the research framework of the Centre d'Analyse et de Traitement Informatique du Français Québécois (CATIFQ) and, more particularly, of the FRANQUS research group from the Université de Sherbrooke. The objective of the FRANQUS project is to develop an original nomenclature of the French language used in Quebec, mainly of its standard usage. The current lecture is explicitly related to the initial exploitation of the *Banque de données textuelles de Sherbrooke (BDTS)* (The Sherbrooke textual database) using a tool that was developed and that is known as BDTS-concordances. First, we will introduce the BDTS, and then the BDTS-concordances from a methodological perspective. Second, using examples, we will illustrate how BDTS high-frequency words are treated in order to write lexicographic articles. It appears that contexts, for terms with a frequency rate inferior to 100, can use an "ordinary" treatment, meaning that the memory retention of the information was possible. On the other hand, close to 20% of the BDTS terms show a frequency rate of 100 or more. Those consequently received a uniform treatment based on an original model. This constancy is achieved through a Definition of Type of Document (DTD-XML). The data is entered directly in a computerized form which is displayed on the DTD in four main sections: 1-entry, 2-co-occurences and constructions, 3-comment, and 4-writing/prog. We will present a detailed description of this file and illustrate, using examples of high-frequency words, how the BDTS-concordances works in its present state, its contributions and limits. We will also point out the possible avenues considered for its optimization.

Résumé

La présente communication s'inscrit dans le cadre des travaux du Centre d'analyse et de traitement informatique du français québécois (CATIFQ), et plus particulièrement du groupe de recherche FRANQUS de l'Université de Sherbrooke. Le projet FRANQUS vise à l'élaboration d'une nomenclature originale du français en usage au Québec, essentiellement de son usage standard. L'objet du présent exposé concerne explicitement une première exploitation de la Banque de données textuelles de Sherbrooke (BDTS) grâce à l'outil développé communément nommé BDTS-concordances. Nous présenterons dans un premier temps la BDTS, puis la BDTS-concordances d'un point de vue méthodologique. Dans un deuxième temps, nous illustrerons, à l'aide d'exemples, le traitement des mots de fréquence élevée de la BDTS aux fins de la rédaction des articles lexicographiques. Il est apparu que les contextes, pour les vocables ayant une fréquence inférieure à 100, se satisfaisaient bien d'un traitement « ordinaire », c'est-à-dire que la rétention en mémoire des informations était possible. Par ailleurs, près de 20 % des vocables de la BDTS présentent une fréquence de 100 ou plus et ont fait ainsi l'objet d'un traitement uniforme selon un modèle original. C'est une Définition de Type de Document (DTD-XML) qui assure cette constance. Les informations sont saisies directement dans la fiche informatisée qui se déploie sur la DTD en quatre grandes sections : 1- entrée, 2 –cooccurrents et constructions, 3- remarque et 4- rédaction. Nous présenterons une description détaillée de ce fichier et nous illustrerons, grâce à quelques exemples de mots fréquents, le fonctionnement, les apports, de même que les limites de la BDTS-concordances telle qu'elle fonctionne actuellement. Nous ferons également état des pistes envisagées pour son optimisation.

Mots-clés : corpus, base de données textuelles, analyse, contextes, concordances, redondance, structuration, DTD, XML, informatisation.

1. Introduction

La présente communication s'inscrit dans le cadre des travaux du Centre d'analyse et de traitement informatique du français québécois, et plus particulièrement du groupe de recherche FRANQUS (français québécois : usage standard) de l'Université de Sherbrooke, responsable du projet « Nomenclatures, description et application dans les technologies de l'information et de la communication ». Ce projet est échelonné sur cinq ans; le groupe entreprend sa troisième année de travail. Le calendrier des activités prévoyait, au cours des deux premières années, la conception de tous les outils linguistiques (notamment, la constitution des bases documentaires) et informatiques (notamment, la construction de la plateforme informatique et des modèles de traitement) visant à l'élaboration d'une nomenclature originale du français en usage au Québec, essentiellement de son usage standard, et au support de l'activité de rédaction des articles lexicographiques. L'objet du présent exposé concerne explicitement une première exploitation de la Banque de données textuelles de Sherbrooke (BDTS) grâce à l'outil développé communément nommé BDTS-concordances. L'exploitation de banques de données textuelles est devenue un incontournable en linguistique, notamment en lexicographie.

De fait, la plupart des banques de données textuelles, tant anglaises que françaises, ont été constituées dans le cadre de projets visant à l'élaboration de dictionnaires. Néanmoins, un problème se pose : comment traiter ces masses de données textuelles, particulièrement pour les mots de fréquence élevée, de manière à faciliter le travail des rédacteurs et rédactrices et à respecter un échéancier dans un projet dictionnairique. Nous présenterons, dans un premier temps la BDTS, puis la BDTS-concordances d'un point de vue méthodologique, ce dernier point, appuyé de quelques exemples de traitement de mots de fréquence élevée de la BDTS aux fins de la rédaction des articles lexicographiques, en prenant soin de faire état des apports et des limites de l'ensemble ainsi que des projections de développement.

2. Banque de données textuelles de Sherbrooke (BDTS)

La BDTS est un corpus de textes mis sur pied par une équipe de chercheurs du Centre d'analyse et de traitement informatique du français québécois (CATIFQ). Il s'agit d'une banque de textes représentatifs des différents usages du français en usage au Québec[1], de diverses situations de communication et de divers registres de langue; elle contient plus de 37 millions d'occurrences, ceci, sans compter les corpus complémentaires[2] requis par son caractère volontairement évolutif. La BDTS est conçue de manière à réunir le plus grand nombre possible de thèmes, de discours et de niveaux de langue du français contemporain oral et écrit en usage au Québec. La diversité des textes sélectionnés vise à une « représentativité » du français en usage au Québec dans un contexte nord-américain. Il importe de préciser toutefois qu'il ne s'agit pas d'une représentativité au sens statistique du terme (échantillonnage non probabiliste raisonné par quota) : en effet, la notion de représentativité de la BDTS doit plutôt être associée à la variété des textes qui la composent et qui reflètent la langue générale (orale et écrite) de même que la langue littéraire, journalistique et relativement plus spécialisée utilisée au Québec dans différentes situations de communication. Il est entendu que tous les

[1] Pour la notion de français en usage au Québec, nous faisons référence aux textes stockés aux fins de diverses analyses et rédigés dans l'espace géographique du Québec. Ces textes sont considérés dans leur totalité.

[2] De fait, la BDTS mise à jour totalise aujourd'hui quelque 45 M de mots, ce qui augmente d'autant la tâche de traitement et justifie la mise en œuvre de moyens destinés à en réduire l'impact.

emplois lexicaux ou sémantiques rencontrés dans les textes de la BDTS n'appartiennent pas nécessairement à la langue dite « standard »[3].

2.1. *Typologie et composition de la BDTS*

Les 37 millions d'occurrences de la BDTS sont répartis selon cinq types de discours : 59 % de textes spécialisés (près de 22 millions d'occurrences), 16 % de textes littéraires (près de six millions d'occurrences), 14 % de textes journalistiques (plus de cinq millions d'occurrences), 6 % de textes didactiques (plus de deux millions d'occurrences) et 5 % de textes oraux (près de deux millions d'occurrences)[4]. La ventilation des pourcentages associés aux diverses catégories de textes correspond à l'objectif visé : avoir accès aux différents registres de langue, notamment le niveau standard, actuellement en usage au Québec. Les 37 millions d'occurrences de la BDTS sont répartis dans plus de 10 000 textes contemporains, couvrant les années 1960 jusqu'à 2002.

Le corpus de base contient environ 206 000 formes différentes ou quelque 192 000 formes si l'on exclut les chiffres. Comme le but du projet de description lexicographique concerne le français contemporain en usage au Québec, la plupart des textes sont postérieurs à 1960 (date choisie de façon arbitraire, mais correspondant au début de la Révolution tranquille et marquant le début du Québec moderne). Cependant, certains textes littéraires antérieurs à 1960 ont été sélectionnés en raison de leur valeur et de leurs qualités intrinsèques reconnues par les Québécois. La répartition des vocables selon les types de source s'énonce ainsi :

Le sous-ensemble de langue spécialisée contient des textes qui fournissent des mots de langue générale et le vocabulaire de base de nombreux domaines spécialisés. Il renferme en outre des textes nombreux et diversifiés : monographies, mémoires, thèses, rapports, livres, documents administratifs, etc. **Le sous-ensemble de langue littéraire** comprend des textes tirés de romans, d'essais, de poèmes, de nouvelles, de récits, de contes, de pièces de théâtre, etc. **Le sous-ensemble de langue journalistique** contient des articles tirés de différentes publications québécoises, comme des quotidiens (*La Presse, Le Devoir, Le Soleil, Le Droit,* etc.), des magazines ou périodiques spécialisés (*L'Actualité, Voir, Québec Science, Franc-Vert, Interface*, etc.). **Le sous-ensemble de langue didactique** comprend des textes tirés de manuels scolaires (de niveau secondaire, collégial, universitaire), de logiciels informatiques (textes de présentation et guides de l'utilisateur), de manuels pour une formation technique, etc. Enfin, **le sous-ensemble de langue orale** contient, d'une part, des échantillons de langue parlée spontanée, c'est-à-dire des transcriptions d'enquêtes sociolinguistiques orales réalisées dans différentes régions du Québec, et d'autre part, des échantillons de langue moins spontanée (contes, monologues, téléromans, textes radiophoniques et télévisés, tribunes téléphoniques, témoignages, etc.).

2.2. *Brève caractérisation de la BDTS*

La BDTS, si elle est une banque de taille relativement modeste en comparaison de celles des grandes banques européennes construites essentiellement à des fins lexicographiques (par

[3] Le mot *standard*, employé dans le contexte de ce projet de description lexicographique, désigne une catégorie de textes dont le niveau de langue est généralement soigné ou soutenu (textes littéraires reconnus, documents administratifs et législatifs soignés, manuels scolaires primés, productions scientifiques, livres et essais de tous genres publiés par des intellectuels, textes journalistiques, etc.); il s'oppose principalement à des textes oraux de niveau « familier » ou « très familier », etc.

[4] Nous tenons à remercier Marie-Claude Lavallée, Geneviève Labrecque, Chantal Fontaine, Lucie Lahaie et Katherine Pérusse, chercheures et rédactrices au groupe FRANQUS pour toutes les données et les informations qu'elles nous ont fournies.

exemple le *Trésor de la langue française*), présente un intérêt particulier en vertu de sa composition lorsqu'on la compare aux autres banques de données textuelles de la francophonie : Frantext est composée de 80 % de textes littéraires et 20 % de textes techniques, Beltext, de 100 % de textes littéraires et oraux, Québétext, de 100 % de textes littéraires et Suistext, de 100 % de textes littéraires.

Elle se distingue également à deux autres titres. En effet, outre son caractère évolutif évoqué plus haut, la BDTS est entièrement informatisée et indexée, ce qui en permet une consultation efficace et l'obtention rapide de toutes les références, les concordances et les attestations des mots. Ce n'est pas son informatisation et son indexation proprement dites qui la distinguent, mais plutôt les choix qui ont été faits au moment de son retraitement. En effet, entamée il y a plusieurs années, sa constitution s'était faite dans une coquille « propriétaire » opérable exclusivement en mode texte et selon des fonctionnalités prédéfinies. Si la manipulation des données, dans cette coquille, était économique et efficace, il demeure que cette manipulation était réservée aux initiés et que les résultats obtenus se montraient très peu « collaboratifs » (avec d'autres applications), autant de limitations difficilement acceptables. Pour ces raisons, et compte tenu du fait que tous les chercheurs et les rédacteurs de FRANQUS ne sont pas des informaticiens, mais doivent cependant interroger quotidiennement la BDTS, compte tenu aussi de notre objectif de développement d'une plateforme intégrée de traitement et d'une interface unifiée d'interrogation (toutes deux actuellement fonctionnelles et en usage), il nous fallait adopter un langage à la fois capable de soutenir une structuration (au besoin) très fine des informations, leur hiérarchisation, tout en étant non propriétaire (ISO) et convivial. Il y a 6 ans déjà, nous faisions le choix du langage de balisage structurel extensible XML dans cet esprit. Toutes nos bases, sans exception, sont structurées ainsi, les préexistantes, telles que la BDTS ayant fait l'objet d'une structuration et d'une conversion parallèlement à la poursuite de son développement. La BDTS XMLisée, dont les sorties sont d'autant plus lisibles qu'elles sont mises en forme à l'aide de feuilles de style XSL, se prête à tous les besoins puisqu'il nous est maintenant possible de développer « à la carte » toutes les routines informatiques nécessaires (en langage JAVA), le tout, sur nos propres serveurs hautement sécurisés.

3. La BDTS-concordances : une première exploitation de la BDTS

Malgré la convivialité supérieure du XML, la difficulté cognitive et fonctionnelle de la manipulation/consultation du corpus dans l'identification, la recension et l'analyse des contextes pour les mots de fréquence élevée est rapidement apparue, qu'ils s'agisse d'une sortie imprimée ou d'une sortie sur écran. Il fallait exercer un contrôle sur la masse documentaire, laquelle était imposante à cause de la redondance des informations. La BDTS-concordances constitue une première exploitation de la BDTS. Elle fournit une première analyse des contextes pour chaque mot destiné à faire partie de la nomenclature du dictionnaire en cours d'élaboration à Sherbrooke. Elle a pour but de réduire la tâche d'examen des contextes de deux manières travaillant en conjonction, soit une structuration de leur présentation de manière à faciliter l'identification des constructions et des cooccurrents les plus récurrents ainsi qu'un contrôle (au moins partiel) de la redondance des contextes pour les mots de fréquence élevée.

La première étape de ce processus a consisté en des balayages successifs de la BDTS à l'aide de la liste des vocables déjà identifiés comme devant faire partie de notre nomenclature. Cette procédure, réalisée par nos informaticiens, était destinée à alléger la tâche des rédacteurs en leur fournissant, sous une commode forme textuelle, toutes les occurrences d'un vocable, mises en évidence et « encadrées » de leurs contextes. Parallèlement, elle permettait de vali-

der les informations contenues dans les dictionnaires français et québécois (et compilées dans la Base Thésaurus documentaire) et de repérer les nouveautés, c'est-à-dire les informations spécifiques de la BDTS. L'ensemble se présentait ainsi sous la forme d'une liste non structurée :

```
un songe éveillé. Comment ne pas se |*griser d'illusions, comment ne pas renaît
i des compensations l'incitait à se |*griser des colères qu'il pouvait faire à
rait! Jamais plus il ne pourrait se |*griser de_l espérance d'un bonheur venant
 chaleur de_la maison acheva de les |*griser en battant aux tempes et contre le
it du fond du seau finissait par me |*griser. Puis le lait s'ajoutant au lait,
eine. L'inconnu ne pouvait que nous |*griser. C'étaient des images d'un audacie
à l'autre, à l'ivre enfant Qui vous |*grisa de mots, certain soir triomphant. C
geaient, qui la soutenaient, qui la |*grisaient du fond de_la glace. Plus tard,
quelques gorgées de vin mousseux la |*grisaient plus_qu un demi#litre de whisky
|P73 saveur aux aliments. Ce qui la |*grisait surtout, c'était dans une glace p
en pointe, l'amiral, timidement, se |*grisait tous les soirs. Les deux |P140 so
et au_fur_et_à_mesure que le vin me |*grisait. Cela devenait grossier. En elle
is Chateaubriand», et cette idée me |*grisait. J'avais le sentiment de communie
 Il pourrait lui parler. Déjà il se |*grisait d'images et d'attouchements. Mais
ui peut se révéler particulièrement |*grisant pour ceux qui apprennent à écrire
en va et, en même temps, c$est très |*grisant parce qu'on est les premiers à le
mptueux que le Forum, ait été moins |*grisant, c$est compréhensible. Que Joe Ca
banques ont mieux résisté au climat |*grisant des marchés financiers d'alors. M
```

Figure 1. Exemple du traitement du vocable « griser » et de ses formes à la première étape

Après des tests auprès de nos rédacteurs, il est apparu que les contextes, pour les vocables ayant une fréquence égale ou inférieure à 100, se satisfaisaient bien d'un tel traitement, c'est-à-dire que la rétention des informations en mémoire, un élagage en temp réel de la redondance de même que l'identification des informations d'intérêt étaient possibles. Ce constat a conduit à la mise de côté temporaire d'un traitement plus poussé de tels mots, ce qui permettait de concentrer l'attention sur les mots de fréquences moyenne et élevée, nettement plus problématiques

Avec près de 20 % des mots de la nomenclature présentant une fréquence de 100 ou plus, le modèle de structuration à retenir pour la construction proprement dite des documents XML composant la BDTS-concordances devait permettre un traitement uniforme et conforme à nos besoins de même qu'au type de données. C'est une Définition de Type de Document (DTD-XML) qui assure cette constance et permet la construction de documents non seulement bien formés, mais valides.

Les contraintes logistiques étaient nombreuses puisque l'ensemble des bases regroupées dans la plateforme intégrée (BDI) comptent plusieurs dizaines de milliers de documents XML, dont plusieurs centaines peuvent être en circulation simutanément. D'un point de vue informatique, pour assurer la complétude du traitement de la liste de vocables composant notre nomenclature, pour connaître l'état d'avancement de chaque document (et pour éviter des erreurs telles que l'écrasement de fichiers), un fichier « squelette », préstructuré et comportant déjà certaines informations était prégénéré pour chacun des fichiers de cette liste. On remarquera que ces fichiers XML se sont vu attribuer une extension .CNC permettant de les différencier des fichiers « éponymes » des autres bases, d'où leur surnom de *fiche CNC*. Les autres informations étaient saisies directement dans la fiche informatisée qui se déploie sur la DTD. Celle-ci structure et regroupe les informations sous 4 grandes sections : (1) ***Entrée***, (2) ***Cooccurrents et constructions***, (3) ***Remarque*** et (4) ***Rédaction***. Elles se présentent ainsi :

« Entrée »

Elle se subdivise en trois sous-sections (dans les deux premières, les informations sont générées automatiquement). Il s'agit respectivement, plus précisément :

- du mot en entrée, de la catégorie grammaticale du vocable (appelé ici mot-pivot) dans le Thésaurus documentaire (TD) et de sa fréquence dans la BDTS. Il convient de préciser que le Thésaurus documentaire comprend une compilation des réseaux sémantiques des principaux dictionnaires usuels du français;
- des formes présentes en contextes. Dans cette partie se trouvent les formes du mot-pivot présentes dans les concordances, avec leur fréquence;
- des expressions, exemples et locutions (appelés « syntagmes » aux fins de notre recherche). Sous cette section sont groupés ces syntagmes rattachés au mot-pivot et répertoriés dans le Thésaurus documentaire (donc dans les dictionnaires), et présentés de façon hiérarchisée selon leurs sens et sous-sens, etc.

« Cooccurrents et constructions »

Cette partie de la fiche est celle où le rédacteur intervient le plus. Elle comporte un tableau qui rend compte des cooccurrents du mot-pivot, donnant, pour chaque cooccurrent, sa fréquence, son lemme et ses formes, extraits, cette fois, de la BDTS.

« Remarque »

Au bas de la fiche, un espace (de taille indéfinie) est prévu pour accueillir les commentaires des rédacteurs de fiches de la BDTS-concordances.

« Rédaction »

Tout comme dans le Thésaurus documentaire et les autres fichiers de la Base de données intégrée (BDI), la fiche comporte une partie « prog » dans laquelle est encapsulée l'information nécessaire au fonctionnement de la BDI.

4. Un exemple de traitement : le verbe *atterrir*

4.1. *Les modalités de la construction de la fiche CNC (BDTS-concordances)*

À une première étape, il s'agit de récupérer tous les sens déjà encodés dans la fiche XML correspondante stockée dans la Base Thésaurus Documentaire (TD) lorsqu'une telle fiche existe. En effet, tous les mots retenus à la nomenclature du dictionnaire propriétaire ne disposent pas d'une telle fiche puisque celle-ci, se rappelle-t-on, est une compilation des réseaux sémantiques établis dans un ensemble de dictionnaires existants; par conséquent, les mots spécifiques à la BDTS ne s'y trouvent pas représentés. Il s'agit d'abord de coller, dans les cases « cooccurrents » de la fiche CNC, les sens tirés du TD, afin de créer (à des fins de structuration et de hiérarchisation) les rubriques de la fiche. En une seconde étape, il s'agit, dans la case « constructions » associée à chaque rubrique, de coller un nombre de contextes (tirés de la BDTS) suffisant pour avoir une bonne idée de l'usage du mot dans ce sens.

4.2. *Le résultat du traitement et la fiche CNC ainsi constituée*

Une fois constituée, la fiche, qui fait état de la fréquence totale du mot-pivot (156), de la distribution de ses formes, du nombre et de la distribution de ses cooccurrents (259), qui ordonne également les « syntagmes » selon les sens et les sous-sens, etc., se présente ainsi lorsque automatiquement mise en forme à l'aide de la feuille de style (XSL) dans l'interface unifiée :

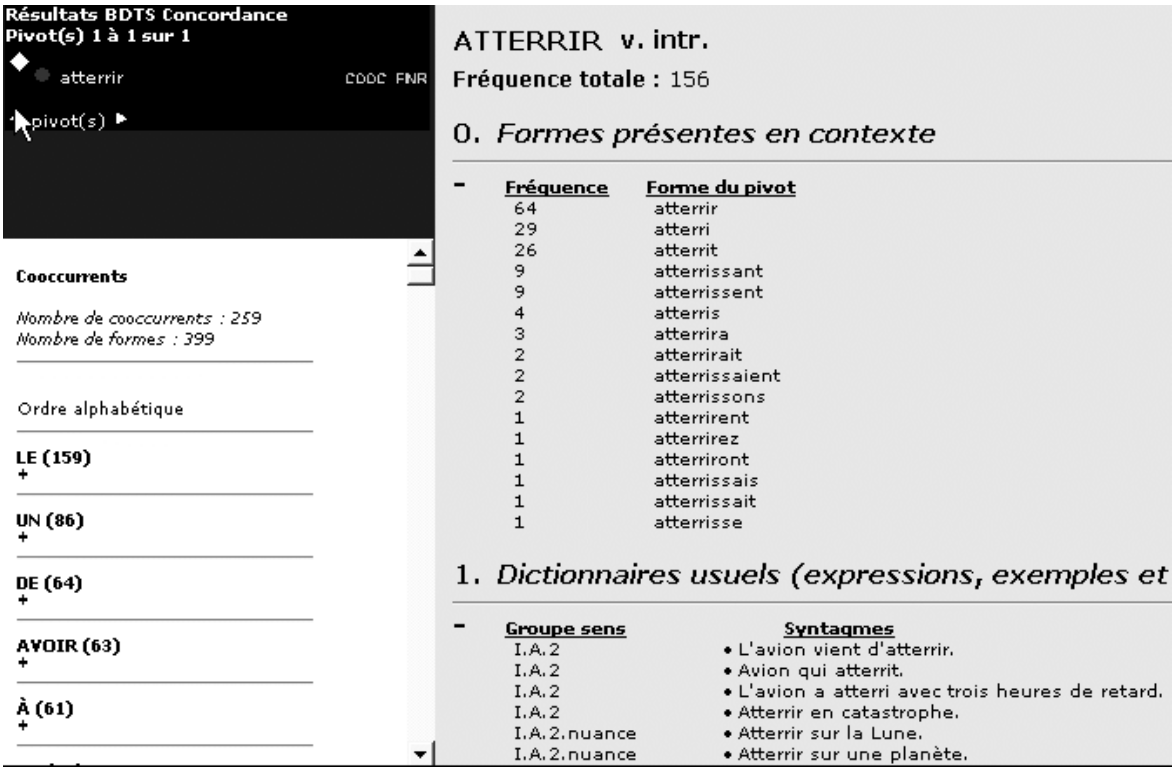

Figure 2. Vue très fragmentaire de la fiche CNC (BDTS-concordances) du verbe « atterrir »

5. Intérêt et limites de la BDTS-concordances

La construction des premières centaines de fiches CNC a été l'occasion de tester le modèle, la méthode, les instruments de même que le protocole de réalisation de la tâche et de vérifier l'adéquation et l'intérêt de cet ensemble. Les conclusions de ce processus pourraient se moduler comme suit.

5.1. *L'intérêt des fiches CNC (BDTS-concordances)*

En plus de sa très grande lisibilité et de sa richesse informationnelle, le document (la fiche informatisée complétée) qui en résulte présente l'avantage de pouvoir être intégré dans la sous-base de données (BDTS-concordances) qui lui est consacrée dans la super Base intégrée (BDI). Cette intégration en permet l'interrogation dans une interface unique, concurremment d'ailleurs, à celle de documents d'autres sous-bases également intégrées. Un tel document fournit alors rapidement, grâce à la liste structurée de cooccurrents intégrés à la fiche, des pistes pour les cooccurrences et collocations fréquentes.

À l'usage, ces fiches CNC/XML s'avèrent plus particulièrement utiles pour le traitement (la rédaction des articles) de certaines catégories de mots, notamment des mots outils. En effet, dans ceux-ci, souvent hyper-fréquents, la redondance des cooccurents est particulièrement marquée et se prête bien non seulement en soi à une structuration mais, surtout, à la synthèse, le tout sans perte d'informations (ou, à tout le moins, avec une perte calculée comme minimale et perçue comme tolérable). Ici, toujours à l'usage, il est apparu préférable de travailler à partir d'une synthèse plutôt qu'avoir l'ensemble des contextes : il est en effet peu probable qu'un rédacteur ou un réviseur souhaite relire les 27 651 contextes du mot *avant*!

5.2. *Les limites des fiches CNC (BDTS-concordances)*

En ce qui concerne le traitement des mots autres que les mots outils, il semble que certaines nuances doivent être exprimées et que cette approche constitue un choix moins intéressant pour 2 principales raisons :

- La fiche CNC ne conserve pas de traces du travail fait sur les contextes ;
- Le rédacteur ne dispose pas de l'ensemble des contextes or l'expérience a montré que tous les rédacteurs souhaitaient voir l'*ensemble* des contextes, même si la fiche CNC existe.

Pour conserver son intérêt, il est apparu que le document synthèse de l'analyse des contextes de la BDTS que constitue la fiche CNC doit être élaboré par le rédacteur à qui est confié un mot aux fins de rédaction de l'article. Ce faisant, il dispose d'une vision globale du mot et est au courant des analyses réalisées sur les contextes, une condition qui est apparue difficilement négociable quand il s'agissait d'accorder sa confiance aux résultats de tris et à l'organisation des informations. Dans cette perspective, il est apparu que la spécialisation du travail n'était pas souhaitable puisque le rédacteur, en dépit de l'existence d'une fiche CNC réalisée par un autre que lui-même, ressentait le besoin de retourner dans la Banque pour faire son travail. Également, il est apparu que le retour différé sur un article donné (quelques mois plus tard, par exemple) pouvait même requérir de procéder à nouveau à l'analyse des contextes, la synthèse « inscrite » dans la fiche CNC ne satisfaisant pas nécessairement le réviseur.

En dernier lieu, l'examen des pratiques des rédacteurs a démontré que le travail dans l'éditeur XML dans lequel se déploient les DTD sous la forme de fiches informatisées n'était pas également aisé pour tous et que certains préféraient traiter manuellement les contextes dans un éditeur de texte. Mais ce qui est au cœur de la réorientation de la méthode de travail en ce qui concerne le traitement des contextes est beaucoup plus fondamental. C'est le résultat d'une « projection » de la tâche de rédaction des articles au moment du traitement des contextes des verbes, combinaison qui a marqué un changement progressif de perspective en faveur d'une approche d'analyse et de structuration des informations fondée sur la sémantique plutôt que sur la syntaxe et ce, pour tous les types de vocables, y compris les mots outils. Il s'agit là d'une façon de faire innovatrice dans le domaine de la lexicographie appliquée.

5.3. *Une perspective sémantique plutôt que syntaxique : la méthode « alternative » de traitement des contextes*

Lors du traitement des contextes des verbes notamment, plusieurs rédacteurs ont progressivement développé (et préféré employer) une autre méthode qui leur permettait, celle-là, non seulement de créer une synthèse de l'analyse des contextes, mais encore de laisser des traces du travail d'analyse ayant conduit à cette synthèse. Moins axée sur la réduction de la masse documentaire que sur la structuration de cette masse, la procédure s'est progressivement cristallisée autour d'un tri fondé sur la sémantique plutôt que sur la syntaxe. Cette méthode initialement « alternative » fait appel au dossier-sens et aux fichiers textuels de contextes (voir à ce sujet le point 2 et la Figure 1).

Il s'agit d'abord de procéder à l'extraction des paires sens-construction des dictionnaires (disponibles dans les fiches du Thésaurus Documentaire) et de créer un dossier-sens. L'étape suivante consiste à trier les contextes selon les paires contenues dans ce dossier et selon les spécificités de la Banque, façon de faire qui présente l'avantage de permettre l'enrichissement du dossier-sens par l'intégration des « nouveautés » identifiées dans la BDTS.

Il en résulte un nouveau fichier textuel dans lequel **tous** les contextes sont présents, mais regroupés par paires sens-construction (voir la Figure 3, plus bas). Ce fichier est complété par un second document, le dossier-sens lui-même, qui propose la synthèse de l'analyse : paires sens-construction avec fréquences associées, regroupées par blocs sémantiques, avec commentaires au besoin (voir le tableau ci-dessous). La complétude du tandem ainsi formé est telle qu'il redevient envisageable de spécialiser certains rédacteurs dans leur construction et ce, à la satisfaction de la majorité des autres rédacteurs qui peuvent alors se concentrer sur la tâche de rédaction des articles.

GRISER [fréquence = 52 après tri des contextes]

###[(rendre un peu ivre) GRISER QQN : A] [v. trans. dir.] **[5]**
###[(rendre un peu ivre) SE GRISER : S] [v. pron. réfléchi] **[3]**
###[(fig. enivrer de, repaître de) GRISER QQN DE : A] [v. trans. dir.] **[17]**
###[(fig. enivrer de, repaître de) ÊTRE GRISÉ DE, PAR : S] [voix passive] **[11]**
###[(fig. enivrer de, repaître de) GRISÉ : S] [part. passé adj.] **[2]**
###[(fig. enivrer de, repaître de) SE GRISER DE QQCH : A] [v. pron. réfléchi] **[13]**
###[(ternir) GRISER QQCH : S] [v. trans. dir.] **[1]**
###[IGNORÉS] **[2]**

Tableau 1. Exemple du contenu du dossier-sens pour le vocable « griser »

Il convient de remarquer qu'il est toujours également possible de consulter une fiche CNC qui aurait été construite pour avoir un aperçu des cooccurrents avant de commencer le tri des contextes; cependant, dans la plupart des cas, le fait d'avoir les contextes alignés sur le pipe « |* » permet de repérer ces cooccurrents rapidement, visuellement, ce qu'on ne pouvait faire avant d'avoir les contextes sous cette forme.

Comme il est possible de le constater dans la saisie d'écran (très fragmentaire) suivante, non seulement les différentes formes du vocable s'affichent et se regroupent sous le sens approprié, mais aussi le nombre de contextes, sous chacun, est-il rapporté. De plus, il est possible à l'aide de quelques commandes de retrier à volonté les informations.

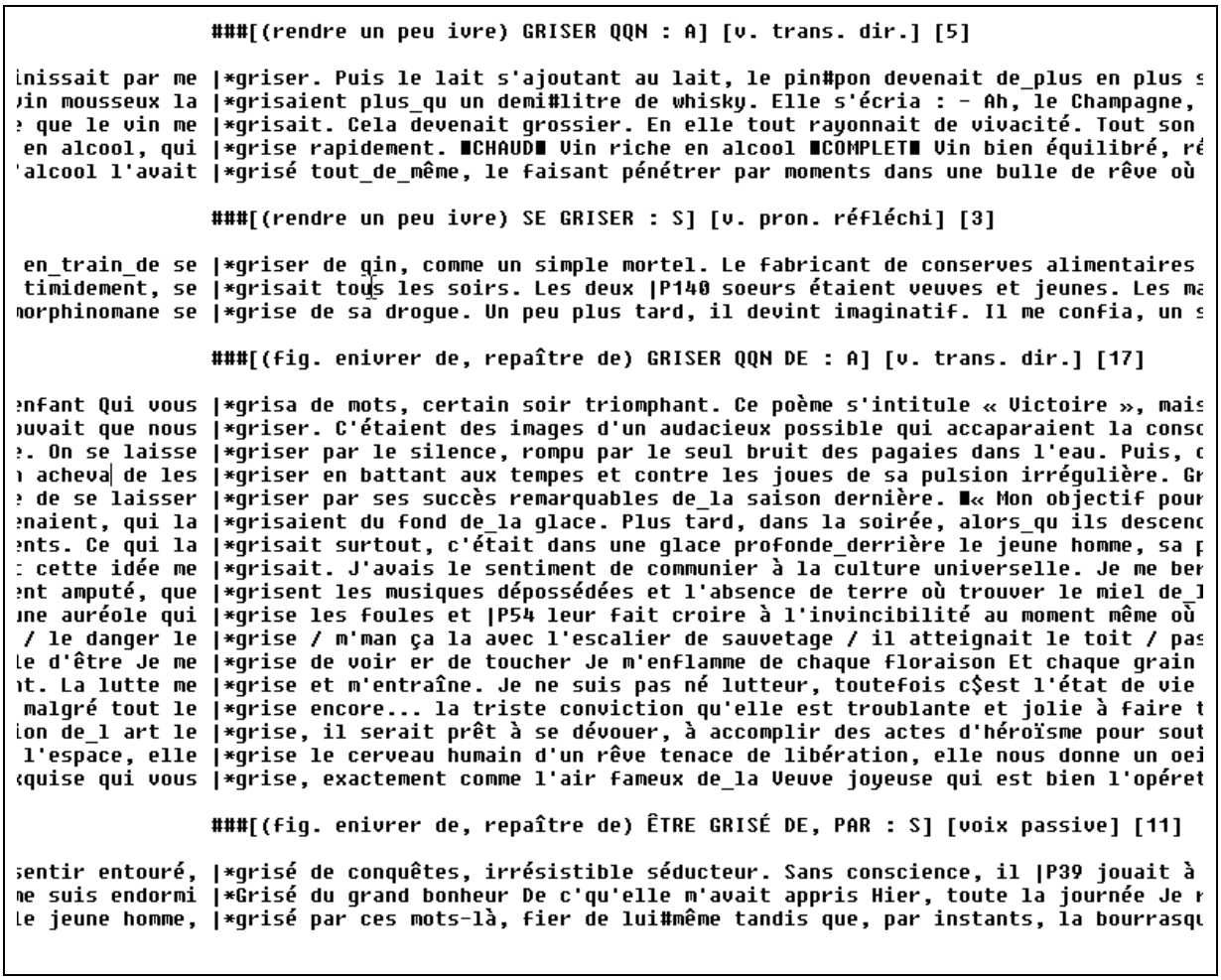

Figure 3. La présentation des contextes triés par sens

5.4. *Intérêts et limites de la méthode « alternative »*

Cette méthode est apparue à la fois rapide, simple et efficace. Elle permet de garder une trace claire du travail d'analyse effectué et de conserver l'accès à tous les contextes. Ceux-ci sont maintenant également triés par paires construction-sens et regroupés en blocs sémantiques, ce qui constitue une avancée très intéressante et nous permet d'envisager, à moyen terme, la constitution d'un dictionnaire de cooccurrents sémantiquement groupés.

Avec, en main, la synthèse et l'ensemble des contextes triés, les rédacteurs éprouvent peu le besoin de retourner dans la Banque, sauf dans les cas où le contexte de 100 caractères sur lequel ils souhaitent se pencher particulièrement n'est pas suffisant ou pour puiser dans le corpus de citations. Elle peut également, par la suite, alléger le travail de révision des articles dans la mesure où il n'est pas nécessaire de refaire les analyses, mais seulement de les évaluer. Tous ces avantages nous indiquent qu'il y a lieu de favoriser cette méthode et de trouver les moyens de pallier sa seule carence (majeure, cependant) : les documents ainsi produits ne sont pas accessibles par la BDI, alors que cette intégration constitue l'une des raisons d'être de la plateforme informatique.

6. Conclusion

La BDTS-concordances constitue, en soi, un outil important et ce, à plusieurs titres. L'allègement et l'optimisation du traitement des contextes des mots outils et des autres mots

de fréquence très élevée identifiés dans la Banque de données textuelles est un impératif. Cependant, dans la pratique, il est apparu qu'avec leur forme actuelle, l'utilité des fiches développées selon le protocole initial n'est pas aussi grande que ce qui était anticipé lorsqu'il s'agit du traitement de vocables d'un autre type dans le cadre d'une tâche dictionnairique. En effet, les rédacteurs, s'ils désirent accéder aisément aux cooccurrents et aux constructions, réclament néanmoins la possibilité de voir la totalité des contextes. Il fallait donc se poser la question de l'intérêt actuel de la poursuite de la constitution de fiches CNC telles quelles, ce qui pouvait exiger une suspension de l'idéal de lisibilité et, surtout, d'accessibilité et de collaborativité tant locales (fiches consultables dans l'interface unique de la plateforme informatique) qu'étendues que permet l' « universalité » du langage de balisage structurel XML (ISO). Ces objectifs d'accessibilité et de collaborativité qui ont gouverné (et gouvernent encore) le développement de la plateforme intégrée propriétaire de FRANQUS ne devaient pas aller à l'encontre de notre conviction suivant laquelle il revient à l'informatique d'instrumenter une tâche et non à celle-ci de se plier à des contraintes informatiques, une perspective réductrice.

La lexicographie est une pratique qui doit absolument tenir compte des impératifs cognitifs de ceux-là mêmes qui en sont les artisans. Une solution possible au maintien de la **procédure** initialement prévue nous vient des rédacteurs eux-mêmes alors qu'ils énoncent qu'elle n'est réellement « intéressante » que lorsqu'ils prennent eux-mêmes en charge la construction de la fiche CNC du vocable dont la description leur est confiée. Dans cet esprit, il est apparu qu'il n'y avait pas lieu de spécialiser les rédacteurs par type de tâches (par exemple, rédaction des fiches CNC, rédaction des articles), mais plutôt de favoriser une approche dont l'homogénéité serait fondée et maximisée par une spécialisation thématique et/ou grammaticale plutôt (type de partie du discours).

L'ouverture avec laquelle la question a été abordée s'est traduite par des gains alors que les rédacteurs ont développé, parallèlement, une façon de faire permettant de combler leurs besoins tout en privilégiant une procédure fondée sur des critères de structuration sémantiques plutôt que syntaxiques. La voie est ouverte vers une BDTS-concordances « nouvelle et améliorée », plus riche en termes de contenu et porteuse de nouveaux développements, par exemple la constitution, à moyen terme, d'un dictionnaire des cooccurrents groupés de manière sémantique. Il s'agira maintenant pour nous de développer les modèles, les stratégies et les routines nécessaires à une prise en charge informatique correspondant à nos objectifs.

Références

Brunet Ét. *Hyperbase v. 5.4 : logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels*.

Fellbaum C. (1998). *WordNet, an Electronic Lexical Database.* MIT Press.

Gorcy G. (1990). Le Trésor de la langue française. Son originalité et les voies ouvertes pour son informatisation. In *CNRS/INaLF Dictionnairique et lexicographie. Autour d'un dictionnaire : le Trésor de la langue française, témoignages d'atelier et voies nouvelles*. Didier Érudition : 187-207.

Leroy-Turcan I. et Wooldridge R. (1998). *Quelques exemples des acquis de la base informatisée de la première édition du Dictionnaire de l'Académie française (1694).* Document hypertexte. http://www.chass.utoronto.ca/~wulfric/academie/acad1694/quebec298.htm.

Masson C.-É. (2001). *Le traitement des substantifs dans Le Robert sur CD-ROM : modélisation, formalisation et proposition méthodologique en vue de son informatisation*. Thèse de doctorat, Université de Sherbrooke.

Pruvost J. (2000). *Dictionnaires et nouvelles technologies.* Presses Universitaires de France.

Tutin A. et Wionet C. (1998). *Informatisation du Dictionnaire universel de Furetière revu par Basnage de Bauval (1702) : premier bilan.* Document hypertexte. http://www.chass.utoronto.ca/~wulfric/siehlda/dicta1998/tw_tab.htm.

University of Virginia. *TEI Guidelines for Electronic Text Encoding and Interchange*. Document hypertexte. http://etext.lib.virginia.edu/bin/tei-tocs?div=DIV2&id=DIEN.

Véronis J. et Ide N. (1996). Encodage des dictionnaires électroniques : problèmes et propositions de la TEI. In Piotrowski D. (Ed.), *Lexicographie et informatique. Autour de l'informatisation du Trésor de la langue française. Actes du Colloque international de Nancy*. Didier Érudition : 239-261.

Wooldridge R. *Baliser un texte, c'est le penser : le cas du Dictionnaire de l'Académie française.* Document hypertexte. http://www.chass.utoronto.ca/~wulfric/articles/gehlf597/index.html.