Structuration et balisage sémantique des définitions du *Trésor de Langue Française informatisé* (TLFi)

Lucie Barque & Alain Polguère
OLST—Université de Montréal
C.P. 6128, succ. Centre-ville,
Montréal (QC) H3C 3J7 — Canada
lucie.barque, alain.polguere@gmail.com

Abstract

We present an ongoing lexicographic project which aims at providing an explicit and formalized structuring for the definitions of the French electronic dictionary *Trésor de la Langue Française informatisé* (TLFi). Section 1 deals with the general issue of formal structuring of lexicographic definitions; section 2 introduces the notion of analytical definition and gives an overview of its use in French lexicography, particularly in the TLFi; section 3 details our project, its methodology and consecutive steps.

1 Problème de la structuration des définitions lexicographiques

Il n'existe pas à l'heure actuelle pour le français (et, à notre connaissance, pour aucune autre langue) de base de données lexicale, libre d'accès et à très large couverture, proposant pour chaque unité lexicale décrite une définition explicitement structurée. Par *définition explicitement structurée*, nous entendons une définition lexicographique munie d'un balisage formel (XML ou autre) indiquant :

1. la structure en composantes définitionnelles de type genre prochain et différences spécifiques ;
2. le rôle joué par chaque composante dans la caractérisation du sens de l'unité lexicale définie.

Examinons, à fin d'illustration, la définition (1) ci-dessous de l'acception de base de TROTTINETTE, tirée du *Trésor de la Langue Française informatisé*, désormais *TLFi* (Dendien & Pierrel, 2003).

(1) TROTTINETTE (acception **A.1**) : Jouet composé d'une plate-forme allongée montée sur deux petites roues et d'un guidon à direction articulée, que l'enfant fait avancer en s'aidant d'un pied qu'il pose régulièrement par terre pour donner l'impulsion ou en actionnant une pédale en un mouvement de va-et-vient.

Une version structurée de cette définition devrait indiquer explicitement :
– la présence de trois composantes définitionnelles majeures ;
– le fait que *jouet* est la composante centrale — identifiant le genre prochain —, qui caractérise l'unité lexicale décrite comme dénotant un type de 'jouet' ;
– le fait que les composantes (i) *composée … direction articulée* et (ii) *que l'enfant … va-et-vient* sont des composantes périphériques — différences spécifiques —, qui spécifient la composante centrale, respectivement, en termes de 'parties constituantes' et 'mode d'utilisation'.

Il est clair qu'une base de données de définitions lexicographiques structurées d'une telle façon serait extrêmement précieuse, aussi bien pour la communauté du traitement automatique de la langue (Barnbrook, 2002, pages 2–7), que pour la recherche en linguistique et la lexicographie.

Dans le présent article, nous présentons un projet, en cours d'exécution, qui vise la conversion des définitions du TLFi sous une forme structurée. Ce projet s'appuie sur les acquis du projet BDéf de base de définitions lexicales formalisées (Altman & Polguère, 2003; Barque, 2008). Contrairement à ce qui a été fait avec la BDéf, nous visons une formalisation se limitant à la caractérisation de la structure des définitions en termes de blocs définitionnels, alors que la DBéf s'attachait aussi à la formalisation des composantes elles-mêmes en termes de propositions élémentaires écrites dans un langage contrôlé. Cependant, le présent projet est plus ambitieux dans la mesure où il vise le développement à moyen terme d'une ressource lexicale à très large couverture, alors que le travail sur la BDéf s'est concentré sur la formalisation d'un sous-ensemble du lexique français : celui décrit dans les quatre volumes publiés du *Dictionnaire explicatif et combinatoire du français contemporain* ou DEC (Mel'čuk et al., 1984, 1988, 1992, 1999).

Il y a deux raisons pour lesquelles nous appuyons notre travail sur le TLFi, plutôt que sur tout autre dictionnaire du français. Premièrement, les définitions de ce dictionnaire sont dans leur grande majorité des définitions, appelées *définitions analytiques*, qui respectent certaines contraintes de forme et de contenu permettant leur structuration formelle en termes de composantes définitionnelles [1]. Deuxièmement, ce dictionnaire, développé durant plus de trente ans à l'INALF (maintenant ATILF), est la seule modélisation complète de haute qualité du lexique français librement accessible sous forme électronique à fin de recherche [2].

Dans ce qui suit, section 2, nous spécifierons ce qu'est une définition analytique, puisque c'est sur ce type particulier de définition lexicographique que nous travaillons. Ensuite, section 3, nous décrirons le projet actuellement en cours : corpus traités, stratégie adoptée et principales étapes de traitement des données.

2 Notion de définition analytique

2.1 Délimitation de la notion

Il existe une ample littérature sur la définition lexicographique, littérature concernant aussi bien les aspects strictement lexicographiques de la question — par exemple, (Wierzbicka, 1987; Dostie et al., 1999; Barnbrook, 2002; Rundell, 2008) — que ses aspects plus théoriques — par exemple, (Fodor et al., 1980; Wierzbicka, 1996; Riemer, 2006). Il ne serait pas pertinent de faire ici une synthèse de tout ce qui a été dit sur cette notion et nous nous contenterons, tout d'abord, d'une caractérisation approximative de ce que l'on peut entendre par *définition lexicographique*.

> Une définition lexicographique d'une lexie [3] L est (i) une analyse de son sens, (ii) qui prend la forme d'une paraphrase linguistique de L.

Bien qu'assez vague, cette caractérisation exclut du domaine de la définition lexicographique cetaines modélisations sémantiques des lexies, comme par exemple celles du *Lexique génératif* (Pustejovsky, 1995), ou toutes autres modélisations fondées sur le langage de la logique formelle, puisqu'il ne s'agit pas à proprement parler d'**énoncés linguistiques** paraphrasant les lexies décrites.

La caractérisation ci-dessus exclut aussi les descriptions qui ne seraient pas des **analyses** sémantiques. Ainsi, Atkins & Rundell (2008, pages 36–38) considèrent que, parmi les trois descriptions sémantiques de l'acception de base du vocable anglais DISTURB données ci-dessous, seules les deux dernières ont le statut de paraphrases définitionnelles :

(2) a. to intrude on ; interrupt [*Collins English Dictionary*, 2006]

b. If you disturb someone, you break their rest, peace or privacy [*Collins School Dictionary*, 2006]

c. To interrupt someone and stop them from continuing what they were doing [*Macmillan English Dictionary for Advanced Learners*, 2002]

1. Nous traiterons brièvement des cas de définitions non analytiques à la fin de la section suivante.

2. Nous sommes infiniment reconnaissants à Jean-Marie Pierrel, directeur de l'ATLIF, de nous avoir donné accès aux fichiers source du TLFi.

3. C'est-à-dire, une unité lexicale. Dans cet article, nous utilisons la terminologie de l'approche de la Lexicologie Explicative et Combinatoire, telle que présentée dans (Mel'čuk et al., 1995) ou (Polguère, 2008).

Examinons brièvement les caractéristiques propres à chacune de ces descriptions :

– Contrairement à ce que disent Atkins & Rundell, la description (2a) a recours au paraphrasage, puisque le quasi-synonyme est la plus élémentaire des paraphrases. Ils ont cependant raison de considérer qu'une telle description ne **définit** pas le sens, car elle ne l'analyse pas. L'équivalent quasi-synoymique n'est donc pas à proprement parler une définition lexicographique (une analyse du sens) [4].
– La description (2b), quant à elle, définit véritablement DISTURB, en mettant au jour son contenu sémantique. On pourrait se questionner quant à la nature paraphrastique de la description proposée. Nous sommes cependant d'accord pour considérer avec Atkins & Rundell qu'il s'agit bien là d'un paraphrasage. En effet, l'énoncé linguistique *If you disturb someone...* peut être facilement reformulé sous forme d'une égalité du type « definiendum = definiens » :
  X disturb Y = X break Y's rest, peace or privacy
  Le format choisi par le dictionnaire est simplement considéré par les rédacteurs comme plus approprié, d'un point de vue pédagogique, que celui que nous allons maintenant examiner.
– La description (2c) est bien entendu elle aussi une analyse paraphrastique ; contrairement à la précédente, elle ne prend en compte que le définiens. Elle semble plus rigoureuse, plus « académique » d'un certain point de vue, mais on peut noter qu'elle présente l'inconvénient de ne pas introduire de façon explicite la structure actancielle complète de la lexie définie et sa correspondance dans la paraphrase, contrairement à (2b), qui a recours aux deux pseudo-variables *you* and *someone* pour nommer, respectivement, le premier et le second actant de DISTURB [5].

Nous voyons donc que le critère de paraphrase (pris dans un sens large) n'est pas suffisant pour caractériser un type de définition considéré comme satisfaisant : il faut aussi que la définition analyse le sens de l'unité lexicale définie en termes de sens plus simples. Un tel principe est bien entendu appliqué dans nombre de dictionnaires grand public, ainsi qu'en lexicographie théorique — notamment, en Lexicographie Explicative et Combinatoire (Mel'čuk et al., 1995) et dans l'approche du *Natural Semantic Metalanguage* (Wierzbicka, 1996).

La dernière contrainte couramment appliquée aux définitions lexicographiques qui vont nous intéresser ici — et qui n'est pas présente dans la caractérisation des définitions lexicales donnée plus haut – est le principe aristotélicien de décomposition par genre prochain et différences spécifiques : on doit retrouver au cœur de la définition une composante centrale (= genre prochain) qui correspond au noyau sémantique hyperonymique de l'unité définie, accompagnée d'une ou plusieurs composantes périphériques (= différences spécifiques) qui particularisent l'unité définie par rapport à son genre prochain et à ses cohyponymes. Sémantiquement, la composante centrale de la définition est la composante sémantiquement dominante de la représentation sémantique de celle-ci [6] ; elle a de ce fait une fonction classifiante, dans la mesure où elle permet de regrouper des lexies ayant un même noyau sémantique. Syntaxiquement, il s'agit d'un lexème ou d'un syntagme qui est au sommet de la structure syntaxique de la définition. Ainsi, *to interrupt* est la composante centrale de la définition (2b) ci-dessus.

Les définitions lexicographiques qui respectent tous les critères qui viennent d'être examinés — analyses paraphrastiques formulées en termes de genre prochain et différences spécifiques — sont appelées *définitions analytiques* dans (Polguère, 2008, page 183). C'est désormais spécifiquement à ce type de définitions que nous nous intéresserons dans la suite de cet article.

4. On notera que ce point de vue n'est pas nécessairement partagé par tous. Ainsi, Benson et al. (1986, pages 203–204) considèrent le recours aux (quasi-)équivalents synonymiques comme un type particulier de définition lexicographique, qu'ils appellent *synonym definition*.
5. Voir (Rundell, 2008), pour une brève comparaison entre les définitions sous forme de phrase complète (angl. *full-sentence definitions*), du type (2b), et les définitions strictement paraphrastiques, du type (2c).
6. Sur la dominance communicative dans les réseaux sémantiques, voir (Polguère, 1997).

2.2 Pratique lexicographique

Le recours à la définition analytique est la norme dans les dictionnaires de référence du français tels le TLFi [7] ou le *Nouveau Petit Robert*. On doit cependant souligner trois cas fréquents de transgression de ce principe définitionnel dans ces dictionnaires.

Premièrement, il est bien connu que les dictionnaires décrivent le contenu sémantique et/ou syntaxique des lexies grammaticales (articles, prépositions régies du type À et DE, etc.) par des énoncés qui ne sont en aucun cas des paraphrases analytiques, comme l'illustrent les deux descriptions ci-dessous de l'article indéfini français, tel qu'employé dans *Marc veut avoir **une** petite sœur* :

(3) a. UN (acception **II.A.1**) : Désigne un objet, un élément distinct mais indéterminé. [*Petit Robert*, 2007]

b. UN[2] (acception **I.A.1**) : [Empl. spécifique. *J'ai acheté un livre pour enfants.* Ce qui est dit est vrai d'un seul livre, pris sur l'ensemble des livres pour enfants. Certes, p. oppos. à *le livre*, ce livre n'appartient pas encore au thème de l'énoncé ; il est indéterminé, mais il s'agit d'un livre précis] [TLFi]

Dans le cas de la description (3b), l'usage des crochets indique une caractérisation ≪ sémantico-grammaticale ≫ clairement distinguée d'une définition lexicographique. Il arrive cependant que, dans le TLFi, les caractérisations de ce type et les définitions non analytiques cohabitent pour un même article de lexie, comme l'illustre la caractérisation ci-dessous de la préposition DE, telle qu'employée dans *naviguer **de** Brest à Halifax* :

(4) DE[2] (acception **I.A.1**) : [Orig. spatio-temporelle.] Le point de départ se situe dans l'espace ou dans le temps (s'oppose à la prép. *à*, parfois à *en*, plus rarement à *jusqu'à*), par rapport à un point d'aboutissement dans l'espace ou dans le temps. [TLFi]

Le deuxième cas de non respect du principe de définition analytique est tout simplement le recours à des descriptions par quasi-synonymes, qui ne sont pas, comme nous l'avons vu en 2.1, des analyses du sens. Le TLFi utilise fréquemment les ≪ définitions ≫ par synonymes, comme illustré sous (5) :

(5) FUNAMBULESQUE (acception ≪ figurée ≫) : Fantaisiste, bizarre.

On trouve aussi dans le TLFi beaucoup les reformulations successives d'un même sens ou de sens très proches, comme illustré sous (6) :

(6) a. FRANC[3] (acception **I.A.3**) : Qui est moralement libre ; qui agit de sa propre volonté.

b. EN ÊTRE POUR SES FRAIS (acception ≪ figurée ≫ de la locution, sous FRAIS[2] **C.3**) : S'être mis inutilement en peine, être déçu.

c. FUSTIGER (acception **B**) : Attaquer, combattre, critiquer violemment.

Ce type de définitions pose problème car il faut pouvoir distinguer les cas où il s'agit vraiment d'une reformulation (FRANC[3] **I.A.3**), des cas où il s'agit d'une unité vague, c'est-à-dire d'une unité qui peut dénoter l'un et/ou l'autre des deux sens figurant dans la paraphrase (EN ÊTRE POUR SES FRAIS). Les reformulations peuvent en outre donner lieu à des ambiguïtés de rattachement concernant certaines composantes. On est par exemple en droit de se demander si la composante *violemment*, dans la définition de FUSTIGER **B**, est à rattacher seulement à *critiquer* ou à la composante plus large *attaquer, combattre, critiquer*.

Finalement, le dernier cas de non respect du principe de définition analytique est plus difficile à identifier. Il concerne les définitions paraphrastiques qui, de par leur structure syntaxique, présentent comme genre prochain une composante qui devrait normalement se retrouver parmi les différences spécifiques. Nous avons rencontré un tel cas avec la définition de TROTTINETTE du TLFi — voir (1), en début d'article. La structure syntaxique de cette définition met en évidence le fait que TROTTINETTE est classifiée sémantiquement

7. Voir (Frassi, 2007) pour une étude détaillée des définitions du TLF.

dans le TLFi en tant que ‘jouet’, et non en tant que ‘véhicule (à deux roues)’. Le fait qu’une trottinette est avant tout un véhicule — et seulement secondairement un jouet — n’est qu’implicitement suggéré dans la définition, à travers la description du mode d’utilisation. Au moment de la rédaction de cette définition, les trottinettes étaient utilisées comme jouets et n’avaient pas encore connu leur seconde vie (sous une forme « relookée ») en tant que mode de déplacement urbain branché. Cependant, depuis toujours, la trottinette est avant tout un type de véhicule (‘artefact servant à se déplacer’) et n’est un jouet que de façon secondaire : une trottinette que l’on utilise pour se rendre au travail, et non pour jouer, reste une trottinette.

Ici s’achève cette section sur la notion de définition analytique et son implication dans la structuration **implicite** des définitions des dictionnaires de langue. Nous allons maintenant aborder la présentation de notre projet de structuration formelle explicite de ce type de définitions, projet fondé sur l’utilisation du TLFi en tant que ressource lexicale.

3 Vers une base de définitions formalisées dérivée du TLFi

Nous abordons maintenant la présentation de notre projet de structuration des définitions du TLFi. Après avoir donné quelques informations sur la taille du corpus traité (section 3.1), nous présenterons la stratégie générale adoptée pour ce projet (section 3.2). Puis, nous décrirons les trois étapes principales de sa réalisation : mise en évidence, au moyen d’un balisage XML, de la structuration des définitions en composantes centrale et périphériques (section 3.3), marquage sémantique normalisé de ces composantes (section 3.4) et mise en œuvre d’une base lexicale sémantique dérivées du TLFi (section 3.5).

3.1 Corpus traité

L’identification de la nomenclature de toutes les entités lexicales définies du TLFi n’est pas aisée à faire. En effet, le TLFi compte 54 281 entrées principales (vocables potentiellement polysémiques), auxquelles s’ajoutent 18 096 entrées « enchâssées » correspondant à des dérivés morphologiques d’une entrée principale (par exemple, FRUITARISME sous FRUIT) [8]. Le TLFi contient aussi 59 168 syntagmes définis, syntagmes qui peuvent être de deux types. Tout d’abord, le syntagme défini peut être une locution : par exemple, FRUIT DÉFENDU sous FRUIT[1] ; il s’agit là d’une pratique généralisée dans les dictionnaires de langue [9]. De façon plus originale, nombre de collocations sont aussi définies comme des touts lexicaux, soit dans l’article de la base de la collocation (7a), soit dans celle de son collocatif (7b).

(7) a. *Vol domestique* (sous VOL[2] **A.1**) : Vol commis par un domestique, un employé de maison, soit envers son employeur à l’intérieur des locaux que celui-ci possède et dans lesquels il l’accompagne, soit envers des personnes se trouvant dans l’habitation de son employeur.

b. *Troupes fraîches, chevaux frais* (sous FRAIS[1] **C.1b**) : Troupes, chevaux destinés à remplacer ceux qui sont fatigués.

Pour notre projet, nous avons extrait automatiquement les 271 164 définitions décrivant les sens de ces différents types d’entités lexicales [10].

3.2 Stratégie adoptée

Notre stratégie de développement de la base de définitions est avant tout lexicographique. Nous entendons par là qu’elle s’appuie en premier lieu sur un travail manuel effectué par des lexicographes et des annotateurs [11]. On peut mettre en opposition notre approche d’extraction de données lexicographiques struc-

8. Un peu comme le dictionnaire *Lexis* de Larousse, le TLF subordonne la description de nombreux dérivés morphologiques à celle de leur source morphologique. Il semblerait toutefois que seuls les dérivés rares ou techniques soient traités de cette façon.

9. Même s’il est clair que chaque locution devrait en théorie posséder son entrée propre dans la nomenclature, comme cela est le cas dans le DEC (Mel’čuk et al., 1984, 1988, 1992, 1999).

10. Cette extraction est aisée dans la mesure où les fichiers informatiques du TLFi contiennent un balisage structurant les articles lexicographiques ; ainsi, toutes les définitions sont encadrées par les deux balises `<def>` et `</def>`.

11. Voir, à ce propos, notre guide d’annotation (Barque & Polguère, 2009).

turées avec celle adoptée dans (Barnbrook, 2002) pour l'analyse des définitions du *Collins Cobuild Student's Dictionary* (Sinclair, 1990), qui repose sur des procédures entièrement automatiques. Notre stratégie nous semble plus appropriée pour le présent projet dans la mesure où nos premières observations nous conduisent à penser que l'ensemble des définitions du TLFi forme un tout moins uniforme que l'ensemble des défnitions de dictionnaires commerciaux tels que le *Collins Cobuild* ou, mieux, le *Longman Dictionary of Contemporary English* LDOCE (cf. son métalangage définitionnel contrôlé) [12].

Toutefois, nous n'écartons pas la voie de l'automatisation du traitement, mais en tant que support à l'analyse manuelle. Le corpus à traiter étant énorme (voir *supra*), nous testons actuellement une pré-annotation automatique au moyen du logiciel MACAON, développé par Alexis Nasr [13]. Ce prétraitement présente en outre un intérêt lexicographique, puisque MACAON nous permet de décrire la grammaire des définitions du TLFi et de voir ainsi dans quelle mesure le métalangage définitionnel du TLFi se rapproche d'un sous-langage (Kittredge, 1982) et dans quelle mesure il pourrait être amélioré (voir plus bas, section 3.5).

3.3 Étape 1 : structuration des définitions

La première étape de notre projet, en cours de réalisation, consiste à structurer les définitions du TLFi. Plus précisément, il s'agit de délimiter la composante centrale et les éventuelles composantes périphériques de la paraphrase définitionnelle. Considérons, pour commencer, la définition non structurée de BLINDÉ$_{\text{N}}$:

(8) BLINDÉ$_{\text{N}}$ (nom décrit à l'intérieur de l'article du participe BLINDÉ **II.A**) : Un véhicule militaire blindé.

La tâche la plus délicate consiste à délimiter la composante centrale, qui doit être, rappelons-le, une composante sémantiquement classifiante. Dans notre exemple, il est possible d'hésiter entre *véhicule*, *véhicule militaire* ou, même, *véhicule militaire blindé*, puisque chacune de ces expressions peut se retrouver comme composante centrale de la définition de plusieurs lexies françaises. En l'occurrence, dans le TLFi, *véhicule militaire blindé* ne se retrouve que dans la définition de BLINDÉ$_{\text{N}}$ **II.A** ; mais le contenu sémantique de cette expression se retrouve, exprimé différemment, dans la définition d'autres lexies, comme CHAR :

(9) CHAR (acception **C.3**) : **Engin de guerre motorisé et blindé**, monté sur chenilles et doté d'un armement (mitrailleuses, canons, etc.) et que manœuvrent des soldats placés à l'intérieur.

On choisira donc ici le syntagme le plus long comme composante centrale, sans composante périphérique. Voyons maintenant une définition dont la structuration sera plus standard, celle de BROUETTE :

(10) BROUETTE (acception **B.1**) : Véhicule à une roue et à deux brancards servant au transport des matériaux.

On peut se poser ici le même type de question que dans le cas précédent : faut-il choisir *véhicule* ou *véhicule à une roue* comme composante centrale ? Dans ce cas-ci, la composante centrale doit être *véhicule* car le TLFi ne contient pas de définition d'autres lexies dénotant des véhicules à une roue [14]. Une fois la composante centrale (*véhicule*) délimitée, il devient relativement simple d'isoler les différentes composantes périphériques. Nous présentons ci-dessous la version structurée de cette définition : la composante centrale y est indiquée au moyen d'une balise <CC> et les composantes périphériques au moyen de balises <CP>.

(11) BROUETTE (acception **B.1**)

```
<PARAPH>
  <CC>Véhicule</CC>
  <CP>à une roue et à deux brancards</CP>
  <CP>servant au transport des matériaux</CP>
</PARAPH>
```

12. À propos du traitement automatique des définitions du LDOCE, voir (Fontenelle, 2009), qui vient de nous être signalé et que nous n'avons malheureusement pas encore eu la possibilité de consulter.

13. http://pageperso.lif.univ-mrs.fr/~alexis.nasr/macaon/index.html

14. Il existe bien une entrée SIDE-CAR. Cependant, l'acception de base dénote une ≪ **Caisse** carossée monoplace à une roue ≫. Seule la seconde acception, métonymique, dénote un ≪ **Véhicule** formé par la réunion d'une motocyclette et de cette caisse ≫.

On remarque ici que la séquence *à une roue et à deux brancards* ne forme pas deux mais une seule composante périphérique. Les deux éléments de la conjonction participent en effet de la même façon à la spécification du sens : il s'agit d'indiquer des parties caractéristiques du véhicule. Comme on le voit, nous procédons à une segmentation minimale en composantes périphériques. La règle que nous appliquons consiste à ne délimiter, dans la mesure du possible, que des composantes périphériques ayant un apport de nature distincte vis-à-vis de la composante centrale.

3.4 Étape 2 : Marquage sémantique et construction d'une hiérarchie d'étiquettes sémantiques

Une fois identifiée la structure générale de la définition, nous procéderons au marquage sémantique de ses composantes. Le marquage sémantique consiste tout d'abord à attribuer à la composante centrale une **étiquette sémantique**, c'est-à-dire une expression linguistique normalisée qui rend compte de la valeur sémantique de la composante centrale (Polguère, 2003). Pour ce faire, nous extrairons de notre base de définitions segmentées des ensembles de définitions qui ont une composante centrale identique ou proche, par exemple la composante *véhicule* et toute autre composante elle-même définie par *véhicule* (*voiture*, etc). Puis, dans le cadre du développement de la hiérarchie des étiquettes sémantiques du TLFi [15], nous créerons le jeu d'étiquettes sémantiques qui servira à marquer ces différentes composantes centrales. Dans la série de définitions annotées présentée ci-dessous, les deux étiquettes servant à marquer les composantes centrales sont `véhicule` et `voiture` (étiquette fille de `véhicule`, dans la hiérarchie).

(12) a. BROUETTE : `<CC etiq=`**`véhicule`**`>Véhicule</CC> ...`

b. CARAVANE : `<CC etiq=`**`véhicule`**`>Roulotte</CC> ...`

c. BERLINE : `<CC etiq=`**`voiture`**`>Voiture automobile</CC> ...`

d. ACCÉLÉRIFÈRE : `<CC etiq=`**`voiture`**`>Sorte de diligence ou voiture publique</CC> ...`

Notons que les étiquettes sémantiques doivent être désambiguïsées par le numéro d'acception qui leur correspond dans le TLFi et que leur sens doit être décrit par la définition correspondant à cette acception, comme illustré sous (13a-b).

(13) a. VÉHICULE (acception **II.A**) : Engin constitué d'un châssis muni de roues, à traction animale ou autopropulsé, servant au transport routier ou ferroviaire.

b. VOITURE (acception **B.2**) : Véhicule automobile servant à transporter un nombre réduit de personnes ou des objets de faible encombrement.

Une fois la composante centrale marquée d'une étiquette, il reste à indiquer le **rôle** que joue chaque composante périphérique par rapport à cette composante centrale [16]. La définition de l'étiquette sémantique nous permet déjà de proposer une première série de rôles de composantes périphériques. Par exemple, la définition de l'étiquette `véhicule`, donnée ci-dessus en (13a), présente les **parties** caractéristiques de l'engin (*constitué d'un châssis muni de roues*), son **mode de fonctionnement** (*traction animale ou autopropulsé*) et, enfin, sa **fonction** (*servant au transport routier ou ferrovière*). L'observation des définitions des lexies étiquetées `véhicule` nous permet de confirmer ces différents rôles de composantes périphériques, comme on le voit dans les définitions de REMORQUE (14) et LANDAU (15), et, le cas échant, d'en proposer d'autres : les définitions de BOLIDE (16) et TACOT (17) nous indiquent, par exemple, que la **vitesse** est un type de différence spécifique à standardiser pour le champ sémantique des véhicules.

15. Cette hiérarchie sera développée à partir de celle déjà élaborée dans le cadre du projet de base de données lexicale DiCo, présentée dans (Polguère, 2003).

16. On peut envisager deux approches de la délimitation des **rôles**. Il est possible de considérer que ces rôles forment un ensemble restreint et prédéfini. C'est l'approche adoptée par le *Lexique Génératif* (Pustejovsky, 1995), qui propose une structure lexicale composée de quatre types de traits appelés *qualia* (`formal`, `constitutive`, `telic` et `agentive`). Ainsi, quelle que soit la partie du discours de l'unité définie et quel que soit son type sémantique, l'article représentant son sens sera constitué de ces quatre rôles, auxquels on attribuera ou non une valeur. On peut à l'inverse considérer, et c'est l'approche que nous adoptons, que ces rôles forment un ensemble fini, mais susceptible d'être si vaste qu'il ne pourra être connu qu'à l'issue de la description du lexique.

(14) REMORQUE (acception **B**)

```
<PARAPH>
  <CC étiq=véhicule>Bateau ou véhicule à roues</CC>
  <CP rôle=mode_fonctionnement>dépourvu d'un moyen de propulsion propre</CP>
  <CP rôle=fonction>et employé pour le transport des marchandises et/ou
      des voyageurs</CP>
</PARAPH>
```

(15) LANDAU (acception **B**)

```
<PARAPH>
  <CC étiq=véhicule>Voiture d'enfant</CC>
  <CP rôle=parties>munie de grandes roues, d'une capote pliante
       surmontant une caisse suspendue garnie d'une literie</CP>
  <CP rôle=fonction>et qui permet de promener un tout jeune enfant, notamment
      en position allongée</CP>
</PARAPH>
```

(16) BOLIDE (acception **B**)

```
<PARAPH>
  <CC étiq=véhicule>Véhicule (avion, automobile)</CC>
  <CP rôle=vitesse>allant à très grande vitesse</CP>
</PARAPH>
```

(17) TACOT (acception **B**)

```
<PARAPH>
  <CC étiq=voiture>Voiture</CC>
  <CP rôle=apparence>démodée</CP>
  <CP rôle=état>en mauvais état (mécanique, carrosserie)</CP>
  <CP rôle=son>qui fait du bruit</CP>
  <CP rôle=vitesse>et n'avance pas</CP>
</PARAPH>
```

3.5 Étape 3 : construction d'une base dérivée

La version entièrement structurée des définitions du TLFi pourra être utilisée comme « fonds lexicographique » à partir duquel sera entrepris un nouveau travail de rédaction d'articles pour le français contemporain, travail visant la production d'une base dérivée formelle exploitable informatiquement. La nomenclature de l'actuel TLFi contient un nombre important de lexies n'appartenant pas ou plus au français courant. Nous ne sélectionnerons, pour la nouvelle base de données, que les sens utilisés en français contemporain, en leur associant systématiquement des phrases d'exemple tirées de corpus de langue usuelle.

Une fois la nouvelle nomenclature mise en place, il nous faudra procéder à l'explicitation de la structure actancielle des lexies prédicatives, condition nécessaire pour une bonne définition de ces unités. Par exemple, nous indiquerons que le nom RANCUNE a trois actants : la personne X qui éprouve ce sentiment, la personne Y qui fait l'objet de ce sentiment et l'événement Z qui a suscité ce sentiment. Il est du reste important que les informations contenues dans une base de données sémantique puissent être mises en parallèle avec des informations morphologiques et syntaxiques. On pourra ainsi indiquer que l'adjectif RANCUNIER caractérise le premier actant du nom RANCUNE, ou encore que le second actant de ce nom peut être introduit par différentes prépositions (*rancune de X **vis-à-vis** de Y, rancune de X **à l'encontre** de Y*, etc.) [17].

L'explicitation de la structure actancielle des lexies servira de support à la réécriture des définitions dans un langage contrôlé. Certaines règles de réécriture pourront être appliquées de manière automatique. Par exemple, les composantes périphériques du type `fonction` (comme dans le cas des définitions de lexies étiquetées `véhicule`) seront systématiquement introduite par le syntagme *servant à* au lieu des différents *destiné à*, *qui sert à*, *spécialement équipé pour*, etc., relevés dans les définitions du TLFi. La plupart des règles de réécriture de définitions nécessiteront toutefois un travail important de la part des lexicographes, secondés par des applications d'aide à l'encodage des données (Barque & Nasr, à paraître). Notons, fi-

17. Voir les descriptions des liens de fonctions lexicales de la Lexicologie Explicative et Combinatoire.

nalement, que les formes lexicales utilisées dans les définitions seront désambiguïsées par le numéro de l'acception auxquelles elles correspondent dans la nouvelle nomenclature.

4 Conclusion

Le développement d'une base de données sémantique dérivée du TLFi est un projet ambitieux qui ne pourra se réaliser qu'au terme de plusieurs années de travail lexicographique. Les premières étapes de sa réalisation, qui seront menées à bien dans un délai beaucoup plus court, donneront déjà des résultats importants que nous énumérons ici.

- Valorisation de l'actuel TLFi : l'ajout d'information concernant la structure des définitions du TLFi rend la base encore plus attractive en permettant aux utilisateurs d'effectuer des requêtes plus fines que celles qu'il est possible d'effectuer actuellement. Par exemple, on pourra connaître l'ensemble des lexies du français dénotant un sentiment, ou encore connaître l'ensemble des lexies du français qui dénotent un véhicule et dont la fonction de ce véhicule est mentionnée dans leur définition, etc.
- Définition d'un format pour les définitions lexicographiques, dans le cadre du développement de lexiques pour le traitement automatique de la langue (TAL). En effet, la nouvelle norme ISO de structuration des données lexicales informatisées, *Lexical Markup Framework* (LMF), a défini un standard d'encodage des informations sémantiques (entre autres) des lexiques destinés au TAL, sans toutefois proposer de structuration **interne** pour les définitions lexicographiques (LMF, 2008) [18]. Ces dernières sont pourtant présentes dans les lexiques les plus utilisés en TAL — notamment WordNet (Fellbaum, 1998) et FrameNet (Fillmore et al., 2003), pour l'anglais — et mériteraient d'être rendues exploitables informatiquement suivant un modèle général standardisé de structuration.
- Une ressource pour la recherche en sémantique lexicale : la hiérarchie des étiquettes sémantiques développée pour le balisage des définitions du TLFi sera probablement complète avant même la fin du processus de structuration des définitions. Elle offrira une description des principaux sens lexicaux du français, organisés en hiérarchie, et indiquera les propriétés lexicograpiquement pertinentes associées à ces sens généraux.

Remerciement

Nous remercions Jean-Marie Pierrel et Pascale Bernard de l'ATILF pour leur aide sur le TLFi. Un grand merci aussi à Claudia Fecteau, Anne-Laure Jousse et Olivier Taïs pour leur implication dans le travail de balisage des définitions. Nous sommes très reconnaissants au réviseur de MTT'09 pour ses commentaires sur la première version de cet article. La recherche présentée ici est financée par des subventions du Fonds de recherche sur la société et la culture du Québec (FQRSC) et du Conseil de recherches en sciences humaines du Canada (CRSH).

Bibliographie

Altman, Joel, et Alain Polguère. 2003. La BDéf : base de définitions dérivée du *Dictionnaire explicatif et combinatoire*. *Proceedings of the First international conference on the Meaning- Text Theory (MTT'2003)*, Paris, 43–54.

Atkins, B. T. Sue and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford, UK.

Barnbrook, Geoff. 2002. *Defining Language. A local grammar of definition sentences*. Benjamins, Amsterdam/Philadelphia.

Barque, Lucie. 2008. *Description et formalisation de la polysémie régulière du français*. Thèse de doctorat, Université Paris 7.

18. Notons qu'il est toutefois possible d'utiliser le cadre formel de LMF pour encoder divers types de structuration des définitions, comme celui de la BDéf (Francopoulo, 2005, pages 19–21).

Barque, Lucie, et Alexis Nasr. à paraître. Un modèle formel de descriptions lexicales : du formalisme BDéf aux structures de traits typées. *Traitement Automatique des Langues (T.A.L.)*, 50(1).

Barque, Lucie, et Alain Polguère. 2009. *Guide des annotateurs pour la structuration des dénitions du TLFi*. OLST, Université de Montréal.
[Accès en ligne : `http://www.olst.umontreal.ca/pdf/guideAnnoTLFi.pdf`]

Benson, Morton, Evelyn Benson, and Robert Ilson. 1986. *Lexicographic Description of English*. Language Companion Series 14, Benjamins, Amsterdam/Philadelphia.

Dendien, Jacques, et Jean-Marie Pierrel. 2003. Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues (T.A.L.)*, 44(2) :11–37.

Dostie, Gaétane, Igor Mel'čuk, et Alain Polguère. 1999. Méthodologie d'élaboration des articles du dictionnaire explicatif et combinatoire du français contemporain. Dans I. Mel'čuk, N. Arbatchewsky-Jumarie, L. Iordanskaja, S. Mantha et A. Polguère (dir.) : *Dictionnaire explicatif et combinatoire du français contemporain : recherches lexico-sémantiques IV*, 11–28. Les Presses de l'Université de Montréal, Montréal.

Fodor, Jerry A., Merrill F. Garrett, Edward C. T. Walker, and Cornelia H. Parkes. 1980. Against definitions. *Cognition*, 8 :263–367.

Fellbaum, Christiane. 1998. *WordNet : An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Fillmore, Charles J., Chris Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet *International Journal of Lexicography*, 16 :235–250

Francopoulo, Gil. 2005. *Extended examples of lexicons using LMF (auxiliary working paper for LMF)*. Rapport technique, INRIA-Loria.

Fontenelle, Thierry 2009. Linguistic research and learners dictionaries : the *Longman Dictionary of Contemporary English*. Dans A. P. Cowie (dir) : *The Oxford History of English Lexicography*, vol. II, 412–435. Oxford University Press, Oxford.

Frassi, Paolo. 2007. *La definizione nel* Trésor de la langue française *: studio tipologico e metalinguistico*. Thèse de doctorat, Università degli Studi di Verona, Vérone.

Kittredge, Richard I. 1982. Variation and homogeneity of sublanguages. Dans R. Kittredge and J. Lehrberger (dir) : *Sublanguage : Studies of Language in Restricted Semantic Domains*, 107–137. de Gruyter, Berlin.

Language resource management — Lexical markup framework (LMF). 2008. ISO/TC 37/SC 4 N453 (N330 Rev.16).
[Accès en ligne : `http://www.lexicalmarkupframwork.org`]

Lexis. Larousse de la langue française. 2002. Sous la direction de Jean Dubois. Larousse/VUEF, Paris.

Mel'čuk, Igor, et al. 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain : recherches lexico-sémantiques I–IV*. Les Presses de l'Université de Montréal, Montréal.

Mel'čuk, Igor, André Clas, et Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Louvain-la-Neuve.

Nouveau Petit Robert. 2007. Sous la direction de Josette Rey-Debove et Alain Rey. Dictionnaires Le Robert, Paris.

Polguère, Alain. 1997. Meaning-Text Semantic Networks as a Formal Language. Dans L. Wanner (dir) : *Recent Trends in Meaning-Text Theory*, 1–24. Language Companion Series 39, Benjamins, Amsterdam/Philadelphia.

Polguère, Alain 2003. Étiquetage sémantique des lexies dans la base de données DiCo. *Traitement Automatique des Langues*, 44(2) :39–68.

Polguère, Alain. 2008. *Lexicologie et sémantique lexicale. Notions fondamentales*. Les Presses de l'Université de Montréal, Montréal.

Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA & London, UK.

Riemer, Nick. 2006. Reductive Paraphrase and Meaning : A Critique of Wierzbickian Semantics. *Linguistics and Philosophy*, 29 : 347–379.

Rundell, Michael. 2008. More Than One Way to Skin a Cat : Why Full-Sentence Definitions Have Not Been Universally Adopted. Dans T. Fontenelle (dir) : *Practical Lexicography. A Reader*, 197–209. Oxford University Press, Oxford, UK.

Sinclair, John M. (dir.). 1990. *Collins Cobuild Student's Dictionary*. Collins, London & Glasgow.

Trésor de la Langue Française informatisé (TLFi). 2004. CNRS Éditions, Paris.
[Accès en ligne : `http://atilf.atilf.fr/tlf.htm`]

Wierzbicka, Anna. 1987. *English Speech Act Verbs. A Semantic Dictionary*. Academic Press, Sydney et al.

Wierzbicka, Anna. 1996. *Semantics : Primes and Universals*. Oxford University Press, Oxford, UK.