À quoi servent les lexiques sémantiques généralistes ? Discussion et proposition

Mathieu Valette[1]
ATILF (CNRS, Nancy)

Abstract

The paper proposes a discussion about non-specialised semantic lexicon. It is made of two parts: first, we'll outline a critical view of existing lexical semantic resources by asking their theoretic background. We also address the issues of their purposes. Second, we'll present a project of semantic lexicon which where epistemological prerequisites cover textual practices. In this project, we focus on infra-lexical and textual description levels rather than addressing lexical and propositional description levels.

Keywords: semantic lexicon, linguistic resources, text semantics, text lexicology

Résumé

Ce papier propose une réflexion sur les lexiques sémantiques généralistes. L'exposé s'articule en deux parties : nous esquissons un parcours critique sur les ressources lexico-sémantiques existantes en questionnant leur arrière-plan théorique et en tentant de répondre à la question de leurs finalités ; puis, en guise d'ouverture, nous présentons un projet de constitution de lexique sémantique dont les présupposés épistémologiques se singularisent par la volonté de prendre en considération les pratiques textuelles d'une part et par le choix des niveaux de description (infralexicaux et textuels, plutôt que lexicaux et propositionnels) d'autre part.

Mots-clés : lexique sémantique, ressources linguistiques, sémantique textuelle, lexicologie textuelle

1. Faire ressources des langues

On peut distinguer plusieurs critères d'évaluation des ressources lexico-sémantiques informatisées. Ces critères apparaissent, en premier lieu, *formels* et liés aux ressources en tant que matériau physique, c'est-à-dire indépendamment de ce pourquoi elles sont constituées. La qualité des ressources est alors affaire de disponibilité, de couverture, de maintenance (pérennité, évolutivité) et de comparabilité[2]. Le respect de ces critères

1 ATILF (CNRS, Nancy), mvalette@atilf.fr.

2 Ces critères nous ont été inspirés par l'argumentaire de la journée d'étude ATALA. « Des ressources sémantiques existantes ? un FrameNet français : (Contre-)arguments, ressources, méthodes et outils », S. Salmon-Alt et G. Pitel (éds.), 13 mai 2006. Cf. aussi le site du CLARIN (Common Language Resources and Technology Infrastructure), http://www.mpi.nl/clarin/.

donne lieu, d'un point de vue scientifique, à des initiatives de normalisation telles que la *Text Encoding Initiative*[3] ou, plus récemment, le *Lexical Markup Framework* (Salmon-Alt 2006)[4]. D'un point de vue académique, on peut évoquer les infrastructures telles que le CLARIN (Common Language Ressources and Technology Infrastucture)[5], ou, dans un contexte francophone, le tout récent CNRTL (Centre National de Ressources Textuelles et Lexicales)[6]. Il existe également des critères *fonctionnels* pour l'évaluation des ressources ; il s'agit en l'occurrence de critères linguistiques. On mesure alors la *qualité lexicographique* des ressources, c'est-à-dire la pertinence de la description supposée être faite d'une unité lexicale. Celle-ci dépend de ce à quoi la ressource est destinée. Ainsi, la finalité d'une ressource linguistique permettrait de faire la distinction entre des critères formels et des critères fonctionnels, les premiers n'étant pas directement conditionnés par l'application ou la gamme d'applications.

On sait qu'en ce qui concerne les textes, la distinction faite entre une *base textuelle* et un *corpus* concède à ce dernier un statut d'objet d'étude. Le corpus est constitué en fonction d'un objectif particulier tandis que la base textuelle a pour principale finalité d'être à large couverture, disponible, pérenne, évolutive, etc. – bref, l'évaluation d'une base textuelle dépend essentiellement des seuls critères formels tandis que le corpus exige une qualité particulière déterminée par la tâche assujettissante. Si l'on projette cette description sur les ressources lexico-sémantiques, on sera tenté d'opposer les lexiques généralistes aux lexiques d'un domaine particulier. Ce serait cependant priver les premiers quasi mécaniquement d'une pertinence linguistique à laquelle les bases textuelles ne peuvent guère prétendre. Mais l'on peut aussi, préjugeant de leur légitimité, se poser la question suivante : à quoi servent les lexiques généralistes ?

Pour répondre à cette question, nous proposerons d'articuler notre exposé en deux parties : nous esquisserons un parcours critique sur les ressources lexico-sémantiques existantes en questionnant leur arrière-plan théorique et en tentant de répondre à la question de leurs finalités ; puis, nous définirons un projet de constitution de lexique sémantique en nous positionnant par rapport aux conclusions de notre analyse de l'existant.

2. Les lexiques sémantiques généralistes : du mot à la phrase

Le domaine des lexiques sémantiques généralistes est bien portant. Il est susceptible de connaître une évolution rapide qui menace d'obsolescence les analyses qui en sont faites. Toutefois, si l'on s'en tient aux ressources existantes, on distingue deux

3 http://www.tei-c.org
4 http://tagmatica.fr/doc > Doc > ISO TC37 documentation.
5 http://www.mpi.nl/clarin/
6 http://www.cnrtl.fr

approches du sens présidant à leur constitution. Circonscrivons-les brièvement avant d'aborder leur utilité.

2.1. Les approches paradigmatiques

D'inspiration philosophique et terminologique, et principalement représentées par les thésaurus et les ontologies telles que WordNet (Fellbaum 1998), les approches qui nous qualifierons ici de paradigmatiques proposent une représentation close du monde ou du domaine, où la signification des items dépend de relations hiérarchiques (hyperonymie, hyponymie, etc.) construites en fonction des référents qu'ils désignent. Centrées sur la référence et non sur les usages en discours, les ontologies relèvent peut-être autant, sinon davantage de la philosophie que de la linguistique. Plus légitimes dans la perspective terminologique des ontologies de domaine, elles apparaissent insuffisantes pour rendre compte du sens des unités lexicales dès lors qu'elles sont actualisées dans des textes où l'ambiguïté est licite (textes littéraires par exemple), à l'inverse, des textes techniques, notamment, où un terme désigne un concept ou un objet en principe sans équivocité.

2.2. Les approches syntagmatiques

D'inspiration logico-grammaticale, les approches syntagmatiques sont fondées sur le comportement cotextuel du lexique et donnent une large place à la syntaxe. FrameNet (Fillmore, Baker & Sato 2002) en est un excellent représentant. Les unités lexicales y sont livrées accompagnées d'une notice d'actualisation décrivant la combinatoire syntaxique (arguments) et sémantique (actants) de leurs différentes acceptions. Les approches syntagmatiques apparaissent plus pertinentes, linguistiquement parlant, que les approches paradigmatiques parce qu'elles tiennent compte de l'énonciation. En d'autres termes, elles ont pour objet le sens de l'unité lexicale dans l'énoncé plutôt que sa référence. Elles reposent toutefois sur une vision grammaticale du sens, où le syntagme et la phrase (dont le parangon est la proposition logique) constituent les seules unités prises en compte, et adaptent les acteurs traditionnels de la sémantique de l'énoncé (thème/rhème, agent/patient, etc.). Or, si la signification d'une unité lexicale peut sans encombre être rapportée au lexique, son sens dépend, bien au-delà de la phrase, du texte dans son unité et du corpus dont celui-ci dépend, autrement dit, des *usages* socialement codifiés et linguistiquement organisés en discours ; cf. (Rastier 2001).

– À noter qu'un lexique tel que le DiCo (Jousse & Polguère 2005) opte explicitement pour une approche mixte, à la fois paradigmatique (dérivation sémantique) et syntagmatique (consignation des principales collocations). S'il présente une qualité lexicographique indéniable, son objet demeure circonscrit aux mêmes unités moyennes que la plupart des lexiques d'inspiration logico-grammaticale : mot, syntagme, phrase. À la différence de WordNet ou de FrameNet, sa couverture est, pour

l'heure, encore insuffisante pour répondre exhaustivement aux critères formels signalés précédemment.

3. À quoi servent les lexiques sémantiques généralistes ?

Quoiqu'elle puisse paraître saugrenue, la question mérite d'être posée dans la mesure où la légitimité des lexiques sémantiques est logiquement soumise à leurs usages, et que ceux-ci ne sont pas toujours faciles à appréhender. Faisant figure d'exception de par sa notoriété, WordNet est relativement à l'abri des critiques concernant sa finalité (c'est bien utile, lit-on souvent ; – comprendre : on n'a guère d'alternative). Sa transposition EuroWordNet, quoique plus sophistiquée, est cependant contestée, tant d'un point de vue fonctionnel, c'est-à-dire linguistique (Slodzian 1999) que formel (ni pérenne, ni évolutive pour des raisons techniques, la ressource échouerait à l'examen du critère de maintenance). Dès lors, les projets de développement ou de transposition des ressources (anglo-saxonnes) en d'autres langues comme le français, tels que FR.FrameNet[7], où Fillmore et l'université de Berkeley sont impliqués, posent des problèmes non triviaux d'objectifs. Distinguons trois usages possibles : l'équipement d'une théorie, la lexicographie, les applications en TAL.

3.1. L'équipement d'une théorie

Selon toute vraisemblance, la réalisation d'un lexique à large couverture inspirée d'une théorie vise, *ab initio*, à décrire systématiquement et formellement le lexique d'une langue pour équiper la théorie d'une ressource idoine. C'est le cas par exemple de FrameNet et de la sémantique des cadres de (Fillmore 1976), du DiCo et de la théorie Sens-Texte de (Mel'čuk, Clas & Polguère 1995), des classes d'objets et du Lexique-Grammaire construits par, ou à la suite des travaux de (Gross 1975, 2005). Au-delà du truisme apparent, la réalisation d'un lexique à large couverture est une entreprise passablement ambitieuse qui n'est pas forcément initiée par les « usagers » des ressources, mais plutôt par les promoteurs des théories. Les usagers, en général, ont des besoins ponctuels ; soit ils empruntent les ressources existantes et s'en satisfont peu ou prou, soit ils en produisent eux-mêmes des lacunaires et néanmoins suffisantes. En fait, qu'une application particulière mobilise un lexique « général » apparaît difficilement soutenable[8] : une tâche relève d'un domaine et d'un usage déterminé auquel correspond un discours spécifique. N'est alors requis que le seul lexique particulier dudit discours. Se restreindre au seul lexique pertinent pour la tâche permet d'éviter l'opération de désambiguïsation qu'implique l'usage d'un lexique général (Rastier & Valette, à paraître).

7 http://libresource.inria.fr/projects/framenet/

8 À moins bien entendu qu'il ne s'agisse d'une application lexicographique réflexive.

3.2. Les projets lexicographiques

Si les ressources lexico-sémantiques généralistes sont des objets lexicographiques, elles prétendent à une forme de prototypicité susceptible d'être dérivée en d'autres projets lexicographiques. FrameNet, par exemple, propose aux lexicographes et à toutes fins utiles une caractérisation des propriétés grammaticales et combinatoires des entrées. Le DiCo est, quant à lui, présenté comme un cadre pour dériver des dictionnaires à l'usage du grand public tels que le Lexique Actif du Français (Polguère 2000). En somme, à l'instar de bon nombre des disciplines des sciences du langage, la lexicographie réalise des modèles formalisés et leur attribue conséquemment un intrinsèque potentiel générateur.

3.3. Les applications TAL

Les applications TAL constituent le meilleur débouché pour les ressources lexico-sémantiques. Toutefois, là encore, à l'exception notable de WordNet, on observe un certain fossé entre l'effectif et le volitif. Ainsi, on lit dans (Jousse et Polguère, 2005) qu'« on doit pouvoir produire automatiquement à partir du DiCo (par compilation de celui-ci) des lexiques de systèmes de traitement automatique de la langue (TAL) » mais pour l'heure, la plupart des travaux réalisés demeurant à l'état expérimental. Si l'on questionne FrameNet, une ressource plus avancée dans sa constitution[9], sur ses usages en TAL, la FAQ du site officiel répond par la liste que nous reproduisons *in extenso* ci-dessous[10] :

> *word sense disambiguation:* FrameNet gives the syntactic and collocational information that in many cases is likely to be uniquely associated with a single sense;
>
> *machine translation:* FrameNets for other languages will reveal cross-linguistic differences in the meanings and grammatical behavior of words belonging to shared frames;
>
> *information extraction:* FrameNet annotations provide a high precision seed for building information extraction patterns;
>
> *question answering:* FrameNet data will facilitate the recognition of semantic frame relations between the language of user questions and that of passages in the documents that answer the question.

Si l'on s'en tient ici aux seules modalités énonciatives, on est bien en peine de discerner ce qui ressortit à l'avéré et ce qui relève du vœu. Deux tâches tendent à première vue vers l'effectif, en premier lieu la désambiguïsation, puis la recherche d'information qui techniquement lui succède, tandis que les deux autres, la traduction

9 En septembre 2006, FrameNet comptait 8900 entrées, le DiCo 588.

10 http://framenet.icsi.berkeley.edu > FAQs > "Who is using the database?".

automatique et les systèmes de question/réponse, sont présentées sur un mode plus intentionnel.

Concentrons-nous sur la désambiguïsation, véritable manne du TAL lexico-sémantique, que la production de lexiques généralistes n'épuise pas, loin s'en faut, puisqu'en subsumant une multitude d'acceptions ou d'emplois sous une seule entrée, ils génèrent eux-mêmes des polysémies et requièrent, par conséquent, que les textes soient désambiguïsés à leur contact.

3.4. Un mot, un sens

Si la polysémie, observable en langue, bénéficie en effet d'une certaine bienveillance, celle-ci est a peu près égale à l'embarras dans lequel l'ambiguïté lexicale, son pendant discursif, plonge les linguistes. Qu'une unité lexicale ait plusieurs acceptions constitue en effet pour certains une « qualité importante » (François, Manguin & Victorri 2003 : 4) et pour d'autre une « source de richesse et de souplesse dans les langues » (Venant 2004 : 1145). En revanche, on admet moins facilement qu'un texte fasse l'objet de plusieurs interprétations ; et sitôt passée en discours, la polysémie, devenue ambiguïté, pose inlassablement « problème »[11]. Ainsi, de façon assez cocasse, on regrette que les occurrences de discours, multiples mais singulières par définition, soient tissées d'équivoques mais on s'émerveille de la polysémie dans la langue, pourtant réputée stable et conservatoire par nature. À la langue, le terme complexe savamment composé de formants grecs ; au discours, les appellations multiples, les mots vulgaires de basse extraction latine, à consonance péjorative[12].

En somme, en faisant l'hypothèse implicite qu'il existe une structure profonde (hiérarchique ou relationnelle) permettant de préciser les significations, les promoteurs des lexiques sémantiques actuels cherchent à atteindre la permanence référentielle et l'univocité. D'un point de vue linguistique, cette position réduit drastiquement le champ des possibles sémantiques, et donne à quelques niveaux d'analyse (le mot, le syntagme, la phrase) une suprématie sémantique discutable.

4. Contexte, texte, intertexte ; tâches et usages

11 Le mot problème apparaît en effet fréquemment dès qu'il est question de désambiguïsation. Par exemple « Pourtant dès que l'on veut automatiser une telle performance, la polysémie devient un véritable problème et elle donne bien du souci aux chercheurs en traitement automatique du langage. » (Venant, *op. cit*).

12 Avant l'invention du mot *polysémie*, on recourrait notamment aux mots *équivoque* et *louche*. À la cour, l'équivocité était évidemment à bannir ; cf. Cl. Favre de Vaugelas, *Remarques sur la langue françoise vtiles à ceux qui veulent bien parler et escrire*, Paris, P. le Petit et la Veuve Camusat, 164 – cité au séminaire PRAXILING (CNRS, Montpellier) du 20 mars 2006 par W. Ayres-Bennett.

Pour réduire la polysémie générée par les lexiques généralistes et établir le sens d'un mot, il est d'usage d'interroger le contexte, c'est-à-dire d'entourer d'un peu de texte le mot qui en a été privé en accédant au statut d'entrée du dictionnaire. Mais le linguiste arpente le sens à la toise de son objet ; le contexte s'évalue différemment selon que l'on pratique une linguistique du mot, de la phrase ou de quelque unité que ce soit.

4.1. Les incolores idées vertes ont la peau dure…

Aux questions (fréquemment posées), « Your lexicon is based on sentences taken one at a time. Why don't you look at longer texts? Also, how can you tell whether an isolated sentence like *Make sure he's dead* is a request to confirm someone's state of health or a command to kill him? », l'équipe de FrameNet répond :

> First of all, we would have to say that the phrasal entry *make sure* belongs to two frames, one related to *find out* and the other to *ensure*. For our purposes it would be important to know that the imperative sentence in the question supports both meanings (which is why it's ambiguous), but it is not our job to know which sense it had in the corpus where we found it.
>
> For purely lexicographic purposes – or rather for the purposes of describing the intrasentential combinatoric properties of lexical items – we see no need to examine structures larger than the sentence. Our main use of the corpus is as evidence for lexicographic description; so far we have not found any words where we cannot find enough clear examples for illustrating all of the word's possibilities.
>
> In fact, we *have* recently begun annotating some continuous texts, and in that case, we *do* pay careful attention to context to determine which LUs (word senses) appear.[13]

Bien qu'elle puisse étonner au vu des débouchés revendiqués précédemment (cf. 3.3), la déclaration d'incompétence (« it is not our job to know which sense it had in the corpus where we found it ») est franche et montre à quel point la réalisation de FrameNet semble conditionnée par la théorie qui la sous-tend davantage que par une tâche précise (cf. 3.1). FrameNet n'accorde aucune place théorique au contexte. Celui-ci acquerra, à la rigueur, quelques lettres de noblesse empiriques.

Pourtant, les questions conjointes de la dimension, de l'hétérogénéité et de la nature du cotexte (i.e. du contexte *linguistique*) semblent cruciales et ont été passablement étudiées, comme l'atteste la célèbre synthèse d'(Ide & Véronis 1998) et plus récemment (Crestan, El-Bèze & de Loupy 2003). Mais si déterminer une limite inférieure à la dimension d'une fenêtre cotextuelle semble possible, les limites supérieures sont plus délicates à établir. Tout se passe comme si, à mesure que le cotexte s'élargit, il se rapproche du contexte, c'est-à-dire du contexte *extralinguistique*, lequel peut être, dans certains cas, assimilé à l'usage socialement déterminé.

---

13 http://framenet.icsi.berkeley.edu > FAQs > "Your lexicon is based on sentences taken one at a time. Why don't you look at longer texts?"

Situé aux confins du linguistique et de l'extralinguistique, le domaine thématique mérite toute notre attention. Relatons ici ce qu'en ont dit (Ide & Véronis 1998 : 21-22). Après avoir établi diverses qualités de cotexte, ils rapportent plusieurs expériences relatives à la désambiguïsation contextuelle par le domaine thématique. Précisons qu'elles se concluent souvent par des échecs, bien qu'elles reposent sur l'hypothèse *a priori* cohérente qu'il n'y a pas de polysémie dans un domaine donné, et qu'elles fassent appel à des microglossaires. Ide et Véronis font notamment état de la théorie des scripts de (Schank & Abelson 1977) qui consiste à identifier dans un texte des scénarii préétablis pour choisir les emplois correspondant à la pratique ainsi formalisée (par exemple, au restaurant : entrer, commander, manger, sortir). Mais ces scripts sont des modèles conceptuels d'inspiration cognitiviste et ne sont pas hérités de l'analyse de corpus. En conséquence, ils ne relèvent pas du cotexte à proprement parler mais d'un *préjugé sur le contexte* (ça se passe comme ça) *projeté sur le cotexte* (cela doit s'énoncer ainsi).

Par ailleurs, (Ide & Véronis 1998), pour caractériser les recherches en désambiguïsation à partir du domaine thématique, juxtaposent, c'est-à-dire ne distinguent pas formellement, des approches IA telles que la prescription du domaine évoquée ci-dessus, des approches lexicologiques telles que la distance prototype / emplois et des approches davantage textuelles mais non identifiées comme telles (« The influence of domain likely depends on factors such as the type of text (how technical the text is, etc.) » (*op. cit.*). En définitive, les travaux que les auteurs ont présentés sont dominés par une vision lexicale sinon dictionnairique du sens, même lorsqu'il s'agit d'élargir le cotexte.

Ailleurs, (Kleiber 1997 : 23), qui s'essaie également à la métrologie, n'envisage pas de dépasser la « totalité du texte » comme limite supérieure. On peine à interpréter son sentiment lorsqu'il affirme : « Si on accepte l'idée d'intertexte, on est même conduit comme le note (Kerbrat-Orecchioni 1996) à une étendue discursive théoriquement illimitée. ». Car enfin, seul l'inexploré est illimité et cette étendue discursive paraît en fait constructible. On peut notamment l'organiser en discours et en genre, et en dégager l'*intertexte*, qui donne corps aux informations discursives et domaniales.

4.2. … Mais les fruits frais brûlent mal

Adoptons le point de vue de la sémantique textuelle selon lequel le sens d'un mot dépend, au-delà du contexte, du texte dans lequel il est actualisé, ledit texte étant déterminé par la pratique sociale qui a présidé à sa réalisation (principe de détermination du local par le global, cf. Rastier 2001). Cette position, outre qu'elle alloue à l'usage un rôle décisif, tend à minimiser sensiblement le risque d'ambiguïté sémantique dans les traitements interprétatifs, dans la mesure où les pratiques conditionnent des discours et des genres normés. Soit, le premier quatrain de *Bohémiens en voyage*, de Ch. Baudelaire :

La tribu prophétique aux prunelles ardentes
Hier s'est mise en route, emportant ses petits
Sur son dos, ou livrant à leurs fiers appétits
Le trésor toujours prêt des mamelles pendantes.

Bien évidemment, l'ambiguïté de *prunelles*, à la fois « fruit du prunellier » et, par analogie, « pupille de l'œil » ou métonymiquement « yeux, regard »[14], est un pur effet de mitoyenneté lexicographique. Si des machines TAL ne s'étonneront peut-être pas de la combustion de petits fruits globuleux par des groupes sociaux annonciateurs d'événements futurs, l'étude des pratiques littéraires nous apprendra que ce qui, en fait, est *ardent*, dans la poésie du 19ème siècle, est, par ordre d'importance, l'*œil*, le *soleil*, le *feu* et le *coeur*, nullement les petites prunes[15]. De fait, ce n'est pas tant le cotexte que le texte lui-même, en tant qu'il relève d'un discours (littéraire), d'un champ générique (la poésie) voire d'un genre (le sonnet italien) et d'une période (le 19ème siècle) qui permet de désambiguïser le mot[16].

4.3. Sources ou ressources ?

La disjonction persistante opérée par la lexicographie informatique entre le lexique et le texte menace d'être dommageable sinon contreproductive. Elle témoigne d'une forme de *ressourcisme* (constituer les ressources pour les ressources) où les critères formels énoncés précédemment dominent largement. Certes le ressourcisme apparaît académiquement encouragé – *tempori servire* – mais sa légitimité doit néanmoins être discutée. Consulter une discipline ressourciste par excellence, en l'occurrence, la terminologie, est, à cet égard, très instructif. Loin des ardeurs théoriciennes, la récente terminologie textuelle a en effet su faire montre de pragmatisme en soumettant la constitution de ressources aux *sources*, c'est-à-dire aux textes. Les inventeurs de la terminologie textuelle expliquent qu'« étant donné un domaine d'activité, il n'y a pas *une* terminologie, qui représenterait *le* savoir sur le domaine, mais autant de terminologies que d'applications dans lesquelles ces terminologies ont été utilisées » (Bourigault & Slodzian 2000 : 30). D'une certaine façon, cette proposition oppose aux

14 D'après le *Trésor de la Langue Française*.

15 Étude des voisinages de la forme *ardent* dans la base textuelle FRANTEXT, textes non catégorisés. Fonction Voisinage de mot +/10 dans un regroupement de textes, genre « poésie » 1800-1900. Nombre d'occurrences du pivot : 231. Taille des voisinages explorés : 4620 occurrences. Nombre de graphies trouvées dans ces voisinages : 1278. À l'inverse, si nous nous interrogeons sur l'environnement cooccurrentiel de *prunelles* dans les traités du 20ème siècle, on y rencontrera *alises*, *calvilles*, *châtaignes*, *cormes*, *figues*, *nèfles*, *noix*, *noisettes*, *pêches*, *poires*, *pommes*, *raisins*, lesquels constituent quelques individus d'une classe sémantique des fruits comestibles qui n'est pas convoquée par *prunelles* dans la poésie du 19ème siècle. Étude des voisinages de la forme *prunelles*, dans la base textuelle FRANTEXT, textes non catégorisés. Voisinage de mot +/10 dans un regroupement de textes, genre « traité » 1918-2000. Nombre d'occurrences du pivot : 5 Taille des voisinages explorés : 241 occurrences. Nombre de graphies trouvées dans ces voisinages : 86.

16 On lira bientôt à propos de la désambiguïsation automatique des textes littéraires un Becker & Valette actuellement en préparation.

lexiques généralistes un principe de réalité textuelle qui dépasse largement les fenêtres contextuelles, quelle que soit leur largeur.

Se réclamant de l'héritage des linguistes terminologues, on proposera de maintenir soudé le rapport lexique-corpus. Ce choix de principe, outre qu'il implique de constituer les corpus selon les exigences de la linguistique de corpus[17], a une incidence notoire sur ce que l'on entendra par lexique. En effet, les présupposés théoriques de la sémantique lexicale sont étrangers à toute *lexicologie textuelle* ; c'est donc vers la sémantique textuelle (Rastier 1989, 2001) que nous nous tournerons pour définir la matière lexicale qui selon nous, peut constituer un lexique sémantique alternatif.

5. Ni mot, ni phrase, mais des traits sémantiques et des textes

Nous évoquerons ici la réalisation en cours d'un lexique sémantique alternatif aux approches décrites précédemment, élaboré à partir des critiques que nous avons formulées à cette occasion. Du point de vue paradigmatique, nous proposons d'envisager les relations entre unités lexicales non pas en termes de construction hiérarchique mais en termes de *classes sémantiques* dont la cohésion est assurée par des descripteurs infralexicaux (que nous qualifions de sèmes). Du point de vue syntagmatique, l'instanciation des unités lexicales dans les textes ne se fait pas au niveau de la phrase ou de l'énoncé, mais au niveau d'unités textuelles dont la taille dépend davantage des applications que de choix théoriques édictés en amont. Le cadre théorique choisi, la sémantique textuelle, induit un certain nombre de propositions liminaires :

a. Le contenu sémantique (sémème ou sémie) d'une unité lexicale est constitué de traits sémantiques (sèmes).

b. Au sein d'un sémème, les sèmes sont organisés et pondérés en fonction des domaines, des discours et des genres textuels des différents corpus dans lesquels il est actualisé, autrement dit, des usages possibles. Ainsi, à une unité lexicale peut correspondre plusieurs sémèmes suivant le corpus dont elle est issue.

c. Les différents sémèmes sont obtenus par apprentissage statistique. Les sèmes sont qualifiés en fonction de leur participation à des *réseaux sémiques intratextuels* dont on observe les régularités sur des corpus homogènes (en genre, discours et domaine). On appelle ces réseaux sémiques des *fonds* et des

17 Sans entrer dans une discussion qui dépasserait notre propos ici, signalons à ce sujet les remarquables travaux définitoires de (Bommier-Pincemin 1999 : 415-427) et (Rastier 2005).

formes sémantiques. L'empan de ceux-ci varie suivant leur nature et celle du texte.

La ressource lexico-sémantique en cours de développement s'apparente donc à un dictionnaire sémique, ou plus précisément, à *une collection de dictionnaires sémiques* relevant de domaines, de genres et de discours particuliers. La ressource, que nous avons nommé DIXEM, est conçue en premier lieu à des fins analytiques : notre objectif est d'étudier l'économie générale du contenu sémantique des unités lexicales dans des corpus de textes structurés, dans une perspective tant diachronique que synchronique. En termes d'application, notre ressource lexico-sémantique se prêtera à l'analyse thématique. Cette recherche s'inscrit donc dans la continuité des travaux actuels sur l'analyse thématique (Rastier éd. 1995), (Bourion, 2001), (Zweigenbaum & Habert, 2004), (Valette 2004), (Loiseau 2005) mais dans une perspective infralexicale.

5.1. Les traits infralexicaux

Les expériences rapportées dans (Valette, Estacio-Moreno, Petitjean & Jacquey 2006) ont donc eu pour objectif d'explorer d'un point de vue paradigmatique un dictionnaire de traits sémantiques infralexicaux de manière à en évaluer la pertinence fonctionnelle. Pour la réalisation d'une telle ressource, nous avons exploité un dictionnaire de langue informatisé : le *Trésor de la Langue Française* (Dendien & Pierrel 2003), désormais *TLF*. Ce dictionnaire est doté de 100 000 mots et de 270 000 définitions. Le corpus fut composé suivant une hypothèse posturale minimaliste dont nous ne discuterons pas l'hétérodoxie ici : une définition est un sémème mis en texte. Ainsi, les mots pleins d'une définition (substantifs, adjectifs, verbes et certains adverbes) sont, une fois lemmatisés, les sèmes potentiels qui constituent le sémème d'une unité lexicale en attente d'actualisation. Par exemple, pour une définition telle que :

> LAURACÉES. Famille de plantes dicotylédones, comprenant des arbres et des arbustes, à feuilles simples, alternes et persistantes, qui croissent dans les régions chaudes et tempérées.

Nous extrayons le sémème suivant :

> LAURACÉES {/famille/, /plante/, /dicotylédone/, /comprendre/, /arbre/, /arbuste/, /feuille/, /simple/, /alterne/, /persistant/, /croître/, /région/, /chaud/, /tempéré/}

Nos premiers travaux ont porté sur la constitution de classes sémantiques au sein d'un domaine donné et sur la structuration des sémèmes en fonction des classes obtenues. Ils ont montré que la réalisation de classes sémantiques à partir des définitions dictionnairiques était possible, même si elle ne pouvait se substituer à un apprentissage sur corpus. Une étude portant sur 588 sémèmes incluant le sème /arbre/ nous a en effet permis de distinguer, par Classification Ascendante Hiérarchique (CAH), des sous-classes pertinentes, tant d'un point de vue gnosique (Essences d'arbre, Parasites) que

praxique (Plantation, Bûcheronnage, Arboriculture) sans que de telles classes n'aient été dessinées *a priori*, ni par nous, ni par les lexicographes du *TLF*.

Par ailleurs, nous avons procédé à une pondération interne des sémèmes à l'aide d'un calcul d'écart réduit à l'intérieur d'une classe comportant 105 définitions. Cette pondération a donné à voir une organisation sémantiquement pertinente où les traits spécifiques sont susceptibles d'être opérationnels, notamment dans la perspective thématique/textuelle qui est la nôtre. Par exemple, les sèmes */jardin/*, */avenue/* et */ornement/* ont été distingués comme les plus caractérisants de l'entrée *« sophora »* (« *Arbre exotique de la famille des Légumineuses, servant à l'ornement des jardins et des avenues* ») au sein de la classe des Essences d'arbre. En d'autres termes, le *sens* est apparu valorisé au détriment de la *référence* (*/arbre exotique/*, */légumineuse/*).

5.2. Les réseaux intratextuels

Nous faisons l'hypothèse que les sèmes « spécifiques » participeront de façon privilégiée à des réseaux de traits sémantiques parcourant un texte. L'apprentissage sur corpus fera donc apparaître la régularité de ces réseaux de sèmes et permettra de stabiliser les sémèmes (en pondérant les sèmes : activation des traits spécifiques, inhibition des traits peu spécifiques ou du bruit), voire de les réorganiser, non plus dans une perspective paradigmatique mais en tenant compte des usages des unités lexicales dans les corpus de textes préalablement constitués. En effet, les dictionnaires, en décontextualisant les mots, en donnent une définition typique et consensuelle qui ne correspond pas nécessairement aux instanciations. Bien que reposant sur une méthodologie scientifique inédite à l'époque, le *TLF* demeure œuvre de lexicographie humaine. Malgré une sensibilité reconnue pour les usages (pas moins de 5000 domaines composés sont référencés) et l'utilisation systématique de corpus de concordances, il n'échappe pas aux préjugés et aux (auto)censures rédactionnelles.

À titre d'exemple, nous avons mis en italique, dans le court texte suivant extrait de *Bouvard et Pécuchet* de G. Flaubert, tous les mots étiquetés dans le *TLF* comme relevant du domaine de la Botanique et nous avons signalé par un exposant les mots dont le sémème comprend le sème /ornemental/, toujours d'après le *TLF*.

> Alors Pécuchet se tourna vers les *fleurs*$^{/o/}$. Il écrivit à Dumouchel pour avoir des *arbustes* avec des *graines*, acheta une provision de terre de bruyère et se mit à l´oeuvre résolument.
>
> Mais il planta des *passiflores* à l´ombre, des *pensées* au soleil, couvrit de fumier les *jacinthes*, arrosa les *lys*$^{/o/}$ après leur floraison, détruisit les *rhododendrons*$^{/o/}$ par des excès d´abattage, stimula les *fuchsias*$^{/o/}$ avec de la colle forte, et rôtit un grenadier, en l´exposant au feu dans la cuisine.
>
> Aux approches du froid, il abrita les *églantiers* sous des dômes de papier fort enduits de chandelle ; cela faisait comme des pains de sucre, tenus en l´air par des bâtons. Les tuteurs des *dahlias*$^{/o/}$ étaient gigantesques ; – et on apercevait, entre ces lignes droites les

> rameaux tortueux d´un *sophora*/o/-japonica qui demeurait immuable, sans dépérir, ni sans pousser.

Cette séquence est d'une homogénéité évidemment exemplaire. Douze fois l'étiquette domaniale Botanique est instanciée, et six fois le sème /ornemental/, en particulier lors des séquences *lys – rhododendron – fuchsias* et *dahlias – sophora*. On peut, de ce fait, penser que le sémème d'*églantier*, pris entre les deux séquences, qui partage avec celles-ci le domaine Botanique, est susceptible d'hériter du sème /ornemental/ dont il est dépourvu. De fait, l'églantier peut être utilisé dans une perspective ornementale. Nous faisons donc l'hypothèse que si la cooccurrence du sème /ornemental/ et du sémème *églantier* est statistiquement significative sur un corpus homogène, ledit sémème peut accueillir ce nouveau sème. Cette hypothèse n'a pas été vérifiée, ce travail est en cours.

6. Bilan

À la question, à quoi servent les lexiques sémantiques généralistes ?, nous avons répondu en esquissant une alternative qui n'est généraliste que dans la mise en œuvre, non dans la ressource proprement dite. Du point de vue de la mise en œuvre, il s'agit bel et bien de définir un protocole *général* de production de ressource lexico-sémantique complexe susceptible de couvrir plusieurs discours, pourvu qu'il en existe des archives convertibles en corpus. La ressource quant à elle ne sera pas une et universelle : nous aurons une *collection* de dictionnaires sémiques *comparables* c'est-à-dire plusieurs lexiques profilés automatiquement à partir de corpus, par apprentissage statistique, en fonction des applications. Par exemple, si la ressource n'a pas été entraînée sur des corpus de textes juridiques, elle sera inappropriée pour des applications ressortissant à cet usage.

Ce projet ne relève pas d'un travail de lexicographe mais d'un travail de *corpiste*, pour reprendre la plaisante expression forgée par Habert : le soin que le lexicographe apporte au lexique, nous l'apporterons pour notre part à la constitution du corpus, laissant aux algorithmes d'apprentissage la tâche d'en extraire la matière sémantique. En cela, la lexicologie textuelle relève peut-être davantage d'une linguistique des textes que d'une linguistique du mot.

– Quant aux critères de qualité formelle énoncés dans le premier paragraphe (disponibilité, couverture, maintenance et comparabilité), on comprendra que nous les excluons globalement de notre problématique ; mais pour y satisfaire néanmoins, le lexique sémantique DIXEM est développé et géré en partenariat étroit avec le Centre National de Ressources Textuelles et Lexicales (CNRTL).

Je tiens à exprimer toute ma reconnaissance à Evelyne Jacquey et François Rastier pour leurs lectures de cet article, leurs observations et leurs suggestions.

Références

BOMMIER-PINCEMIN B. (1999), *Diffusion ciblée automatique d'informations : conception et mise en oeuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat en Linguistique, Université Paris IV Sorbonne, 6 avril 1999.

BOURIGAULT D., SLODZIAN M. (2000), « Pour une terminologie textuelle », in *Terminologies Nouvelles*, n° 19 : 29-32.

BOURION É (2001), *L'aide à l'interprétation des textes électroniques*. Thèse de doctorat, Université Nancy 2 ; publié sur *Texto ! Textes et cultures* (http://www.revue-texto.fr).

CRESTAN É, EL-BEZE M., DE LOUPY CL. (2003) « Peut-on trouver la taille de contexte optimale en désambiguïsation sémantique? », in *TALN 2003*, Batz-sur-Mer, 11-14 juin 2003, http://www.atala.org/doc/actes_taln/AC_0085.pdf.

DENDIEN J., PIERREL J.-M. (2003), « Le trésor de la langue française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence » in *Traitement Automatique des Langues, TAL*, 44 n°2 : 11-37.

GREIMAS A. J (1966), *Sémantique structurale*. Paris : PUF.

FELLBAUM C. (1998), *WordNet: an Electronic Lexical Database*, Cambridge MA, MIT.

FILLMORE CH. J. (1976) "Frame semantics and the nature of language", in Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, Volume 280: 20-32.

FILLMORE CH. J., BAKER C. F., SATO, H. (2002) "The FrameNet Database and Software Tools", in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas: 1157-1160.

FRANÇOIS J., MANGUIN J.-L., VICTORRI B. (2003), « La réduction de la polysémie adjectivale en contexte nominal : une méthode de sémantique calculatoire », in *Cahier du CRISCO*, 14 : 1-43.

GROSS G. (1975), *Méthode en syntaxe*, Hermann, Paris.

GROSS G. (2005), « Un dictionnaire électronique des adjectifs du français ». *Cahiers de Lexicologie*, n°86 : 11-33.

IDE N., VERONIS J. (1998), « Word Sense Disambiguation: The State of the Art", *Computational Linguistics*, 24/1: 1-40.

JOUSSE A.-L., POLGUÈRE A. (2005), *Le DiCo et sa version Dicouèbe, Document descriptif et manuel d'utilisation*, Version du rapport 1.0 – 19 avril 2005, Observatoire de linguistique Sens-Texte (OLST), Université de Montréal.

KERBRAT-ORECCHIONI C. (1996), « Texte et contexte » *in* Schmoll, P. (éd.) "Contexte(s)", *Scolia* 6 : 39-60.

KLEIBER G. (1997), « Sens, référence et existence : que faire de l'extra-linguistique ? », *Langages*, Vol.127 : 9-37.

LOISEAU S. (2005), « Thématique et sémantique contextuelle d'un concept philosophique », in *La Linguistique de corpus. Actes des deuxièmes journées de la linguistique de corpus,* G. Williams (éd.), Rennes, PUR : 129-140.

MEL'CUK I., CLAS, A., POLGUÈRE A. (1995), *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve, Duculot.

POLGUÈRE A. (2000), "Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French". *Proceedings of EURALEX'2000*, Stuttgart : 517-527.

RASTIER F. (1989), *Sens et textualité*. Paris : Hachette.

RASTIER F. éd. (1995), *L'Analyse thématique des données textuelles : l'exemple des sentiments*. Paris : Didier.

RASTIER F. (2001), *Arts et sciences du texte*. Paris : PUF.

RASTIER F. (2005), « Enjeux épistémologiques de la linguistique de corpus », *in La linguistique de corpus*, G. Williams (éd.), Rennes, PUR : 31-45 ; publié sur *Texto ! Textes et cultures* (http://www.revue-texto.fr).

RASTIER F., VALETTE M. (à paraître), « De la polysémie à la néosémie », *in Langue française*, P. Siblot (éd.).

SALMON-ALT S. (2006), « *V[1]Ω a=able* ou Normaliser des lexiques syntaxiques est délectable ». *in* in *Verbum ex machina, Actes de la 13ème conférence sur le traitement automatique des langues naturelles (TALN 06)*. P. Mertens, C. Fairon, A. Dister, P. Watrin (éds). *Cahier du CENTAL*, 2.1, UCL Presses Universitaires de Louvain, Volume 1 : 297-306.

SCHANK R. C. & ABELSON R. P. (1977), *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum, Hillsdale, New Jersey.

SLODZIAN M. (1999), WordNet et EuroWordNet – Questions impertinentes sur leur pertinence linguistique. *Sémiotique*s, n°17, 51-70.

VALETTE M. (2004), « « Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet » in *Approches sémantiques du document numérique. Actes du 7ème colloque international sur le document électronique : CIDE.7*, P. Enjalbert, M. Gaio (éds.) : 215-230.

VALETTE M., ESTACIO-MORENO A., PETITJEAN É., JACQUEY É. (2006), « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens », in *Verbum ex machina, Actes de la 13ème conférence sur le traitement automatique des langues naturelles (TALN 06)*. P. Mertens, C. Fairon, A. Dister, P. Watrin (éds). *Cahier du CENTAL*, 2.1, UCL Presses Universitaires de Louvain. Volume 1 : 357-366.

VENANT F. (2004), « Polysémie et calcul du sens », in *Le poids des mots, Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 10-12 mars 2004, Louvain-la-Neuve (Belgique), G. Purnelle, C. Fairon, A. Dister, (éds.), UCL-Presses Universitaires de Louvain : 1146-1157.

ZWEIGENBAUM P., HABERT B. (2004), « Accès mesurés aux sens », in *Mots. Les langages du politique*, n°74 : 93-106.