Dictionnaires, théorie des graphes et structures lexicales

Sylvain Loiseau, Philippe Gréa, Jean-Philippe Magué

Dictionnaires, théorie des graphes et structures lexicales

Sylvain Loiseau

UMR 7114 Modyco – CNRS/Université Paris Ouest Nanterre La Défense
Université Paris Ouest Nanterre La Défense
200 Avenue de la République
92001 Nanterre
sylvain.loiseau@u-paris10.fr

Philippe Gréa

UMR 7114 Modyco – CNRS/Université Paris Ouest Nanterre La Défense
Université Paris Ouest Nanterre La Défense
200 Avenue de la République
92001 Nanterre
philippe.grea@u-paris10.fr

Jean-Philippe Magué

UMR 5191 Icar – CNRS/Université Lyon 2/ENS de Lyon
Ecole Normale Supérieure de Lyon
15, Parvis René Descartes
BP 7000
69342 LYON cedex 07
jean-philippe.mague@ens-lyon.fr

RÉSUMÉ. *Depuis quelques années, un certain nombre d'auteurs ont tenté de décrire le sens lexical en s'appuyant sur un objet mathématique, les graphes, et en particulier, les graphes issus de dictionnaires. Cependant, ces expériences ont privilégié des dictionnaires de synonymes ou une analyse en termes de synonymie ou de distance sémantique. Dans cet article, nous proposons une analyse des graphes de dictionnaire de langue monolingue. Ceux-ci, à la différence des dictionnaires de synonymes, ne codent pas une seule relation sémantique, à savoir la synonymie, mais des relations sémantiques beaucoup plus hétérogènes. Cette complexité explique la difficulté à construire des graphes sémantiquement interprétables à partir de ces dictionnaires. Cependant, nous montrons que des graphes construits sur les dictionnaires de langue manifestent des propriétés sémantiques remarquables, qui relèvent d'un cadre sémantique onomasiologique. Nous insistons sur la*

prise en compte des propriétés textuelles des dictionnaires et proposons une analyse contrastive de plusieurs graphes.

ABSTRACT. *A number of authors have proposed to describe lexical meaning using graphs as mathematical model, and, in particular, graphs of dictionaries. Most of the experiments however have focused on dictionaries of synonyms. In this paper we analyse graphs built from language dictionaries. Unlike dictionaries of synonyms, language dictionaries do not encode a single semantic relation, the synonymy, but semantic relations much more heterogeneous, and are therefore difficult to use. However, we show that graphs constructed on language dictionaries exhibit remarkable semantic properties in the semantic framework of onomasiology. We insist on taking into account the properties of these dictionaries and propose a contrastive analysis of several graphs.*

MOTS-CLÉS : *dictionnaire, sens lexical, onomasiologie, classe sémantique, graphe petit monde.*

KEYWORDS: *dictionary, lexicon, onomasiology, semantic classes, small world graphs.*

1. Introduction

L'enjeu de l'utilisation, pour la description lexicale et sémantique, des informations accumulées dans les dictionnaires a souvent été souligné. Coseriu (2001 : 218) note ainsi : « on dispose [en lexicologie] des résultats acquis par les dictionnaires unilingues et par les dictionnaires de synonymes et d'antonymes, résultats qui ne sont nullement à dédaigner. » Rey (2008 : 44) note encore : « La sémantique théorique peut cependant utiliser les résultats de l'analyse sémique de dictionnaires malgré sa grossièreté et ses erreurs, à condition de voir dans les produits de la lexicographie un reflet pragmatique, idéologique et inégalement élaboré des faits de langue ». La linguistique de corpus renouvelle cet intérêt pour les dictionnaires en permettant d'observer, à l'échelle du dictionnaire entier, des structurations et des regroupements lexicaux. Dans le cadre de la sémantique structurale, dans lequel nous nous plaçons, une question centrale est celle de la construction et du fonctionnement des classes lexicales, qu'elles soient construites en langue ou en contexte. Pour cette question, la connaissance, à l'échelle du dictionnaire entier, des regroupements régulièrement effectués pourrait être d'un intérêt évident : ces regroupements à grande échelle, qui échappent à l'attention consciente du lexicographe, ont-ils un intérêt sémantique ? Quelles théories du lexique traduisent-ils ? Dans quelles mesures sont-ils relatifs aux conditions socio-historiques du dictionnaire considéré ? Peuvent-ils apporter des matériaux pour la description lexicale ?

L'une des méthodes utilisées pour cette analyse macroscopique des dictionnaires est la représentation du dictionnaire au moyen d'un graphe. L'utilisation de graphes en sémantique lexicale s'appuie sur un regain d'intérêt plus général pour la théorie des graphes en sciences sociales qui fait suite à la mise en évidence d'une famille de

graphes dite « petit monde », (Watts et Strogatz, 1998 ; Barabási & Albert, 2002)[1]. Ces graphes sont connus pour modéliser un grand nombre de phénomènes et s'avèrent particulièrement adaptés pour rendre compte des relations sociales : réseaux de connaissances, de citations, de co-auteurs dans des publications, *etc*.

Les graphes petits mondes sont généralement définis au moyen de deux indicateurs (voir Véronis 2004 pour une présentation succincte ; Newman 2003 pour une présentation complète). Le premier est le *plus court chemin moyen* (désormais, L). Pour chaque paire de nœuds prise dans un graphe, le plus court chemin est le nombre d'arêtes minimal nécessaire pour relier ces deux nœuds. Le plus court chemin moyen est la moyenne du plus court chemin pour toutes les paires de nœuds d'un graphe. Le second indicateur est le *coefficient de clustering* (désormais, C). Il donne une mesure de la tendance des nœuds d'un graphe à être organisés en sous-ensembles à l'intérieur desquels les nœuds sont fortement connectés entre eux, et faiblement connectés aux nœuds extérieurs. Dans un graphe où C est élevé, si deux nœuds sont reliés, alors un nœud relié à l'un des deux aura une probabilité élevée d'être également relié à l'autre.

Ces deux indicateurs permettent d'opposer les graphes petits mondes aux graphes réguliers et aux graphes aléatoires. Dans le cas des graphes dits réguliers (où tous les nœuds ont le même nombre de nœuds adjacents), L est long, et C élevé. A l'inverse, dans les graphes aléatoires (où les arêtes sont distribuées aléatoirement) L et C sont faibles. Contrairement au précédent graphe, les nœuds d'un graphe aléatoire ne s'organisent pas en sous-ensembles.

Les graphes petits mondes se situent à mi-chemin du graphe régulier et du graphe aléatoire et occupent de la sorte une place intermédiaire entre l'ordre et le désordre. Ils se caractérisent par le fait que L est petit et C élevé (présence de sous-ensembles bien identifiés). L'intérêt de tels graphes tient alors dans le fait qu'ils se prêtent facilement à des procédures (des algorithmes) de découverte de ces sous-ensembles fortement interconnectés, dits « communautés ».

On a ainsi pu montrer (cf. Newman, 2003 pour une recension) que des phénomènes aussi différents que les connexions des pages sur internet, les réseaux de co-auteurs d'articles scientifiques, les réseaux d'interactions de protéines, ou encore les connexions entre individus dans tout type d'interaction sociale (connexions téléphoniques, réseau de connaissance, etc.), avaient les propriétés des graphes petits mondes. L'omniprésence de cette notion, et la prétention à en faire une « signature » commune d'une architecture universelle a toutefois pu être critiqué : pour Fox Keller (2005) les caractéristiques des graphes petits mondes ne disent rien sans un examen des contraintes effectivement à l'œuvre dans les réalités considérées. Pour nous, cette représentation en graphe a seulement un statut de

[1] Nous avons plaisir à remercier Pierre Corbin et Nathalie Gasiglia pour leurs relectures attentives et leurs conseils. Nous remercions également l'ATILF qui nous a permis, dans le cadre d'une convention, d'utiliser le TLFi pour ces expériences, ainsi que trois relecteurs anonymes de la revue TAL.

méthode exploratoire, analogue à celui des méthodes factorielles[2] ; nous ne faisons aucune hypothèse sur la plausibilité de ces graphes comme modèle au sens fort (comme modèle cognitif à base cérébraliste par exemple).

Les dictionnaires, qui se prêtent intuitivement à une représentation en grands réseaux, ont été représentés au moyen de graphes selon deux points de vue sensiblement différents : une première approche avance une représentation géométrique du sens lexical en s'appuyant sur la notion de clique[3] ; on parlera alors, selon les sensibilités théoriques, d'« espace sémantique » (Ploux & Victorri, 1999)[4] ou d'« atlas sémantique » (Ploux & Ji, 2003)[5]. Une seconde approche se fonde sur l'utilisation d'algorithmes de classification (Gaume, 2004) et donne lieu à une métrique appelée « proxémie », qui mesure, selon l'auteur, une distance sémantique entre les mots.

Au-delà des divergences sur la nature des algorithmes employés (détection de cliques *vs* marche aléatoire), ces travaux ont toutefois en commun le fait de n'utiliser qu'un seul et même dictionnaire, le dictionnaire de synonymes du CRISCO (dicoSyn). De rares travaux utilisent des dictionnaires de langue, par exemple Muller *et al.* (2006), mais ils illustrent plutôt qu'ils n'infirment la difficulté à transposer cette méthode aux dictionnaires de langue : d'une part, les auteurs doivent construire un graphe selon des règles complexes et *ad hoc* ; d'autre part, le graphe produit ne reçoit aucune interprétation d'ensemble et peut seulement être utilisé pour extraire des « candidats synonymes » d'un mot donné pouvant, selon les auteurs, assister un travail pratique (lexicographique).

La raison principale de ces difficultés rencontrées avec les dictionnaires de langue est sans doute que ces dictionnaires mettent en jeu des relations beaucoup plus hétérogènes que la seule synonymie. Par conséquent, l'arête d'un graphe construit sur un dictionnaire de langue est extrêmement équivoque. De plus, les travaux cités privilégient, dans l'interprétation des graphes produits, la notion de synonymie (ou des notions dérivées comme celle de « similarité sémantique »). Or, sans doute, d'autres notions sémantiques sont nécessaires pour qualifier les graphes produits. Enfin, ces travaux considèrent le dictionnaire comme une « structure » donnant accès à la langue, sans prendre en compte le fait qu'il s'agit d'abord d'un texte relevant d'un genre et de déterminismes culturels et idéologiques.

---

[2] Des investigations de la structure d'ensemble des dictionnaires utilisant des méthodes vectorielles ont d'ailleurs été proposées, cf. Valette *et al.* 2006.

[3] Une clique est un sous-graphe maximum complet, c'est-à-dire un ensemble de nœuds et d'arêtes tel que chaque nœud est relié à tous les autres.

[4] http://www.crisco.unicaen.fr/cgi-bin/cherches.cgi

[5] http://dico.isc.cnrs.fr/dico_html/fr/index.html

La perspective que nous adoptons dans cet article est donc différente. Plutôt que de chercher à construire un graphe qui permette de retrouver une relation lexicale donnée, comme la synonymie, nous cherchons à décrire et qualifier sémantiquement la structure d'un graphe issu d'un dictionnaire de langue. Nous montrerons que les graphes de dictionnaires de langue permettent d'accéder à une information lexicale et sémantique, en dépit de l'hétérogénéité des relations sémantiques entre mots des gloses, à la condition de prendre en compte ses propriétés textuelles. Pour montrer l'importance de ces propriétés textuelles, nous construirons et comparerons différents types de graphes. Cette information est structurée de façon onomasiologique plutôt que sémasiologique. Elle reflète notamment les choix éditoriaux des dictionnaires ainsi que leur contenu culturel et idéologique. Les graphes de dictionnaires s'avèrent particulièrement utiles pour contraster et décrire les dictionnaires eux-mêmes.

Nous nous appuierons sur un dictionnaire de langue monolingue, le *Trésor de la Langue Française* (TLF). Les difficultés énumérées ci-dessus sont particulièrement nettes dans le cas d'un dictionnaire de type trésor[6], comme le TLF, qui propose une somme des connaissances sur le lexique. Si les dictionnaires sont aussi des objets à vocation pratique et répondant à des impératifs commerciaux, dans le cas d'un trésor la dimension descriptive est prépondérante.

Dans la section 2, nous analyserons la complexité de ces dictionnaires et les raisons des difficultés qu'ils opposent à un traitement automatique. Les difficultés de l'utilisation des dictionnaires de langue peuvent être rapportées au fait que les dictionnaires de langue ne sont pas, comme les dictionnaires de synonymes, une énumération de lexèmes, mais des textes. Les dictionnaires de langues ne peuvent donc pas être représentés comme des graphes sans une analyse préalable. Il faut par exemple prendre en compte la complexité de la microstructure, qui distingue différentes sections : glose, exemple, remarque, etc., ainsi que les normes rédactionnelles du lexicographe. Sur la base de ces propriétés, nous exposons différents graphes élaborés à partir du TLF. Chacun de ces graphes capture une partie des informations de ces dictionnaires. Ces graphes sont en particulier contrastés avec les graphes issus d'un dictionnaire de synonymes, dicoSyn, afin de montrer leurs spécificités.

La section 3 expose l'algorithme de marche aléatoire que nous appliquons aux trois graphes issus du TLF.

La dernière section consiste à faire l'étude contrastive des résultats ainsi obtenus. Nous montrerons qu'il est possible de caractériser la différence entre les graphes de dictionnaire de langue et les graphes de synonymes en faisant appel à un couple

---

6 « Inventaire des unités lexicales d'une langue visant à l'exhaustivité » (Mounin, cité par le TLF, s. v. *trésor*);

d'opposition bien connu dans le champ sémantique : l'opposition entre onomasiologie et sémasiologie.

2. Constitution des graphes de dictionnaires de langue : méthodes et difficultés

Il y a de bonnes raisons de faire l'hypothèse qu'un graphe issu d'un dictionnaire de langue soit très différent d'un graphe issu d'un dictionnaire de synonymes (désormais, *graphe de synonymes*). Dans ce qui suit, nous allons donc revenir sur les différences qui séparent les deux types de dictionnaire en comparant dicoSyn (dictionnaire des synonymes) d'un côté et le TLF (dictionnaire de langue) de l'autre.

2.1. Le graphe de synonymes

Dans dicoSyn, la liste des synonymes de *arbre* est la suivante (classé par ordre alphabétique) :

> ARBRE : arborescence, arbuste, axe, barre, bielle, bois, diagramme, essieu, forêt, fût, manivelle, pivot, plante, précipité, tige, végétal, vilebrequin.

Pour construire un graphe à partir de cette entrée on peut connecter le mot d'entrée *arbre* à tous ses synonymes comme, du reste, le font tous les auteurs précédemment cités. Dans le graphe obtenu (figure 1, gauche), qui correspond à ce qu'on appelle le voisinage immédiat de *arbre*, les sommets renvoient aux mots et chaque arête est censée coder une seule et même relation sémantique, la relation de synonymie.

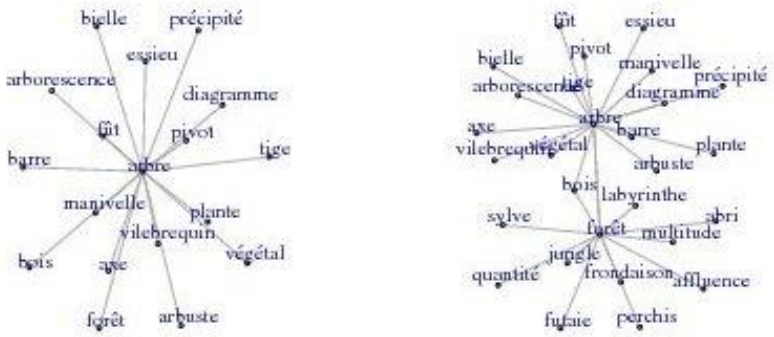

Figure 1 : voisinage de 'arbre' (à gauche), sous graphe du voisinage de 'arbre' et 'forêt' (à droite), dicoSyn.

On peut reproduire cette opération pour le mot d'entrée *forêt*, qui fait partie des synonymes de *arbre*, et pour lequel dicoSyn donne la liste de synonymes suivants :

FORET : abri, affluence, arbre, bois, frondaison, futaie, jungle, labyrinthe, multitude, perchis, quantité, sylve.

Du fait que *forêt* est lui-même connecté à *arbre*, et que les deux listes de synonymes comportent le mot *bois*, les deux ensembles de synonymes sont connectés deux fois (figure 1, droite).

Pour obtenir le graphe complet associé à dicoSyn, il suffit de reproduire cette opération pour tous les mots d'entrée de dycoSyn[7]. Un tel graphe, du fait qu'il manifeste les propriétés caractéristiques des graphes petits mondes, se prête à un ensemble d'algorithmes aujourd'hui bien connus et sur lesquels nous reviendrons *infra*.

2.2. Graphe de dictionnaire de langue : diversité des graphes possibles

Toutefois, les choses se présentent de façon très différente lorsque nous nous intéressons à un dictionnaire de langue. Dans ce cas, en effet, l'article associé à un mot d'entrée est une microstructure hiérarchisée qui donne accès à une information hétérogène. Le cas de *arbre* dans le TLF permet de s'en rendre compte assez facilement. La taille de l'article interdisant toute citation *in extenso*, nous n'en ferons figurer que le début (en tronquant les exemples) dans ce qui suit :

ARBRE, subst. masc.
I. [L'arbre désigne un végétal ou sa représentation]
A. BOT. **Végétal ligneux, de taille variable, dont le tronc se garnit de branches à partir d'une certaine hauteur**. ***De grands arbres, branches d'arbres, troncs d'arbres*** :
1. *Le végétal est si bien composé* [...] BERNARDIN DE SAINT-PIERRE [...]
2. *Entre les quelques champs* [...] L. HÉMON [...]
3. *... et, tout autour du tronc,* [...] GIDE [...]
4. *L'homme était un ouvrier agricole* [...] MONTHERLANT [...]
SYNT. ***Cime, ombre, pied d'un arbre, bouquets d'arbres; abattre, planter, tailler des arbres, grimper aux arbres; arbres dépouillé, mort, en fleurs***.
B. Spécifications de l'arbre
1. Spécifications naturelles ou arboricoles
a) [L'accent est mis sur une caractéristique tirée de l'origine, des propriétés réelles ou supposées d'une espèce d'arbres ou d'arbustes] Arbre à, arbre de : [...]

La construction d'un graphe à partir de cet article se heurte inévitablement à de nombreuses questions. Faut-il tenir compte de tous les mots qui se trouvent contenus dans l'article (toutes sections / catégories syntaxiques confondues) et les connecter au mot d'entrée ? Faut-il au contraire créer des sous-graphes en ne retenant de l'article qu'un certain type d'information (à l'exclusion de tous les autres) ? Cela

7 Le graphe comprend alors 49 149 nœuds et 200 606 arêtes.

permettrait ainsi de construire, par exemple, un sous-graphe uniquement fondé sur les gloses définitionnelles (notée en gras dans la citation précédente), ou un sous-graphe fondé exclusivement sur les exemples (italique), ou bien un sous-graphe fondé sur les syntagmes associés au mot d'entrée (gras italique), ou encore un sous-graphe fondé sur des marqueurs de domaine (en l'occurrence, BOT.), *etc*. Par ailleurs, faut-il créer des sous-graphes qui tiennent compte des catégories syntaxiques et construire de la sorte le sous-graphe des noms, le sous-graphe des verbes, le sous-graphe des adjectifs, *etc*.[8] ?

On le voit, cette transposition d'un dictionnaire vers un graphe, qui semblait évidente pour dicoSyn, devient beaucoup plus délicate avec un dictionnaire de langue (*a fortiori*, un trésor comme le TLF). À première vue, on pourrait penser qu'un graphe réduit aux seules gloses définitionnelles (graphe de définitions) constituerait un bon objet, au moins parce qu'il se limite à ce qui fait la vocation première de tout dictionnaire de langue, à savoir donner le sens d'un mot. Mais il s'agit là d'un préjugé qu'on peut facilement remettre en cause : le graphe construit à partir des seuls exemples du TLF (graphe des exemples), pour prendre une autre possibilité, constitue un objet tout aussi pertinent que le graphe de définitions, en particulier lorsque l'on sait que ces exemples sont généralement choisis par les rédacteurs de telle sorte qu'ils donnent un contexte qui corresponde le mieux au sens glosé, et qui, dans la plupart des cas, renvoient à des situations prototypiques ou des thématiques caractéristiques (pour *arbre*, le premier exemple du TLF est ainsi tiré d'un traité de botanique[9]). Il n'existe donc aucune raison objective de privilégier tel graphe plutôt que tel autre, et il faut se résoudre à les étudier tous en les comparant entre eux.

Dans le cadre du présent travail, nous avons volontairement réduit notre étude à trois graphes : un graphe de définition, un graphe de cooccurrences issu des définitions et un graphe de cooccurrences issu des exemples. La section suivante expose le protocole dont nous nous sommes servis pour les construire.

---

8 Cette dernière question ne se pose pas directement pour le graphe de synonymes : la relation de synonymie implique que les synonymes sont de même catégorie syntaxique. Par exemple, le graphe de noms synonymes et le graphe de verbes synonymes sont déjà séparés dans le dictionnaire.

9 Rappelons que le TLF donne des exemples « cités », et non « forgés » ; la différence cependant n'est pas nécessairement importante du point de vue de l'opposition entre exemples et définitions (« il n'y a aucune raison pragmatique de distinguer ouvertement les exemples produits *ad hoc* des exemples extraits d'un corpus », Rey 2008 : 75).

2.3. Graphe de définitions et graphe de cooccurrences

2.3.1. Graphe de définitions : méthode et difficultés

Afin de limiter notre propos à l'essentiel et de permettre une comparaison directe avec le dictionnaire de synonymes portant sur les noms, nous n'avons retenu de la glose définitionnelle du TLF que les noms. Nous n'avons également conservé que les articles se rapportant à une entrée nominale. L'utilisation de l'ensemble des parties du discours rend en effet la tâche plus difficile. Véronis (2004) par exemple souligne l'effet perturbateur des verbes sur les graphes du fait de leur plus forte polysémie. Malgré cette simplification, plusieurs problèmes spécifiques à l'utilisation d'un dictionnaire de langue demeurent et impliquent un certain nombre de choix.

(i) La question des gloses multiples

Un article de dictionnaire présente très souvent plusieurs acceptions différentes pour un même mot (du fait de leur polysémie, ou de possibles variantes – diachroniques, diatopiques, diastratiques, etc. – ou bien à cause des entrées subordonnées). La question se pose alors de savoir s'il faut tenir compte d'une seule glose, par exemple, celle qui se rapporterait au sens « premier » du mot, si tant est qu'une telle notion soit pertinente et qu'il soit possible de l'identifier parmi les différentes acceptions, ou si l'on doit au contraire tenir compte de toutes les gloses sans faire de distinction. Cette question est importante dans la mesure où la manière d'y répondre aura un impact considérable sur le graphe résultant. Pour illustrer cela, prenons le cas de *blockhaus* dans le TLF :

> BLOCKHAUS :
> A. Vx. **Maison construite en troncs d'arbres.**
> B. TECHN. MILIT.
> 1. Vieilli. **Redoute ou fortin détaché, en bois, de dimension variable, communiquant souvent à un ouvrage principal par des conduits souterrains, et servant dans ce cas, d'ouvrage avancé.**
> 2. Mod. **Ouvrage défensif en béton armé et blindé, muni de pièces d'artillerie.**
> [En parlant d'une chose abstr.] :
> P. ext., MAR. MILIT. **Poste de combat du commandant à bord des navires de guerre.**
> C. Arg. **Haut-de-forme.**

L'article présente cinq gloses distinctes (notées en gras), regroupées sous trois sections différentes : la première fait état d'un emploi archaïque où *blockhaus* est un hyponyme de *maison* (variation diachronique), la seconde traite les acceptions rattachées au domaine militaire, tandis que la dernière rapporte un emploi argotique (variation diastratique) où *blockhaus* est synonyme de *haut-de-forme*. Cette disposition obéit à une convention de rédaction du TLF selon laquelle les acceptions sont présentées dans leur ordre historique. Dans cet état de choses, plusieurs solutions s'offrent à nous : (a) nous pouvons construire le graphe de définitions sur

la base de l'acception historiquement première, c'est-à-dire la première glose qui se présente dans l'article, ce qui aura pour inconvénient de connecter les sommets du graphe selon des relations qui ne sont pas nécessairement pertinentes aujourd'hui ; (b) nous pouvons construire le graphe de définitions à partir des cinq gloses, auquel cas, on mélangera des acceptions qui n'ont aucun rapport entre elles (*haut-de-forme* et *maison* seront ainsi connectés à *blockhaus* sans autre distinction) ; (c) nous décidons de construire le graphe de définitions sur la base de la glose correspondant au(x) sens contemporain(s) (« ouvrage défensif »), mais cette dernière solution est fortement compromise par le fait qu'il n'existe aucun moyen d'identifier automatiquement, dans un article du TLF, la ou les acception(s) contemporaine(s) d'un mot (les marques d'usage, comme « Mod. », moderne, sont trop peu systématiques).

Cette question des gloses multiples trouve néanmoins un début de solution quand on élargit la réflexion à d'autres types de dictionnaires. Si un trésor comme le TLF est en effet susceptible de multiplier les acceptions, un dictionnaire d'apprentissage, à l'inverse, tendra à les réduire en se concentrant sur le sens le plus usuel[10]. De ce point de vue, le traitement (en une seule glose) de *blockaus* dans le Micro-Robert illustre très clairement la différence :

> BLOCKHAUS : Petit ouvrage militaire défensif, étayé de poutres ou fortifié en béton.

Le choix de construire le graphe de définitions sur la base des cinq gloses du TLF, ou bien sur la glose unique du Micro-Robert a donc nécessairement des conséquences sur les graphes résultants et leurs propriétés, conséquences qu'il faudrait pouvoir mesurer et caractériser dans le cadre d'un protocole précis. Malheureusement, au moment ou nous écrivons, nous n'avons pas les moyens de mener les expériences voulues du fait de la difficulté d'accéder à des versions numériques intégrales de dictionnaires qui sont sous droits. C'est pourquoi nous avons dû réduire nos ambitions en n'explorant qu'une seule des trois réponses énumérées *supra* : la prise en compte de toutes les gloses de l'article (graphe de définitions complet).

(ii) L'hétérogénéité des relations sémantiques et leur écrasement dans le graphe

Revenons au cas de *arbre*. Pour simplifier notre exposé, tenons compte de la première glose (en l'occurrence, l'acception végétale). En fonction de la définition citée *supra*, ce graphe se représente de la façon suivante :

---

[10] Habert (2000) note ainsi : « Après transformation, les dictionnaires les plus savants ne sont pas nécessairement les plus utiles comme matière première pour des expériences. »

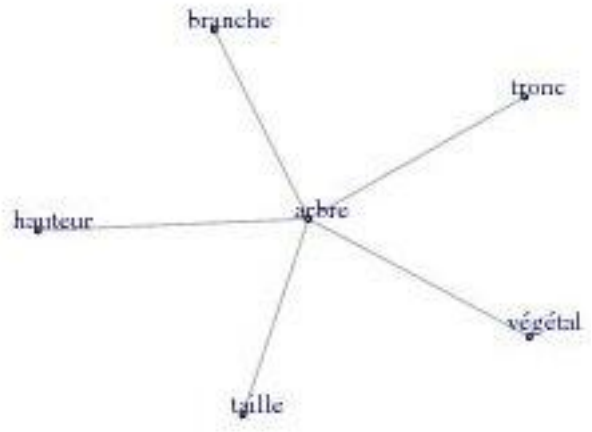

Figure 2: 'Arbre' dans le TLF

Une comparaison de celui-ci avec la Figure 1, où se trouve l'ensemble des unités reliées à *arbre* dans dicoSyn, permet d'illustrer une différence importante. Au-delà des quelques points communs que l'on peut noter (l'hyperonyme *végétal* est ainsi directement connecté à *arbre* dans le graphe de synonymes comme dans le graphe de définitions), il y a surtout une divergence de fond, puisque le graphe de définitions s'appuie sur un grand nombre de relations sémantiques distinctes, en l'occurrence, l'hyperonymie (*arbre – végétal*), la relation partie/tout (*arbre – tronc / branches*), auxquelles s'associent des noms de qualité (*taille*, *hauteur*), et dont on ne trouve aucune trace dans le dictionnaire de synonymes.

Une telle hétérogénéité peut aller beaucoup plus loin dans la mesure où on ne tient pas compte de la directionnalité de la relation de définition et que l'on utilise le même type d'arête pour coder la relation vedette – glose et glose – vedette. Dans ces conditions, en effet, le nœud *arbre* est aussi (et surtout) relié à l'ensemble des mots dont la (ou les) glose(s) contien(nen)t le mot *arbre*. Or, ces derniers sont extrêmement nombreux et entretiennent avec *arbre* des relations de nature très diverse. Il y a par exemple les mots dont *arbre* est l'hyperonyme, comme *châtaignier* ou *aulne*, des collectifs (*forêt*), ou des noms d'activité qui ont à voir (plus ou moins directement) avec les arbres, comme *déforestation* ou *écorçoir*. Mais un nombre important de mots se trouvent connectés à *arbre* sur la base de relations beaucoup plus ténues. C'est par exemple le cas de *corde* qui se trouve rattaché à *arbre* par le biais d'une relation métonymique très secondaire dans la glose :

> CORDE : Assemblage obtenu par torsion de fils de matières textiles (chanvre, coton, laine, soie), synthétiques (nylon), métalliques ou autres (poil, crin, écorce d'arbre, jonc, etc.), de grosseur et de longueur variables, et servant à lier, attacher, soutenir.

Or, cette unité dont la glose contient *arbre* sans que ce dernier soit véritablement indispensable à sa définition, est connectée à *arbre* au même titre que *frondaison* ou *châtaignier* et le graphe ne fera pas du tout état du caractère plus ou moins accidentel des relations sur lesquelles se fondent ces connexions. Pour finir, il est à noter que les définitions par antonymie[11] donnent aussi lieu à une connectivité problématique du point de vue d'une notion de similarité sémantique ou de synonymie.

Ce rapide tour d'horizon des unités connectés à *arbre* nous permet alors de comprendre l'une des conséquences fondamentales de toute transposition d'un dictionnaire de langue dans un graphe, à savoir l'écrasement pur et simple des relations sémantico-lexicales qui s'y trouvent, de leur diversité, de leur importance relative, de leur hiérarchisation, et ce, depuis la synonymie jusqu'à l'antonymie, en passant par les cas d'occurrences quasi-accidentelles ou ayant une importance très secondaire. Le graphe de dictionnaire de langue (et plus spécifiquement, le graphe de définition) a pour effet de réduire toute la complexité d'un dictionnaire de langue à une seule question, celle de la présence ou de l'absence d'une unité dans les gloses, provoquant ainsi une déperdition d'informations considérable. C'est là une différence fondamentale avec le graphe de synonymes.

Malgré ces remarques, nous verrons, dans les sections suivantes, que cet « écrasement », lorsqu'il est bien compris, n'interdit en aucune manière l'étude systématique et l'utilisation avisée des graphes de dictionnaire de langue.

(iii) La polysémie

Avant cela, il nous faut aborder une dernière difficulté liée au graphe de définitions, et qui porte sur l'inévitable question de la polysémie. Nous l'avons déjà rencontré dans le cadre des gloses multiples ; or, elle resurgit à l'autre extrémité du processus. Le cas de *arbre*, pour ne pas changer, permet d'illustrer le problème. Ainsi, parmi les mots connectés à *arbre* dans le graphe de définitions, on remarque la présence de *laryngologie*. Or, il s'avère que cette connexion *arbre – laryngologie* s'appuie sur la glose suivante :

> LARYNGOLOGIE : Étude du larynx et de ses affections; p. méton., branche de la médecine qui traite du larynx, du pharynx et de l'arbre trachéo-bronchique

Ici, c'est la polysémie de *arbre* dans les gloses qui est en cause. Cette connexion dans ce cas précis est renforcée par la présence de *branche* (« branche de la médecine »), qui se trouve déjà connecté à *arbre* pour son acception végétale. Une telle inter-connectivité risque donc de rapprocher artificiellement le domaine

---

[11] Par exemple, dans Le Littré, *herbe* prend la définition suivante : « Toute plante qui, n'étant point arbre ». (Martin 1983, p. 60 et s.) caractérise de telles définitions comme « hyperonymiques négatives ».

laryngologique et l'arboriculture à travers la clique arbre/branche/laryngologie et ne peut qu'avoir des conséquences importantes sur la structuration du graphe.

Si l'on devait utiliser le dictionnaire pour observer des similarités lexicales, ce problème spécifique ne pourrait se régler que par une procédure de désambigüisation du dictionnaire[12]. Cependant, nous montrerons plus loin que les graphes de dictionnaire de langue exhibent une structuration sémantique particulière, différente de celle des graphes de synonymes, et que cette différence permet justement d'éviter le recours à une procédure de désambigüisation.

Au final, le nombre des noms connectés à *arbre* dans le graphe de définitions s'élève à 355. On dira alors que le nœud *arbre* est de *degré* 355. On dira d'un nom connecté à *arbre* qu'il est son *adjacent*. L'ensemble des adjacents de *arbre* sera appelé le *voisinage* de *arbre*.

2.3.2. Graphe de cooccurrences

On peut construire un graphe à partir des gloses définitionnelles selon au moins deux procédures différentes. La première, que nous venons d'étudier en détail dans la section précédente, consiste à relier la vedette aux mots de sa glose (et inversement, des mots de la glose au mot vedette). Nous avons montré quelles difficultés cette approche peut soulever. Une seconde procédure consiste cette fois à relier entre eux les mots de la glose définitionnelle et à construire de la sorte un graphe de cooccurrences (ce qui se traduit par l'exclusion du mot d'entrée). Le graphe résultant est alors très différent du premier : dans le cas d'un graphe de définitions, l'arête note la présence d'un mot à l'intérieur de la glose définitionnelle d'un autre mot ; dans le second cas, en revanche, l'arête note le fait que deux mots apparaissent simultanément dans au moins une glose. La construction de graphes à partir de la relation de cooccurrence a déjà montré son intérêt dans d'autres contextes (Véronis 2004), et nous allons montrer que dans le cas qui nous occupe, l'utilisation d'un graphe de cooccurrences permet de prendre davantage en compte la dimension textuelle du dictionnaire.

Pour illustrer les différences entre un graphe de cooccurrences et un graphe de définitions, revenons une nouvelle fois au traitement de *arbre* dans le TLF. Si l'on tient compte uniquement de la première glose définitionnelle (afin de simplifier les illustrations, rapidement illisibles si l'on tient compte de toutes les gloses), le graphe de cooccurrences a la forme présentée à la Figure 3, à comparer avec la Figure 2 :

12 Comme celle décrite dans Gaume *et al.* (2004).

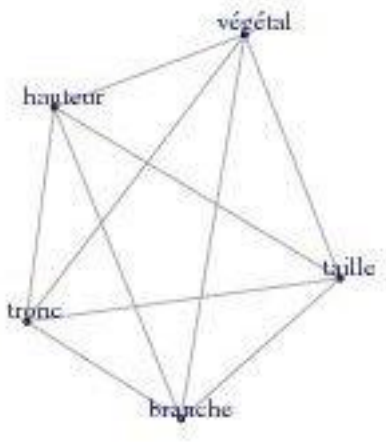

Figure 3: graphe de cooccurrences de 'arbre' dans le TLF

Les éléments localement connectés sont donc très différents selon qu'on fait le choix de l'une ou de l'autre solution. Or, il apparaît que le graphe de cooccurrences offre plusieurs avantages non négligeables par rapport au graphe de définitions :

(i) Dans un graphe de cooccurrences, seuls les mots qui sont dans les gloses définitionnelles sont pris en compte. Les mots qui apparaissent uniquement dans la nomenclature du dictionnaire mais qui n'apparaissent jamais dans aucun article (et plus spécifiquement, dans aucune glose du dictionnaire) ne sont pas pris en compte. En d'autres termes, le vocabulaire utilisé pour construire le graphe est réduit au seul vocabulaire « définissant ». Or, ce dernier est bien moins important que la nomenclature d'un dictionnaire puisque, en moyenne, la richesse lexicale des gloses définitionnelles ne représente qu'environ 30 % de la richesse lexicale de la nomenclature. Le mot *laryngologue* par exemple est un mot qui n'existe que comme unité définie dans le TLF, mais qui n'est jamais utilisé pour définir un autre mot du dictionnaire. Par conséquent, l'unité *laryngologue* n'existera pas dans le graphe de cooccurrences[13]. Le graphe de cooccurrences exclut ainsi un très grand nombre de mots rares.

(ii) Le fait d'utiliser séparément les acceptions d'une même vedette permet d'éviter les connexions aberrantes entre les mots des différentes acceptions, puisqu'on n'interconnecte que des mots apparaissant ensemble dans chaque glose.

[13] En revanche, c'est bien la glose du mot *laryngologue* qui permet de joindre les sommets *spécialiste* et *laryngologie* puisque ces derniers y sont en cooccurrence.

Prenons par exemple deux gloses de *arbre* dans le TLF correspondant à deux acceptions : « Végétal ligneux, de taille variable, dont le tronc se garnit de branches à partir d'une certaine hauteur. » et « Pièce maîtresse, parfois tournante, qui, dans une machine, sert de support à d'autres pièces animées. » Dans le graphe de définition, *machine* se retrouve connecté à *végétal* par l'intermédiaire du mot *arbre*. Dans le graphe de cooccurrences, au contraire, *machine* est connecté aux autres mots de sa glose (qui portent tous sur le domaine //mécanique//), et *végétal* est connecté aux autres mots de sa glose ; dès lors, il n'y a plus de connexion entre *végétal* et *machine*. On règle ainsi la plus grande part du problème de la polysémie. La question des gloses multiples, que le graphe de définitions ne parvient pas à résoudre de façon satisfaisante, trouve ici une solution simple. Pour la même raison, les homonymes ne posent pas de difficultés puisque chaque homonyme introduit des gloses qui sont traitées indépendamment. Naturellement, on pourrait objecter que la polysémie des mots en usage dans les gloses définitionnelles reste toujours un problème. Par exemple, *arbre* peut être utilisé pour désigner tantôt le végétal, tantôt la pièce d'une machine. Cependant, on constate que le genre de la définition implique de n'utiliser les mots que dans quelques acceptions principales ; il y a donc infiniment moins d'acceptions des mots en usage dans les gloses que d'acceptions définies par ces gloses. Par exemple, l'une des gloses d'*arbre* dans le TLF est « P. méton. La fête organisée à l'occasion de Noël. ». Or, il est évident que les lexicographes, eux, n'utilisent pas cette acception spécifique de *arbre* dans les gloses. Dans le graphe de définition, *fête* est donc connecté, via *arbre*, à *végétal* et *machine*. Dans le graphe de cooccurrences, il n'y a pas une telle connexion.

(iii) Un dernier avantage du graphe de cooccurrences tient dans le fait qu'il permet une pondération des arêtes en fonction du nombre de cooccurrences observées sur l'ensemble du dictionnaire. Ainsi, deux mots qui apparaissent simultanément dans une seule et unique glose du dictionnaire donneront une arête de poids « faible » (1). C'est ce qui se passe pour l'arête reliant *branche* et *matière* : ces deux termes n'apparaissent à l'intérieur de la même glose qu'à une seule occasion, à savoir la glose de *thermodynamique*. Mais lorsque deux cooccurrents le sont *n* fois dans l'ensemble du dictionnaire (c'est-à-dire, apparaissent ensemble dans *n* gloses différentes), alors l'arête permettant de les relier peut être « renforcée », et prendre un poids de *n*. Par exemple, la cooccurrence entre *fil* et *soie* apparaît dans 78 gloses différentes. Cela se traduira, dans le graphe résultant, par le fait que l'arête qui les relie aura un poids de 78. Dans le graphe de définition, les arêtes ne peuvent bénéficier de cette variation de poids dans la mesure où une arête ne peut représenter qu'une ou deux cooccurrences (dans le cas où les deux mots sont respectivement dans la glose l'un de l'autre).

Cette façon de procéder a l'avantage de réduire un peu plus les effets de polysémie observés avec *arbre* et *branche*. Ainsi, les arêtes reliant *larynx – arbre* ou *larynx – branche* (grâce à la glose de *laryngologie*, citée *supra*), sont toutes les deux pondérées à 1 : de telles cooccurrences ne se produisent qu'une seule fois dans tout le TLF. A l'inverse, la seule arête *arbre – branche* est, quant à elle, pondérée à 110 : cette cooccurrence est observée dans 110 gloses définitionnelles. La prise en compte

du poids permet donc de reporter structurellement dans le graphe le caractère marginal de la connexion *larynx – arbre* et le caractère plus essentiel de la relation *arbre – branche*. Toutefois, pour des raisons de place, il ne nous a pas été possible d'intégrer dans le présent travail les résultats issus des graphes pondérés.

En résumé, les graphes de cooccurrences ont l'avantage de considérer le dictionnaire comme un texte, plutôt que comme un réseau déjà constitué : dans la mesure où les connexions sont établies entre les mots en usage et non *via* les mots de la nomenclature (en mention), l'arête note une relation de cooccurrence textuelle. Les acceptions représentées dans les gloses et représentées dans le graphe sont un sous ensemble des acceptions possibles, fixé par la pratique lexicographique. Une dimension fréquentielle est donnée par la prise en compte de la fréquence des cooccurrences.

La construction de graphes de cooccurrences peut sembler paradoxale, dans la mesure où l'on peut supposer *a priori* que ce qui est particulièrement « structuré » et intéressant dans un dictionnaire est la relation entre le mot défini et les mots définissant, relation qui se trouve perdue dans le graphe de cooccurrences. Cependant, il apparaît que, dans le cadre de la représentation par graphe, cette relation, localement intéressante, échoue à produire globalement une structure qualifiable, tandis qu'à l'inverse, la relation de cooccurrence permet d'observer des phénomènes plus intéressant, en rapport, on l'a dit, aux propriétés du dictionnaire (cf. *infra*).

On pourrait encore objecter que, dans ces conditions, l'utilisation d'un dictionnaire n'offre plus aucun intérêt : n'importe quel « grand corpus », tels que ceux utilisés en Traitement Automatique des Langues, pourrait être utilisé. Cependant, les relations de cooccurrences observées dans un dictionnaire gardent un intérêt. C'est la notion de genre textuel qui est importante ici : en fonction du genre dont ils relèvent, les textes ne permettent pas d'observer les mêmes relations de cooccurrences. Le texte dictionnairique permet d'observer des relations de cooccurrences en relation avec le genre des définitions ou le genre des exemples, différentes, on peut s'y attendre, des relations de cooccurrences observées dans un roman ou un article d'information.

En résumé, tenir pour structurée la seule relation définissant/défini, c'est ignorer que le dictionnaire est un texte et non un réseau formel, et que l'information sémantique accumulée par les lexicographes s'exprime sous la forme avant tout textuelle des gloses définitionnelles.

3. Parcours des graphes par marche aléatoire

Les graphes ainsi constitués, comptant jusqu'à plusieurs centaines de milliers de nœuds, ne peuvent être décrits qu'à travers des procédures automatisées de découverte. L'intérêt de la construction de graphes de dictionnaire tient à la possibilité d'analyser sa structure globale grâce à ces procédures automatisées.

Une caractéristique des graphes petits mondes est que s'ils sont globalement peu denses, leur nombre d'arêtes étant très petit par rapport à un graphe de même taille complètement interconnecté, ils sont, en revanche, localement denses. Il est en effet possible d'identifier en leur sein des sous-graphes dont les nœuds sont fortement interconnectés. On appelle ces zones denses des *communautés*.

De nombreux algorithmes ont été proposés pour détecter de telles communautés (voir Newman (2003) pour une recension). Ici, nous suivons Gaume (2004) et utilisons des marches aléatoires. Le principe consiste à étudier les caractéristiques de déplacements de nœud en nœud en choisissant aléatoirement, à chaque étape, l'arête empruntée pour passer au nœud suivant. Du fait de la forte densité locale, les premiers pas d'une telle marche tendront à rester dans la communauté du nœud de départ.

Cet algorithme permet de mettre au jour les regroupements et les proximités entre nœuds, proximités basées non pas sur la connexion directe, mais sur l'appartenance des nœuds à des mêmes ensembles, ou communautés. Appelons, à la suite de B. Gaume, *proxémie* la proximité ainsi définie et *proxèmes* les mots les plus proches d'un mot-pôle donné. L'interprétation d'une liste de proxèmes donnée par cet algorithme s'appuie sur le fait que ces proximités se fondent sur la topologie globale du graphe (ou plutôt du sous-graphe à une distance de 3 du mot-pôle, puisque les marches aléatoires ont une longueur de 3). Autrement dit, deux mots peuvent être connectés par une arête, mais être éloignés en termes de proxémie s'ils n'appartiennent pas à un même ensemble. À l'inverse, deux nœuds peuvent ne pas être adjacents mais être néanmoins très proches selon la proxémie s'ils appartiennent à une même communauté.

4. Caractérisation sémantique du graphe de synonymie et des graphes de dictionnaire langue

4.1. Etude contrastive de trois graphes de dictionnaire de langue

Rappelons que trois graphes sont considérés ici : le graphe de définition (gdg) qui connecte l'entrée (nominale) à tous les noms de toutes les gloses définitionnelles (section 2.3.1) ; le graphe de cooccurrences (gcg) qui connecte tous les noms qui apparaissent à l'intérieur d'une même glose définitionnelle (section 2.3.2) ; le graphe des cooccurrences pour les exemples (gce) qui connecte tous les noms qui apparaissent dans un même exemple. Nous avons appliqué un algorithme de marche aléatoire (section 3) pour chacun de ces trois graphes. Le résultat se présente sous la forme d'une liste de termes classés en fonction de leur proximité avec un mot donné. Le mot pivot choisi pour mener à bien cette étude contrastive est le mot *arbre*. Les trois listes sont présentées dans les tableaux 1 à 3.

tronc	végétal	oeil	if	portion

rameau	haie	végétation	lambourde	olivette
bois	dard	massif	coin	soie
branche	rose	plot	support	boîte
souche	plateau	écorce	bille	fruit
sève	mannequin	queue	volant	planche
tige	rosette	noeud	rond	chaudron
couronne	entaille	rame	tympan	assemblage
pivot	gouttière	rondin	volet	coussinet
maille	aile	astragale	allée	plante

Tableau 1 : liste des mots les plus proches d'*arbre* dans le graphe de définition (gdg)

tronc	rameau	réceptacle	vigne	végétal
sève	menuiserie	dizaine	printemps	chêne
cueillette	touffe	racine	boule	gazon
ornementation	région	ébène	alimentation	pré
feuillage	forêt	végétation	taille	pousse
bourgeon	cône	parc	plantation	grosseur
bouquet	fourrage	feuille	maturité	gerbe
jardin	écaille	Nord	brin	branche
souche	hêtre	lamelle	hampe	noir
tige	écorce	enveloppe	excroissance	vert

Tableau 2 : liste des mots les plus proches d'*arbre* dans le graphe de cooccurrences (gcg)

soleil	fenêtre	caillou	voûte	vitre
nuage	sommet	olivier	carré	lit
jet	long	vallée	colonne	travers
pluie	aile	printemps	demi	bas
forêt	acacia	flot	poussière	abri
jardin	plaine	verdure	feuillage	vent
colline	hiver	flamme	été	reflet
mur	ombre	branche	grotte	roc
haut	tronc	toit	fumée	rideau
pente	flanc	parc	hauteur	nu

Tableau 3 : liste des mots les plus proches d'*arbre* dans le graphe des exemples (gce)

À première vue, on observe que le graphe de définition donne une liste de mots moins cohérente mais peut-être plus complète (avec des termes comme *bois*, *fruit*, *gouttière*, qui n'apparaissent pas parmi les 50 premiers mots dans le cas du graphe de cooccurrences) ; le graphe de cooccurrences, à l'inverse, semble globalement plus cohérent mais une partie du vocabulaire lui échappe. Ces quelques observations ne suffisent toutefois pas pour caractériser correctement les différences entre les deux listes. Pour mieux comprendre les différences entre ces deux structurations (relation de définition *vs* relation de cooccurrence) il faut donc porter notre attention sur la distance que les différents proxèmes entretiennent entre eux, et non plus seulement sur celle qu'ils entretiennent avec le mot *arbre*. En effet, les proximités que les mots entretiennent entre eux dans chaque liste, et leur regroupement en ensembles,

donnerait une information plus fine sur la façon dont l'environnement de *arbre* se structure dans chacun des graphes. Ainsi, dans la première liste, on observe par exemple *tronc*, *sève*, *haie* et *massif*. Une question que l'on pourrait se poser serait de savoir si *tronc* et *sève* sont proches l'un de l'autre, tout en étant éloignés, à l'inverse, de *haie* et *massif* (c'est-à-dire si *tronc* et *sève* apparaissent ensemble dans des définitions, mais jamais avec *haie* ou *massif*, qui peuvent être en cooccurrence de leur côté) : une telle répartition, si elle est confirmée par l'ensemble des lexèmes de la liste, serait interprétable comme une structuration de l'environnement de *arbre* entre d'un côté le domaine de son analyse méronymique, de l'autre son utilisation anthropologique et décorative (des exemples d'analyses des répartitions contextuelles de lexèmes associés à un mot-pôle peuvent être trouvés dans Loiseau 2006 ou Blumenthal 2009).

À cette fin, pour les deux graphes concernés (gcg et gdg), nous avons conservé les 200 premiers proxèmes de *arbre*. On utilise la matrice des probabilités de transition (obtenue après une marche aléatoire de 3 étapes) que l'on a réduit aux 200 lignes et 200 colonnes correspondant à ces proxèmes. Cette sous-matrice est soumise à une analyse en composantes principales afin de dégager et résumer les proximités entre éléments proches de *arbre*. Par exemple, *fruit* et *planche* ont tous les deux dans le graphe de définition une proxémie quasiment équivalente par rapport à *arbre*. Or, on peut raisonnablement penser que dans l'ensemble des 200 proxèmes de *arbre*, ils sont en réalité assez éloignés l'un de l'autre. En d'autres termes, peut-être appartiennent-ils dans une certaine mesure à différentes communautés, même si les deux communautés entretiennent des liens avec *arbre*. Nous donnons ci-dessous les résultats de cette analyse en composantes principales pour le graphe de définition et le graphe de cooccurrences.

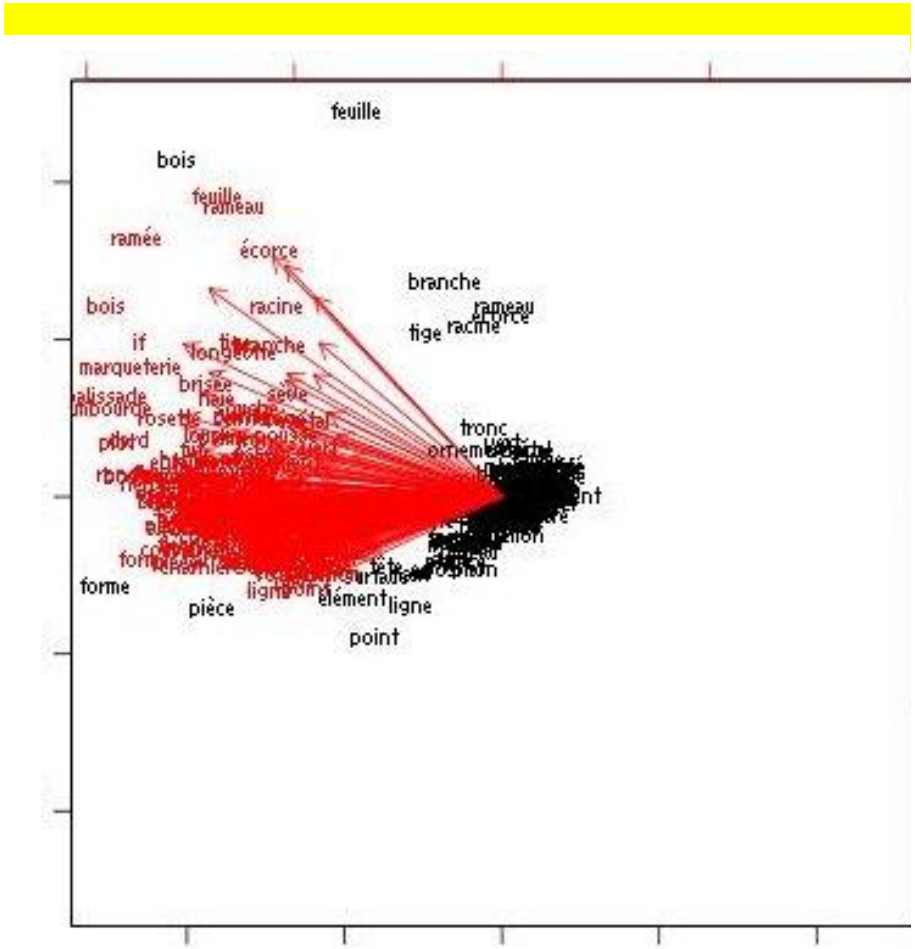

Figure 4: Analyse en composantes principales des distances entre les 200 premiers proxèmes du graphes de définition (individus en noir).

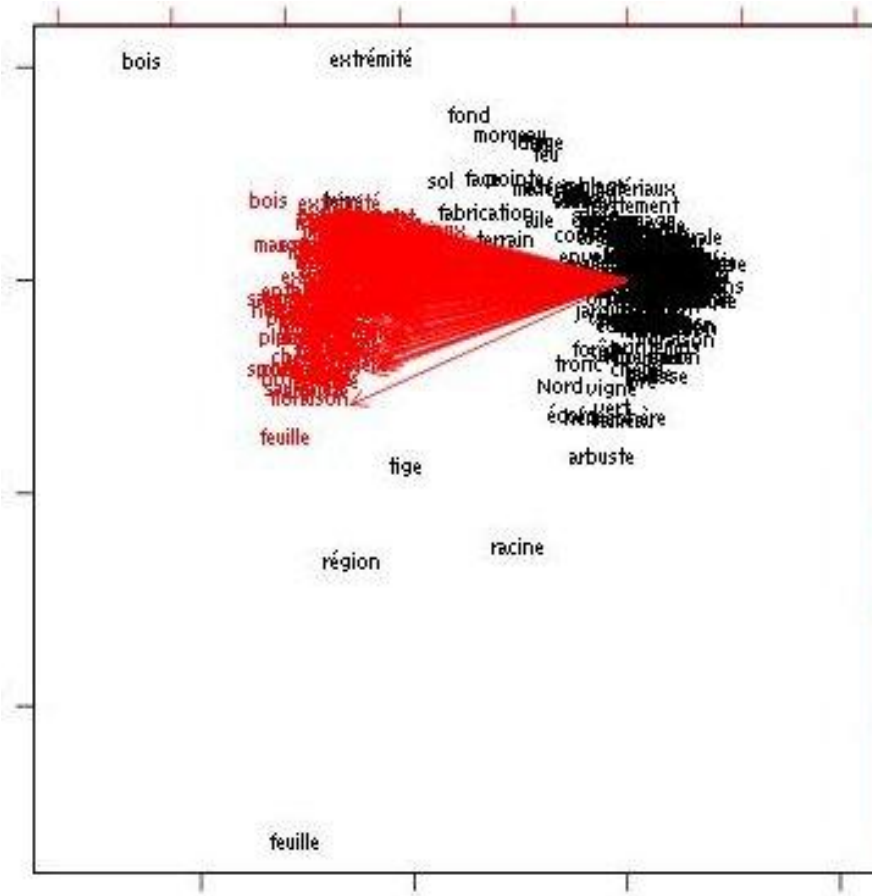

Figure 5: Analyse en composantes principales des distances entre les 200 premiers proxèmes du graphes de cooccurrences (individus en noir).

Les deux figures montrent clairement que les graphes gdg et gcg aboutissent à des structurations très différentes des proxèmes rattachés à *arbre*. Dans le cas du graphe de cooccurrences (Figure 5), le domaine de *arbre* se structure tendanciellement en deux sous-domaines, celui de ses méronymes (en bas : *feuille, racine, tige, tronc, rameau*) et celui de ses usages en tant que matériau (en haut : *bois, matériau, fabrication, morceau*...). Une analyse plus fine ferait peut être apparaître d'autres pôles, par exemple, au milieu du graphique, la proximité entre *floraison*, *printemps*, ou même *forêt*, indique peut-être un sous-domaine //naturalisme//. Le cas particulier de *bois*, qui figure dans le Tableau 1 mais pas dans le Tableau 2 (*bois* n'apparaît en effet qu'à la 71ème position), doit être souligné : il n'est pas conçu dans le cadre d'une relation partie-tout (plus précisément une relation matière-objet) avec *arbre* mais comme simple nom de matière. Cette

observation se trouve largement confirmée par les proxémies de *bois* dans gcg, où *bois* est exclusivement traité comme matière première :

menuiserie	vis	languette	pieu	tablette
planche	meuble	pointe	bordure	fond
rebord	tuyau	fer	semelle	attache
entaille	pourtour	largeur	plafond	revêtement
biseau	boule	anneau	seau	acier
assemblage	planchette	jet	traverse	liège
long	lame	toit	toile	sommier
armature	intempérie	métal	aiguille	vase
fût	cloison	cheminée	ustensile	coffre
trou	extrémité	flèche	cheville	carton

Tableau 4 : liste des mots les plus proches de *bois* dans le graphe de cooccurrences (gcg)

Ce résultat s'explique en partie lorsqu'on s'intéresse aux quelques 2190 gloses du TLF où *bois* apparaît. En effet, c'est fréquemment dans des structures binominales (*pièce de bois*, *morceau de bois*, *copeau de bois*, *tuyau de bois*, *châssis de bois*, *etc*.) où *bois* est bien pris comme nom de matière et non comme partie d'un arbre.

En revanche, il apparaît que dans le graphe de définition (Figure 4), *bois* et *feuille* sont voisins (alors qu'ils étaient opposés dans gcg). *Bois* n'est donc pas éloigné d'un sous-domaine portant sur des parties de l'arbre, avec *feuille, branche, rameau, écorce, racine, tige* (en haut) ainsi que *tronc*, légèrement plus bas. Le côté opposé s'interprète peut être moins facilement que dans le graphe précédent, mais ne relève pas, en tout cas, d'un domaine des artefacts matériels, mais plutôt de l'arbre comme figure (sémiotique graphique) : *point, ligne, surface, niveau, position*.

Ces résultats sont indicatifs et demanderaient des méthodes d'analyse plus élaborées. Ce que nous tenons pour acquis, en revanche, c'est que non seulement les listes de proxèmes diffèrent entre gdg et gcg, mais surtout, que la structuration lexicale en domaines et sous-domaines diffère entre les deux graphes, et que les différences entre les listes de lexèmes ne peuvent être interprétées qu'à la lumière de cette différence de structuration plus globale. De plus, il ne semble pas y avoir de raisons de privilégier un graphe plutôt que l'autre : ils donnent des résultats différents et complémentaires.

Le graphe des exemples (Tableau 3) donne lieu, quant à lui, à une liste de proxèmes assez distinctes des précédentes mais qui confirme, d'une certaine manière, la dimension notionnelle (sur laquelle nous revenons en détail dans la section suivante) de ces graphes. En effet, le domaine naturel / rural s'y trouve fortement représenté (*soleil, nuage, pluie, forêt, jardin, colline*) et se sous-spécifie en différents sous-thèmes : saisons (*hiver, printemps*), fonction associée (*ombre*). On peut aussi souligner la présence de *flamme.* Ces mots étendent donc les « inférences notionnelles » nettement plus loin que dans les graphes précédents. Ils incluent une dimension plus anthropologique que les listes précédentes, centrées autour des

constituants de réalités physiques. Il s'agit cette fois d'implications culturelles ou de connaissances scientifiques/culturelles (tel le rapport météorologie / arbre, qui intéresse typiquement un domaine davantage centré autour de l'agriculture). Ce contraste permet de caractériser les exemples, par oppositions aux gloses définitionnelles, comme mobilisant davantage savoir anthropologique et typicalité. Le graphe des exemples permet donc observer un contenu et une structuration différents des graphes précédents, sans qu'il n'y ait de raisons, à nouveau, de considérer l'un de ces graphes comme plus fondé sémantiquement.

4.2. Graphe de synonymie et graphe de dictionnaire : sémasiologie et onomasiologie

Les travaux portant sur les graphes de dictionnaire divergent sur les méthodes utilisées, puisque (Ploux & Victorri 1998) utilisent une procédure de recherche de cliques tandis que (Gaume 2004) utilise un algorithme de type marche aléatoire. Au-delà de cette différence méthodologique, il n'en reste pas moins que dans les faits, ces deux approches se fondent sur le même objet, à savoir le dictionnaire de synonymes dicoSyn, et surtout, qu'elles partagent une hypothèse de travail commune :

> « De manière générale, si les définitions d'un dictionnaire sont porteuses de sens, c'est au moins par le réseau qu'elles tissent entre les mots qui en sont les entrées. Notre propos est d'exploiter ce réseau de type petit monde en tirant parti de l'hypothèse que les zones de densité fortes en arcs (les agrégats) identifient des zones de sens proches. » (Gaume 2004)

Mais on peut aussi partir de l'idée que le dictionnaire est un objet sémiotique[14] et que si les définitions sont porteuses de sens c'est parce qu'elles sont des textes, et non des réseaux formels. Cette approche se trouve confirmée par le fait que le graphe de cooccurrences, qui a pour effet de court-circuiter toutes les entrées du dictionnaire, n'est pas moins structuré ni moins qualifiable sémantiquement que le graphe de définition. Cela étant dit, nous partageons aussi l'idée selon laquelle la connectivité d'un graphe de dictionnaire entretient, dans une large mesure, un rapport avec l'ordre du sémantique. Mais encore faut-il caractériser le plus précisément possible la nature de ce rapport.

Avec (Ploux & Victorri 1998), le travail sur les graphes de synonymie s'inscrit initialement dans une recherche plus générale sur la polysémie et sa modélisation sous la forme d'un espace sémantique. Dans ce cadre, chaque clique associée à un mot donné et telle qu'identifiée dans le graphe de synonymie correspond à un point de l'espace sémantique associé à ce mot. Grâce à un calcul de distance entre ces différents points, on peut ainsi obtenir une représentation bidimensionnelle de cet

---

[14] « [...] Les dictionnaire [sont] des objets sémiotiques et des textes d'une nature particulière [...] » (Rey 2008 : 52).

espace sémantique[15]. La relation de « proxémie » avancée par Gaume est assez différente en tant qu'elle est le résultat d'un algorithme de marche aléatoire. Elle se présente sous la forme d'une liste de mots classés en fonction d'une distance que Gaume réinterprète directement en termes sémantiques. Selon l'auteur, en effet, la proxémie organise des relations de superordination (hyperonymie / hyponymie) sur un continuum qui peut aller des emplois « intradomaine » (par ex., *éplucher – peler*, qui appartiennent tous les deux au domaine //culinaire//) jusqu'à la prise en compte d'emplois plus figurés (par ex. *éplucher – déshabiller*, qui n'appartiennent pas aux mêmes domaines), prise en compte que l'approche par détection de cliques ne permet pas.

Les deux approches, en tant qu'elles se fondent sur l'exploitation d'un graphe de synonymie, sont toutefois vouées à ne se mouvoir que sur un axe particulier du sens, à savoir l'axe de la polysémie. En effet, que l'on applique à un graphe de synonymie une méthode fondée sur la recherche de cliques ou sur une marche aléatoire, il n'en reste pas moins que ce sont bien différentes acceptions qui sont récupérées, qu'elles soient stabilisées dans le lexique ou qu'elles correspondent à des emplois plus novateurs et donc moins susceptibles d'être listés dans un dictionnaire. Or, depuis le début du 20ème siècle, la tradition sémantique donne précisément un nom à cette dimension : la sémasiologie. À la lumière de ce rapprochement, il est alors possible de caractériser ces différents travaux comme une procédure automatique de découverte du champ sémasiologique associé à une unité particulière.

Afin d'illustrer cela, nous exposons, dans le tableau qui suit, la liste des 50 termes associés à l'exemple conducteur de ce travail (*arbre*), classé par « proxémie » décroissante, et ce, pour le graphe de synonymie des noms (dicoSynNom)[16] :

plante	essieu	forêt	pédale	balancier
manivelle	diagramme	représentation	précipité	cheville
vilebrequin	pivot	broussaille	herbe	couple
tige	graphe	perceuse	manche	bras
axe	barre	fraise	graminée	centre
arborescence	maneton	liane	céréale	baril
bois	bogie	légumineuse	pied	barrique
végétal	fraisoir	foret	plant	futaille
arbuste	fût	futaie	tonneau	châssis
bielle	arbrisseau	nille	buisson	simple

15 Pour une illustration, le lecteur peut se reporter aux différentes études qui s'inscrivent dans ce cadre : (François & Manguin 2004) sur l'adjectif *propre*, (Ozouf 2004) sur le verbe *entendre*, (Venant 2004) sur l'adjectif *sec* ou (Venant 2008) sur *livre*, *etc.*

16 Cette liste est directement accessible sur le site du CNRTL : http://www.cnrtl.fr/proxemie/arbre.

Comme on peut le constater, cette liste présente les acceptions habituelles de *arbre*, à savoir : végétale (*plante*, *bois*, *végétal*, *arbuste*), mécanique (*manivelle*, *vilebrequin*, *axe*), représentationnelle (*arborescence*, *diagramme*, *graphe*). On retrouve aussi des sens fonctionnels plus originaux, comme celui qui se rattache au domaine de la chimie (*précipité*[17]) ou au domaine vinicole (*tonneau*, *baril*, *barrique*, *futaille*), mais qui n'en sont pas moins des acceptions. Sur le site du CNRTL, un calcul supplémentaire permet ainsi d'obtenir une représentation tridimensionnelle de ces différentes acceptions et de leurs distances, sur un modèle comparable – si nous avons bien compris – à celui des espaces sémantiques de Victorri.

Or, nos propres résultats, obtenus à partir d'un dictionnaire de langue comme le TLF et de sa transposition dans un graphe de cooccurrences, sont de nature très différente. Pour le comprendre, il suffit de comparer la liste précédente avec celle du Tableau 2 (proxèmes de gcg).

Le point de divergence le plus évident, c'est que les proxèmes de gcg gravitent autour d'une seule et unique acception, à savoir l'acception végétale. En revanche, si l'acception ne varie pas, les relations qu'entretiennent les mots de cette liste avec *arbre* sont au contraire très hétérogènes : on trouve ainsi des méronymes (*tronc*, *sève*, *feuillage*, *bourgeon*, voire des méronymes de méronymes avec *écaille*, *lamelle* ou *enveloppe*), des collectifs (*bouquet*[18], *forêt*), des hyperonymes (*végétation*, *végétal*) et des hyponymes (*hêtre*, *chêne*). On trouve aussi des activités typiquement associées aux arbres (*cueillette*, *menuiserie*), des lieux qui entretiennent un rapport de métonymie avec *arbre* (*jardin*, *parc*, *pré*), des fonctions traditionnellement rattachées aux arbres (*ornementation*), ainsi que des propriétés typiques (*maturité*, *vert*). Ces unités ne couvrent pas les différents sens de *arbre*. Au contraire, partant de la seule acception végétale, ces mots portent en fait sur différentes parties d'un domaine notionnel unique, celui des arbres. Il est donc légitime de rapprocher cette liste de ce que serait le champ onomasiologique du concept d'arbre. Et de même qu'il était possible de caractériser les travaux portant sur les graphes de synonymie comme une procédure automatique de découverte du champ sémasiologique associé à une unité particulière, notre propre travail peut ainsi se concevoir comme une procédure automatique de découverte du champ onomasiologique associé à un concept particulier.

Cette hypothèse originale trouve une confirmation intéressante lorsqu'on compare nos propres résultats avec les dictionnaires d'orientation onomasiologique, en particulier les dictionnaires analogiques. Dans ces derniers, en effet, et contrairement aux dictionnaires de langue où l'on part d'un signifiant (en entrée) pour en décrire tous les sens (dans l'article), on part d'un concept et l'on décrit la structure du domaine notionnel qui s'y rattache. Dans le cas du concept d'arbre, par

---

17 ARBRE : CHIM. Cristallisation, dépôts plus ou moins arborescents. Arbre de Diane (ou philosophique). Amalgame d'argent obtenu avec du nitrate d'argent et du mercure. Arbre de Jupiter. Étain précipité par le zinc. Arbre de Mars, de Saturne, etc. (TLF)

18 *Bouquet* est en effet marqué comme synonyme de *bosquet* dans le TLF.

exemple, le dictionnaire analogique structure le domaine notionnel par des sous-sections dans lesquelles on trouve, entre autres, la liste des parties (*tronc*, *branche*), le traitement (*arboriculture*, *élagage*), les types de plantation (*allée*, *bosquet*), les maladies (*gouttière*, *gélivure*), les principales espèces (*pin*, *chêne*), *etc*. Or, la moitié des termes du tableau précédent se trouvent justement sous l'entrée *arbre* dans le *Dictionnaire analogique* de Charles Maquet (Larousse) ou dans le *Dictionnaire analogique* de Georges Niobey (Larousse), alors que la proportion baisse considérablement dans le cas du graphe de synonymie, puisque ce n'est plus que le cas de 8 d'entre eux.

Les différences de structuration du domaine en fonction des graphes employés (gcg et gdg), que nous avons mises en évidence dans la section précédente avec *bois*, vont aussi dans le sens d'un tel rapprochement. En effet, si les proxémies de *arbre* peuvent être caractérisées comme couvrant un champ onomasiologique, rien n'empêche de penser qu'ils s'organisent aussi selon un modèle analogue à celui des dictionnaires analogiques. Si c'est le cas, il n'est pas non plus surprenant que certains mots, comme *bois*, soient susceptibles de se rattacher à des domaines différents.

5. Conclusion

Nous avons volontairement réduit notre exposé à un nombre limité de types de graphes. Notre objectif principal, en effet, était de caractériser précisément les graphes de dictionnaire de langue, en particulier en les opposant aux graphes de synonymes. Beaucoup d'autres graphes de dictionnaire de langue mériteraient une étude approfondie. Parmi les différentes possibilités que nous avons évoquées au cours de ce travail, la question des graphes pondérés (dans le cas des graphes de cooccurrence) constitue une direction de recherche intéressante. Mais il existe de nombreux autres procédés qu'il faudrait également comparer. Un premier axe de variation qu'il faudrait tester porterait ainsi sur la prise en compte de la microstructure. Un second axe de variation concernerait l'architecture du graphe proprement dite : si nous avons étudié le graphe de définition et le graphe de cooccurrences, nous n'avons rien dit des possibilités qu'offrirait un graphe orienté, où les relations défini-définissant et définissant-défini seraient différenciées. Un troisième axe de variation que nous souhaiterions développer concerne cette fois-ci différents types de dictionnaires. Nous avons abordé la question en évoquant le cas de *blockhaus* tel qu'il est définit dans le TLF et dans le Micro Robert. Il semble évident que de telles différences dans les définitions proposées ne peuvent avoir qu'un impact considérable sur les graphes résultant et leurs propriétés sémantico-lexicales. Les graphes de dictionnaires pourraient servir, ici, à la caractérisation des textes dictionnairiques et de leurs normes rédactionnelles.

Les dictionnaires onomasiologiques, et plus particulièrement les dictionnaires analogiques, jouissent généralement d'une assez mauvaise réputation. Le reproche principal qui leur est adressé porte sur la rigueur des critères de délimitation du

champ onomasiologique ainsi que sur les choix qui président à la structuration interne du champ conceptuel. Dans ce contexte, notre travail tient une place originale, dans la mesure où il permet, par un détour inédit (l'exploration d'un graphe de dictionnaire de langue, c'est-à-dire d'orientation avant tout sémasiologique), d'apporter une certaine justification à la pratique onomasiologique. En effet, comme nous espérons l'avoir montré avec le cas de *arbre*, les proxémies que nous obtenons, qu'elles soient issues d'un graphe de définition ou d'un graphe de cooccurrences, semblent bien correspondre à un champ onomasiologique. Ce rapprochement se trouve conforté par le fait que ces mêmes proxémies présentent une structuration interne qui n'est pas sans rappeler les organisations thématiques des sous-sections d'une entrée dans un dictionnaire analogique.

De plus, ce travail permet d'insister sur les propriétés textuelles du dictionnaire. Nous avons montré que plusieurs types de graphes, aux propriétés sémantiques différentes et complémentaires, peuvent être tirés d'un dictionnaire : on ne peut en tirer « une » modélisation sémantique univoque. Les graphes issus des différentes sections de la microstructure (graphe des définitions, graphe des exemples) permettent d'observer des contenus et des structurations différents et complémentaires. Ces différences peuvent être rapportées aux normes de la définition et de l'exemple en tant que genres textuels. Si les définitions tendent à neutraliser les implications culturelles et anthropologiques, les exemples permettent de les réintroduire. Cela n'a évidemment rien d'étonnant : les analyses synonymiques réalisées dans les définitions du TLF privilégient un seul type de relation sémantique, les relations « structurées », et excluent les relations culturelles. Néanmoins, le statut des exemples mérite d'être réexaminé : loin d'être une simple mise en discours, ils sont, plutôt, une autre façon de définir.

Bibliographie

Albert, R. & Barabási. A.-L. (2002) « Statistical mechanics of complex networks » *Review of Modern Physics*, 74, pp. 47–97.

Barabási, A.-L. & Albert R. (1999) « Emergence of scaling in random networks », *Science*, 286:509-512.

Blumenthal P. (2009) « Combinatoire des prépositions : approches quantitative », *Langue française*, 157, 37—51.

Clas A., Thoiron P., Béjoint H. (1996) *Lexicomatique et dictionnairique ? Actes du Colloque IVeme Journées scientifiques du réseau thématique « Lexicologie, Terminologie, Traduction »* Lyon, France, 28, 29, 30 septembre 1995, Paris : AUPELF-UREF.

Content A., Mousty P. & Radeau M., (1990) « Brulex, une base de donnée lexicales informatisée pour le français écrit et parlé », *L'année Psychologique*, 90, pp. 551–566.

Corbin P. (2006) *Avec des dictionnaires pour compagnons*, Habilitation à diriger des recherches, Lille 3.

Delesalle S. & Rey A. (éd.) (1979), *Langue française*. 43/1, « Dictionnaire, sémantique et culture ».

Dubois J. & Dubois-Charlier F. (1990) « Incomparabilité des dictionnaires », *Langue Française*, « Dictionnaires électroniques du français », 87, pp. 5–10.

Fox Keller E. (2005) « Revisiting "scale-free" networks », BioEssays, 27-10, pp. 1060—1068.

François, J. & Manguin, J.L. (2004), « La polysémie adjectivale entre synonymie et sélection contextuelle : le cas de propre ». In H. BOUILLON (éd.), *Langues à niveaux multiples. Hommage au Professeur Jacques Perrot à l'occasion de son éméritat* (BCILL 112). Louvain : Peeters.

Gaume B. (2004) « Balade aléatoire dans les petits mondes lexicaux » *I3 Information Interaction Intelligence*, 4(2).

Gaume B., Hathout N., Muller P. (2004) « Word Sense Disambiguation using a dictionary for sense similarity measure » in *Proceedings of the 20th International Conference on Computational Linguistics (COLING 20)*, Geneva, August 23-27 2004, pp. 1194–1200.

Habert B. (2000) « Création de dictionnaires sémantiques et typologie des textes » in Tyvaert J.-E. (éd.) *L'Imparfait, Philologie électronique et assistance à l'interprétation des textes*, Actes des Journées scientifiques 1999 du CIRLEP, Reims : Presses Universitaires de Reims, pp. 171– 188.

Hirst G., St-Onge D. (1998) « Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms » in Fellbaum C. *Wordnet - An electronic lexical database*, Cambridge : The MIT Press, pp. 305–332.

Loiseau S. (2006) *Sémantique du dicours philosophique chez Deleuze : du corpus aux normes*, Thèse de doctorat, Université Paris 10 Nanterre.

Martin R. (1992 [1983]) *Pour une logique du sens*, Paris : Presses Universitaires de France.

Muller P., Hathout N., & Gaume B. (2006) « Synonym extraction using a semantic distance on a dictionary ». NAACL workshop/Textgraphs.

Meschonnic H (éd.) (1991) *Des mots et des mondes, Dictionnaires, Encyclopédies, Grammaires, Nomenclature*, Paris : Hatier.

Newman M. E. J. (2003) « The structure and function of complex networks », SIM Review, 45, 167-256.

Ozouf, C. (2004) « Polysémie lexicale et représentation géométrique du sens : l'exemple du verbe entendre », *Corela*.

Pantel P. & Lin D. (2002) « Discovering word senses from text », *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 613—619.

Ploux S. & Victorri B. (1998) « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », *TAL*, 39/1, pp. 161—182.

Ploux S. & Ji H. (2003) « A Model for Matching Semantic Maps Between Languages (French / English, English / French) », *Computational Linguistics*, 29(2), pp. 155– 178.

Rastier (1987) *Sémantique interprétative*, Paris : PUF.

Rey A. (2008) De l'artisanat des dictionnaires à une science du mot : Images et modèles, Paris : Armand Colin.

Valette M., Estacio-Moreno A., Petitjean E., Jacquey E. (2006) « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques . Pour une approche sémique du sens » *in* Piet Mertens, Cédrick Fairon, Anne Dister, Patrick Watrin (éds), *Verbum ex machina , Actes de la 13ème conférence sur le traitement automatique des langues naturelles*, Volume 1. Pages 357-366.

Véronis J. (2004) « HyperLex: Lexical Cartography for Information Retrieval », *Computer Speech and Language*, « Special Issue on Word Sense Disambiguation », Preiss J., Stevenson M. (éd.), 18-3, pp. 223–252.

Venant F. (2004) « Géométriser le sens », Actes de la 11e conférence sur le Traitement automatique des langues naturelles (TAL-RECITAL 2004), Fès, Maroc, 19-21 avril 2004.

Venant F. (2008) « Représentation géométrique et calcul dynamique du sens lexical : application à la polysémie de livre », *Langages*, 172 : 30-54.

Watts J. W. & Strogatz S. H. (1998) « Collective dynamics of "small-world" networks », *Nature* 393, pp. 440–442.

Widdows D. (2004) *Geometry and Meaning*, Chicago : University of Chicago Press.