SCHRYVER Gilles-Maurice de (dir.), ***A Way with Words : Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks***, Kampala, Menha Publisher, 384 p. (coll. « Menha Linguistics Series », 2010) – ISBN 978-99-7010-101-6.

L'ouvrage est un hommage à Patrick Hanks, linguiste et lexicographe. Il rassemble des articles signés de chercheurs reconnus par la communauté internationale dans le domaine du lexique (lexicologues et lexicographes)[2]. Le volume s'ouvre sur une introduction de G.-M. Schryver, éditeur scientifique du volume. En quelques pages, il rappelle les apports majeurs des travaux de Patrick Hanks. À la suite de John Sinclair, P. Hanks a opté pour une lexicographie de corpus, affirmant que le sens des mots existe, mais en contexte seulement, qu'il consiste en un groupe d'éléments sémantique, et que les dictionnaires ne décrivent que des potentiels sémantiques et non pas des sens. Les publications les plus célèbres de Hanks sont celles qui tournent autour de l'idée que des mesures statistiques sur les collocations sont des outils fondamentaux pour décrire le sens des mots. L'article se poursuit ensuite par un historique de la carrière de P. Hanks, qui l'a amené à la théorie des normes et exploitations (Theory of Norms and Exploitations). Il établit ainsi que deux systèmes de règles régissent les comportements linguistiques : l'un gouverne la phraséologie normale et les mots en usage, l'autre permet aux locuteurs l'utilisation de la phraséologie normale dans des utilisations plus créatives. Pour clore l'article, Shryver revient sur les publications de Hanks.

L'ouvrage se divise en trois grandes parties : il présente d'abord les aspects théoriques de la lexicographie, puis les aspects computationnels de la discipline, et consacre la dernière partie à l'analyse lexicale et à la rédaction de dictionnaires.

Première partie : Aspects théoriques

John Sinclair est généralement reconnu comme le père fondateur de la lexicographie moderne qui s'appuie sur une théorie de la linguistique de corpus. L'article « Defining the definiendum », qui ouvre le volume, est resté inachevé. La version publiée est datée d'à peine deux semaines avant sa mort et n'a pas été modifiée. Cet article essaie de montrer ce que l'on devrait entendre par « définir ». Pour J. Sinclair, définir, c'est spécifier les limites, les frontières d'une unité lexicale plus que décrire son sens. Aussi, l'auteur, en s'appuyant sur la linguistique de corpus, donne toute leur place aux expressions idiomatiques, et considère qu'un *definiendum* doit être un item lexical (et non un mot), c'est-à-dire un segment de texte à laquelle on attribue un sens. Les études de corpus montrent qu'il est absurde de considérer que le sens généralement considéré par les dictionnaires comme le sens principal (*core sense*) d'un item lexical est le sens le plus fréquent. En effet, les locuteurs (sans doute sous l'influence) des dictionnaires) considèrent le sens concrets des items lexicaux comme le sens principal, ce qui n'est pas le cas si on observe l'usage. Ensuite, Sinclair montre comment on peut décrire le sens d'items lexicaux en utilisant des grammaires locales énumérant les constructions dans lesquelles les mots apparaissent et prennent un sens différent du sens habituellement considéré comme principal.

Yorick Wilks republie une communication des années 1970 sur la préférence sémantique : « Very Large Lexical Entries and the Boundaries Between Linguistic and Knowledge Structure », dans la lignée des travaux de Minsky (1975).

---

2 La distinction entre lexicologie et lexicographie est rarement observée par les linguistes de langue anglaise, et ces deux versants sont convoqués au même titre dans ces Mélanges.

Le problème posé est celui de la quantité d'informations que les entrées lexicales d'un système de compréhension automatique du langage doivent contenir, et surtout, celui de la gestion des emplois dont le sens va au-delà de ce qui est habituellement décrit dans un lexique. Wilks montre, en utilisant l'exemple de la phrase « Ma voiture boit beaucoup d'essence », comment la violation d'une contrainte sémantique (ici le fait que le sujet du verbe *boire* n'est pas un animé, comme on l'attend en général) peut être gérée par un système automatique en utilisant un « pseudo-texte » (PT).

L'article de James Pustejovsky et Anna Rumshisky, « Mechanisms of Sense Extension in Verbs », étudie les mécanismes qui permettent de relier les différents sens d'un prédicat. Quel que soit le degré de métaphorisation des différents sens, la manière dont ils sont reliés entre eux est un objet d'étude pertinent. La coercition de type[3], telle qu'elle est décrite dans la théorie du lexique génératif, n'est pas le seul moyen de décrire les extensions de sens des prédicats. Les différents sens proviennent de divers processus opérant sur les prédicats, tels que la généralisation du type des arguments, le changement dans la proéminence relative des arguments ou l'abstraction du sens noyau du prédicat lui même. Tous ces phénomènes sont décrits au travers d'exemples qui permettent de voir clairement tous les sens intermédiaires entre le plus concret et le plus abstrait.

Igor Mel'čuk, dans « The Government Patterns in the Explanatory Combinatorial Dictionary », revient sur certains aspects de la théorie Sens-Texte, et sur des notions fondamentales liées aux dictionnaires explicatifs et combinatoires qu'il a développés dans les dernières décennies. Il explique plus particulièrement les notions de gouvernement et de patron de gouvernement (Government Pattern). Une unité u1 gouverne une unité u2 :

- soit si la forme morphologique de u2 dépend de propriétés intrinsèques de u1 (et non de sa forme) ;
- soit si u1 sélectionne un lexème u2 (et non un grammème) ;
- soit si u1 sélectionne une unité lexicale u2 d'un type grammatical particulier.

Un patron de gouvernement contient alors quatre éléments : la diathèse de l'unité lexicale ; des éléments de structure syntaxique de surface correspondant aux actants de structure profonde ; des moyens linguistiques d'expression de surface des actants syntaxiques profonds ; et enfin des contraintes sur l'expression des actants syntaxiques de surface.

David Wiggins, dans son article « The Paradox of Analysis and the Paradox of Synonymy », revient sur un problème connu : si on peut remplacer sans conséquence un mot (*analysandum*) par une périphrase (*analysant*), alors l'analyse est triviale. S'il y a une différence, alors l'analyse est fausse. Concernant la synonymie, le problème est proche. Wiggins revient alors sur le problème posé par Frege à propos de l'étoile du matin qui est aussi l'étoile du soir, et conclut après avoir expliqué ses relations avec P. Hanks : on n'apprend pas à quel concept correspond un mot, mais on cherche un mot pour exprimer un concept. Le paradoxe de la synonymie montre que chaque mot qui entre dans la langue crée sa propre niche.

Deuxième Partie : Calculer les relations lexicales

L'article de Kenneth W. Church, « More is More », pose le problème de la « qualité » des corpus. Doivent-ils être calibrés ? Doit-on les « nettoyer » ? Aujourd'hui, malgré les protestations de certains, comme Adam Kilgarriff, le choix a été fait : on privilégie la

3 Opération qui consiste à forcer et donc à modifier le type sémantique de l'argument d'un prédicat dans certains cas de polysémie.

quantité et non la qualité. Les corpus doivent être volumineux, mais pas forcément échantillonnés ou nettoyés, et pour Church, ce choix est raisonnable : « plus, c'est plus, malgré les critiques de la Google-ologie[4] ».

Gregory Grefenstette s'intéresse, dans l'article « Estimating the Number of Concepts », aux expressions composées de plusieurs mots. En effet, si les outils de traitement automatique des langues peuvent gérer les mots simples, ils ne sont souvent pas au point concernant la phraséologie, les mots composés, les expressions figées. Les concepts exprimés par plusieurs mots sont en nombre inconnu. Après avoir décrit la méthode avec laquelle il analyse les pages web, et en montrant les réserves à faire la concernant, G. Grefenstette estime le nombre de concepts exprimés par des expressions de deux mots à environ 233 millions.

L'article de David et Louise Guthrie, « Identifying Adjectives that Predict Noun Classes », montre l'utilité des adjectifs dans une tâche de prédiction de la catégorie sémantique du nom qu'ils modifient. Les auteurs utilisent un ensemble de 29 classes de noms, et augmentent les corpus existants grâce à une technique non supervisée d'annotation des syntagmes nominaux. Il est clairement démontré que l'utilisation de très gros corpus est fondamentale ici, et il apparaît que l'étude des adjectifs entourant le nom permet de les désambiguïser efficacement.

Alexander Geyken étudie, dans « Statistical Variation of German Support Verb Construction in Very Large Corpora », des constructions à verbe support en allemand, dites « constructions verbes-nominalisation » (désormais NVG). Il essaie, en étudiant de très grands corpus, de savoir si (1) le nombre de NVG augmente avec la taille des corpus étudiés (autrement dit, si la classe est potentiellement infinie) et (2) ce que donne la comparaison entre les NVG trouvées en corpus et celles qu'on trouve dans les grands dictionnaires « papier ». L'étude montre qu'un corpus doit compter au moins 100 millions de mots pour permettre une réelle observation et constituer la base d'une bonne étude lexicographique : en dessous, on n'y trouve pas de façon significative les NVG recensées dans les dictionnaires. L'étude montre aussi qu'une étude de corpus est indispensable pour compléter les entrées des dictionnaires imprimés.

L'article de Karel Pala et Pavel Rychlý, « A Case Study in Word Sketches – Czech Verb vidět "see" », étudie le schéma (*sketch*) du verbe tchèque *voir* (*vidět*) en corpus. Les schémas ont pour but d'aider les lexicographes à relever en corpus les cooccurrences des unités lexicales afin de décider lesquelles inclure dans un dictionnaire. À partir d'un travail réalisé avec le Sketch-Engine de Kilgarriff[5] et du résultat obtenu pour ce verbe, soit un document d'une page, contenant le schéma du verbe étudié, Pala et Rychly tirent les enseignements des erreurs contenues dans ce schéma. Le premier type d'erreurs provient du mauvais étiquetage du corpus, le second de règles de grammaires manquantes ou incorrectes (parfois, de la combinaison des deux facteurs). Il semble donc nécessaire d'améliorer les performances des étiqueteurs en augmentant la taille des données d'entraînement et en combinant des méthodes statistiques et symboliques.

Dans leur article « The Lexical Population of Semantic Types in Hanks's PDEV », Silvie Cinková, Martin Holub et Lenka Smejkalová rapportent une analyse du Pattern

4 *google-ology* dans le texte.

5 A. Kilgarriff est l'un des plus grands noms de la lexicographie internationale. Il est reconnu pour ses travaux à l'intersection de la linguistique de corpus, la lexicographie et la linguistique informatique.

Dictionary of English Verbs (PDEV), réalisé – entre autres – par Patrick Hanks. Le PDEV a pour but de faire correspondre des sens à des schémas d'usage des verbes. L'idée du travail décrit ici est d'assigner des types sémantiques aux noms qui cooccurrent avec les verbes. Le résultat permet d'associer des verbes au type sémantique le plus probable. Il permet aussi de repérer les désaccords entre annotateurs, et à terme, il permettra d'augmenter le PDEV.

Elisabetta Jezek et Francesca Frontini, dans l'article « From Pattern Dictionary to Patternbank », décrivent le lien entre les types sémantiques (TS) associés aux arguments des verbes et leurs définitions (LS, pour *lexical sets*). En analysant les discordances entre les TS et les LS, les auteurs proposent d'étendre la méthode élaborée par P. Hanks, la Corpus Pattern Analysis (CPA), et l'appliquent à la construction d'une « patternbank » pour l'italien, c'est-à-dire une base de données dans laquelle on trouve les schémas des différents verbes (informations sémantiques et syntaxiques sur les arguments des verbes).

Troisième partie : Analyse lexicale et rédaction de dictionnaires

Dans « Words that Spring to Mind : Idiom, Allusion, and Convention », Rosamund Moon présente une étude en corpus de l'expression anglaise *Spring to mind* (venir à l'esprit). L'auteur décrit le traitement lexicographique de cette expression, en classe les différentes utilisations et en commente certains usages. Elle conclut par les généralités suivantes, en accord avec les théories de P. Hanks : les mots et les expressions n'ont pas seulement le sens que décrivent les dictionnaires ; même leurs occurrences a priori déviantes doivent être prises en compte car elles font partie des utilisations des unités lexicales et permettent la description complète de schémas d'usage.

Sue Atkins, dans l'article « The DANTE Database : Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data », décrit les deux bases lexicales, DANTE (Database of Analysed Texts in English), qui contient environ 42 000 entrées lexicales, et Framenet, qui en contient environ 10 000. Elle note qu'une partie des informations contenues dans les deux bases se recoupent (unité lexicale, catégorie grammaticale, expressions avec des verbes support…), et que les autres informations pourraient se compléter (collocations, par exemple, pour DANTE et rôles sémantiques pour Framenet), et conclut qu'il semble possible (et utile), de trouver un moyen d'unifier automatiquement ces deux bases de données afin d'en créer une seule, plus complète.

L'article d'Adam Kilgarriff et Pavel Rychlý, « Semi-Automatic Dictionary Drafting », revient sur la théorie « Norms and Exploitation » de P. Hanks et montre qu'elle peut être utilisée, avec un grand corpus, afin d'améliorer les techniques automatisées de désambiguïsation des mots. Les auteurs présentent un logiciel, SADD, allant dans ce sens.

Paul Bogaards pose la question de l'existence d'une théorie lexicographique dans l'article « Lexicography : Science without Theory ? ». Il semble en effet communément admis qu'il n'existe pas réellement de théorie en lexicographie, même si la création d'un dictionnaire nécessite une théorie linguistique, une théorie psycholinguistique et une théorie de l'information, afin que les données lexicales soient décrites de façon à être comprises par le lecteur. Cela rend la lexicographie dépendante d'autres disciplines, ce que Bogaards ne considère par forcément comme souhaitable. Finalement, selon lui, peu importe les théories utilisées si le but reste le même : écrire de bons dictionnaires.

Mirosław Bańko, dans « The Polish COBUILD and its Influence on Polish Lexicography », revient sur les origines du COBUILD polonais (ISJP). Il raconte comment, après avoir découvert le COBUILD anglais dont P. Hanks est à l'origine, il a travaillé de façon à produire le même type d'ouvrage pour le polonais. Il revient sur les écueils à éviter,

et pour finir, expose ce qui différencie le COBUILD de P. Hanks et le ISJP, particulièrement le fait qu'il soit destiné à des locuteurs natifs.

Jonathon Green republie un article de Critical Quaterly : « ARGOT : The Flesh Made Word ». Il considère que cet article est celui qu'il aurait écrit s'il avait pu satisfaire à la demande de P. Hanks, qui lui avait demandé une histoire de l'argot français pour son *Encylcopédie du langage et de la linguistique* (*Encyclopedia of language and Linguistics*). Il revient sur l'étymologie incertaine du mot, sur ses premières attestations, et son histoire du XV^e^ siècle à aujourd'hui, même si certains affirment qu'il n'a pas survécu à la seconde guerre mondiale.

L'article de Michael Rundell, « Defining Elegance », porte sur la notion d'élégance en lexicographie. Il revient sur le sens du mot, plutôt positif, son étymologie, et son importance dans tous les domaines, même dans les sciences dites « dures ». En lexicographie, un dictionnaire doit être à la fois concis, informatif et compréhensible pour être qualifié d'élégant. Rundell revient alors sur ce que signifie élégance du point de vue de la microstructure et de la macrostructure du dictionnaire, et termine par des éléments plus généraux : le fait, entre autres, que la vision des dictionnaires a changé avec l'importance grandissante des dictionnaires électroniques, et que l'élégance consiste à tenir compte du fait que l'utilisateur n'est pas linguiste.

Dans l'ensemble, cet impressionnant ouvrage aborde tous les domaines qui touchent au lexique, grâce à des articles très théoriques ou au contraire très pratiques et concrets. Il constitue par conséquent un ouvrage de référence pour des lecteurs déjà aguerris dans le domaine de la lexicologie/lexicographie.

Hélène MANUÉLIAN
LDI (UMR 7187)
Université de Cergy-Pontoise et Université Paris 13 Nord
helene.manuelian@u-cergy.fr

L'auteur du compte rendu remercie John Humbley de lui avoir transmis les notes qu'il avait prises sur cet ouvrage.