Quelques problèmes observés dans l'élaboration de dictionnaires à partir de corpus

Alexander Geyken
Académie des Sciences de Berlin (Berlin-Brandenburgische Akademie der Wissenschaften)

Résumé : Ce travail s'interroge sur les critères quantitatifs et qualitatifs qui président à la construction de corpus électroniques (élaborés dans le cadre de la confection ou de la mise à jour de dictionnaires). Il questionne les notions de corpus équilibrés et opportunistes. Il est montré qu'il existe des phénomènes linguistiques intéressants qui ne sont pas présents dans les plus grands corpus équilibrés disponibles actuellement. Grâce à la disponibilité de journaux sous forme électronique les corpus opportunistes arrivent à une taille considérablement plus grande. Toutefois, différentes explorations à propos du genre et des archaïsmes, etc. montrent que les résultats varient de façon significative en fonction de la taille et de l'échantillonnage : la fréquence n'est donc plus un critère fiable, ce qui pose un problème quant à l'objectivité des corpus opportunistes.

Abstract : This work investigates the quantitative and qualitative criteria that preside over the construction of electronic corpora in the context of the elaboration or the update of dictionaries. In particular the concepts of balanced and opportunistic corpora are addressed. It is shown that there are interesting linguistic phenomena that are not present in the largest balanced corpora currently available. Opportunistic corpora are many times bigger due to the availability of large quantities of electronic newspaper text. However, different studies conducted e.g. on the gender distribution or on archaisms show that the results vary considerably depending on the size and the sampling of the corpora. Hence, frequency is no longer a reliable criterion which poses a problem for opportunistic corpora with regards to their objectivity.

0. Introduction

Un des domaines dans lesquels, dès le départ, les corpus électroniques ont joué un rôle important est l'élaboration de dictionnaires. A la différence de la linguistique générale, la lexicographie, dans sa méthodologie, a toujours été une discipline empirique dans le sens où elle s'est toujours appuyée sur des énoncés attestés. Des grands corpus de citations, permettant d'élaborer de plusieurs centaines de milliers à plusieurs millions de fiches, ont ainsi servi de base à la construction des entrées des dictionnaires. C'est le cas par exemple des grands dictionnaires monolingues comme le Oxford English Dictionary (OED), le Littré ou le Wörterbuch des frères Grimm (DWB). D'un autre côté, l'introspection a toujours coexisté avec l'empirisme, non seulement pour évaluer des phrases issues des corpus mais aussi pour compléter la description lexicographique quand il n'y avait pas d'exemples attestés. On est ainsi amené à se demander ce qu'apportent les grands corpus électroniques par rapport aux « ressources traditionnelles » des lexicographes. Commençons tout d'abord par la caractérisation du terme « corpus ». John Sinclair, un des fondateurs de la « Corpus Linguistics », a défini la notion de corpus de la façon suivante : « A corpus is a collection of *pieces of language* that are selected according to explicit linguistic criteria in order to be used as a sample of the language » (Sinclair 1994 : 4). Cette caractérisation est étendue par Habert (2000 : 13) : « un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extralinguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue. » Il est important de noter ici que la sélection des données langagières repose non seulement sur des critères linguistiques comme la richesse du vocabulaire ou la variabilité syntaxique, mais aussi sur des critères

extralinguistiques, par exemple le choix de types de textes constituant un corpus (par ex. Clear 1992, Biber 1994). Selon les deux caractérisations données ci-dessus, ce qui différencie les corpus des collections arbitraires, comme le web, est le fait que les corpus sont constitués sur la base de critères explicites. De plus, un corpus est censé être un échantillon ; autrement dit, il tente d'être représentatif de l'usage effectif de la langue, de telle sorte que l'on puisse établir des généralisations sur son fonctionnement (par ex. Jacques 2005). Comme il est impossible de mesurer ou de vérifier qu'un corpus est représentatif – il faudrait en effet connaître les proportions d'usage des genres dans la langue considérée –, on affaiblit la contrainte de la représentativité, et on la remplace par la notion d'équilibrage, *balancedness*, par rapport aux types de textes (Kilgarriff et Grefenstette 2003). Le British National Corpus (BNC) est sans doute l'exemple le plus connu d'un tel corpus : les types de textes les plus importants y sont représentés de manière équilibrée (Rundell 2002).

Actuellement, les corpus électroniques ont pris une grande importance dans l'élaboration des dictionnaires. Pour la grande majorité des dictionnaires actuels, les corpus électroniques servent à presque tous les niveaux de la description lexicographique. Ainsi le Trésor de la Langue Française (TLF), en plus des fiches lexicographiques de l'IGLF (Inventaire général de la langue française), doit sa richesse en exemples surtout au corpus électronique Frantext (Gorcy 1992). Le New Oxford Dictionary of English a été essentiellement rédigé à partir du BNC (NODE, préface). Dans certains cas, les corpus sont devenus même la ressource unique pour la rédaction du dictionnaire : ceci est le cas des dictionnaires Cobuild qui ont été rédigés exclusivement à partir de corpus électroniques (Sinclair 1996b).

Selon les définitions de corpus rappelées ci-dessus, les corpus se veulent un échantillon représentatif de l'usage de la langue. La fréquence d'un énoncé dans le corpus devrait donc indiquer un certain rapport à l'usage : plus un énoncé est fréquent plus il est usuel, et vice-versa, moins il apparaît dans le corpus moins il est usuel. Autrement dit, il y a deux présupposés sous-jacents à la représentativité : (a) si un mot/une expression fait partie de la langue, il/elle doit apparaître dans le corpus ; et de même pour la contraposée : si le mot/l'expression n'apparaît pas, il/elle n'existe pas ou n'est du moins pas important ; (b) la fréquence d'un mot/d'une expression dans le corpus est le reflet de sa fréquence réelle dans la langue. De nombreuses publications concernant la construction de dictionnaires de langue générale témoignent de cet usage des corpus (par ex. Kilgariff 2002, Rundell 2002). Il semble alors que les fiches lexicographiques ou les informateurs non officiels des lexicographes soient désormais dépassés et qu'ils doivent être entièrement remplacés par les corpus.

Cet article analyse ces présupposés. Dans quelle mesure peut-on affirmer que les corpus actuels fournissent une base empirique suffisante en ce qui concerne les propositions (a) et (b) ci-dessus ? Pour répondre, nous donnerons d'abord quelques chiffres pour illustrer le rapport entre la taille d'un corpus et du nombre d'entrées d'un dictionnaire (§ 1). Nous montrerons ensuite qu'il y a des phénomènes linguistiques intéressants qui ne sont pas présents dans les plus grands corpus équilibrés disponibles actuellement (§ 2). Les grandes bases de journaux sous forme électronique ont permis récemment la création de corpus d'un ordre de grandeur supérieur au BNC. Avec ces corpus dits opportunistes, on évite certes en partie le problème de la rareté de données, mais en revanche le critère de fréquence (cf. (b) ci-dessus) n'est plus un critère fiable, ce qui pose des problèmes méthodologiques pour la rédaction de dictionnaires (§ 3). On est alors amené à se demander si ces problèmes ne disparaîtront pas lorsqu'on disposera de corpus équilibrés de taille suffisante (§ 4.). Toutefois la question de savoir quelle taille de tels corpus devraient avoir n'a pas de réponse simple, du moins sur la base des corpus actuellement disponibles. En particulier, nous montrerons en nous fondant sur une étude que nous avons menée sur de très grands corpus que l'accroissement du vocabulaire, même s'il

devient plus modeste au fur et à mesure qu'on ajoute des nouveaux textes, ne semble pas converger.

1. Corpus et dictionnaires : une comparaison quantitative

Nous traitons ici de la question de la comparaison du nombre d'entrées des dictionnaires et de la taille des corpus. La taille des dictionnaires est déterminée par le nombre d'entrées annoncées ; dans le cadre du travail présenté ici nous ne faisons pas la distinction entre entrées et sous-entrées, car cette différence alourdirait davantage l'argumentation sans pour autant en changer les lignes principales. La taille des corpus est généralement déterminée par le nombre de *tokens* (c'est-à-dire une chaîne de caractères entre deux blancs) ainsi que par le nombre de *types* (c'est-à-dire des tokens différents dans le corpus). Si on regarde les tableaux 1 et 2, force est de constater que les corpus de « première génération » comme le corpus Brown (cf. l'article de Jacqueline Léon dans ce numéro) ou son pendant allemand, le corpus LIMAS (Hausser 1998 : 3) sont, avec leur million de tokens et respectivement 50 000 et 110 000 types, beaucoup trop petits pour pouvoir entrer en compétition avec un grand dictionnaire monolingue, sans parler même du fait que le nombre de types n'est pas en soi comparable avec le nombre d'entrées de dictionnaires, un sujet que nous allons aborder plus loin.

Nom du corpus	Nombre de tokens	Nombre de types
Brown Corpus (angl.)	1 million	50 000
Limas Corpus (all.)	1 million	110 000
BNC (angl.)	100 millions	650 000
DWDS (all.)	100 millions	2,2 millions
DWDS-E (all.)	1 milliard	9 millions

Tableau 1

Dictionnaire	Nombre d'entrées	Langue
Littré	80 000	français
Duden-GWB	200 000	allemand
DWB	297 000	allemand
OED	500 000	Anglais

Tableau 2

Les rapports de grandeur s'inversent pour le BNC qui, avec ses 650 000 types, dépasse le nombre d'entrées du OED. Les différences sont encore plus marquées si on observe les corpus allemands qui contiennent beaucoup plus de types que les corpus anglais (voir tableau 1 ci-dessus), et ce à cause du phénomène de la composition en allemand. C'est ainsi que le corpus DWDS, un corpus de référence de la langue allemande du XX^{e} siècle, développé à l'Académie des Sciences de Berlin et du Brandebourg, avec ses 2,2 millions de types (Geyken 2007) dépasse de plus de sept fois le nombre d'entrées du plus grand dictionnaire monolingue allemand, le dictionnaire des frères Grimm (Deutsches Wörterbuch - DWB). Un autre corpus, basé principalement sur des textes de journaux des grands quotidiens et hebdomadaires nationaux (entre autres : Frankfurter Allgemeine Zeitung (faz), Frankfurter Rundschau (fr), Neue Zürcher Zeitung (nzz), Spiegel, Süddeutsche Zeitung (sz), die tageszeitung (taz), Die ZEIT), lui aussi établi à l'Académie des Sciences, le DWDS-E (DWDS étendu) contient plus de 9 millions de types, soit quatre fois plus que le DWDS. Si d'un côté cela tend à démontrer que l'augmentation du nombre de tokens engendre l'augmentation du nombre de types (cf. § 4.1.), de l'autre il importe de savoir comment comparer ce grand nombre de types avec

le nombre d'entrées d'un dictionnaire : dans le cas du DWDS-E, le nombre de types est tout de même 30 fois plus grand que le nombre d'entrées du dictionnaire DWB.

En regardant de près les différents types du corpus, on s'aperçoit rapidement que la notion de *type* recouvre quelque chose de totalement différent de la notion d'entrée. Les *tokens* ainsi que les types sont, comme on vient de le voir, des chaînes de caractères séparées par des blancs ; leur nombre est donc le résultat d'un comptage informatique. Afin de faciliter la comparaison entre dictionnaires et corpus, nous utilisons par la suite le terme *mot-forme (*angl. *word-form)* pour désigner des *tokens* analysables morphologiquement et nous utilisons *lexème* pour désigner une forme regroupant des mots-formes qui ne se distinguent que par leur flexion (Polguère 2003). En analysant le DWDS-E on voit que ce corpus contient, à côté des mots-formes de l'allemand, des mots-formes d'autres langues (souvent sous forme de citations), des chiffres, des dates, des noms de marque, etc. Le corpus contient également un nombre important de noms propres, qui ne sont pas candidats à constituer une entrée d'un dictionnaire de langue générale. Étant donné l'importance de la composition en allemand se trouvent également de nombreux mots composés qui, compte tenu de leur transparence sémantique, n'apparaissent pas nécessairement dans des dictionnaires. Par exemple, le DWDS-E contient différents composés incluant *Tür* 'porte', dont des composés transparents considérés comme inintéressants d'un point de vue lexicographique : *Badezimmertür* 'porte de la salle de bains', *Holztür* 'porte en bois', *Stahltür* 'porte en acier', *Schlafzimmertür* 'porte de la chambre à coucher', *Stalltür* 'porte de l'écurie', *Wohnungstür* 'porte de l'appartement', *Wohnzimmertür* 'porte du salon', *Zimmertür* 'porte de la chambre', etc. Une fois que l'on a éliminé toutes ces formes, on peut se demander si les corpus comportent encore du matériel intéressant d'un point de vue lexicographique (cf. § 2. et 3.).

A ce problème des mots composés transparents s'ajoute le problème de la rareté des occurrences. Selon la Loi de Zipf la majorité des types d'un corpus apparaissent très rarement. Comme on considère généralement que des mots-formes ayant un faible nombre d'occurrences ne doivent pas être pris en compte dans l'élaboration du dictionnaire, le nombre de types susceptibles à être intégrés dans un dictionnaire diminuent davantage. Pour illustrer les conséquences de cet argument, nous avons effectué le comptage des types du corpus DWDS-E : plus de la moitié des types du DWDS-E ne sont ainsi attestés qu'une seule fois, 15 % des types sont attestés 2 fois, 6 % des types sont attestés trois fois, etc. (tableau 3). Toutefois, même si l'on supposait qu'on ne doive considérer comme base, pour la constitution d'un dictionnaire de langue générale, que des entrées, dont les types correspondants sont attestés plus de 10 fois dans le corpus – ce n'est généralement pas le cas, car on écrit des articles sur la base de moins d'exemples – alors le nombre de types s'élève toujours à plus d'un million, soit trois fois plus que le nombre d'entrées du dictionnaire DWB.

Nombre de types	**Nombre d'occurrences**
5 378 322	1 fois
1 183 751	2 fois
532 415	3 fois
315 535	4 fois
1 036 590	>=10 fois

Tableau 3 : effectifs du corpus DWDS-E

On peut donc bien supposer que des très grands corpus révèlent des lacunes lexicographiques dans les dictionnaires. Nous traitons cette question plus loin (§4). Au centre de notre travail est la question si les corpus peuvent être la base unique pour des grands dictionnaires monolingues, une question à laquelle nous tentons de répondre par la suite.

2. Corpus équilibrés

2.1 Mots simples ou composés

On affirme parfois dans la littérature (par ex. Hausser 1998, Teubert 2004) que les grands dictionnaires monolingues contiennent toujours des entrées qui ne sont pas présentes dans les corpus – ni en tant que mot-forme, ni en tant que lexème – et que, de ce fait, ils ne concordent pas avec les informations présentes dans les « fiches » d'exemples attestés collectés par les lexicographes. Ainsi, Hausser constate, en comparant le Webster au BNC, qu'il existe une série de mots présents dans le dictionnaire qui ne sont pas attestés dans le BNC (par exemple *aspheric, bipropellant, dynamotor* etc.), et ce malgré la taille du BNC, qui, rappelons-le, est un corpus équilibré et représentatif.

De la même manière, la comparaison entre la liste des entrées du dictionnaire de la langue allemande contemporaine « Wörterbuch der Gegenwartssprache » (WDG) en 6 volumes avec la liste des mots-formes du corpus équilibré DWDS montre qu'une centaine d'entrées (connues d'un locuteur natif) ne sont pas présentes en tant que lexème dans le corpus. Ces entrées représentent les lacunes systématiques du corpus. Par exemple : des mots du langage enfantin *Heiabett* 'lit pour dormir' ou *Puthenne* 'poule', des variantes régionales comme *Spaßettln* variante autrichienne/bavaroise pour *Scherz* 'blague', *Powidel* variante autrichienne pour 'compote de prunes' ou *pitschepatschenaß* 'très mouillé' ou des mots relevant de domaines de connaissance particuliers, tels que *Abdrusch* agriculture : 'battage', *Thomasverfahren* qui renvoie à un procédé de production de l'acier, ou encore *Pomeranzenlikör* qui renvoie à un liquide utilisé pour la pâtisserie.

De plus, on trouve dans les dictionnaires des acceptions qui ne sont pas attestées dans les corpus, et ce pour des mots très fréquents. Par exemple, le mot anglais *dope* dont l'emploi adjectival est attesté dans le NODE (*That suit is dope*) n'apparaît pas dans le BNC[1].

En ce qui concerne les mots composés, le manque d'attestations dans le corpus est encore plus flagrant. Dans une étude statistique sur les noms composés, Senellart (1996) montre qu'il faudrait disposer d'un corpus représentant 50 années du journal *Le Monde* si l'on voulait « voir » attester au moins une fois tous les mots composés recensés selon les critères définitoires du lexique-grammaire (Leclère 2002).

2.2 Expressions figées

Nous avons fait une expérience sur la répartition et le nombre d'occurences de certaines expressions figées dans un grand corpus, le DWDS-E (Geyken *et al*. 2004). Le but de ce travail était de décrire l'accroissement des nombres d'occurences des expressions figées pour déduire la taille minimale qu'un corpus doit avoir pour pouvoir constituer une base d'études solide sur les expressions figées.

46 expressions idiomatiques verbales ont ainsi été choisies au hasard dans un dictionnaire d'expressions idiomatiques, le Duden-11, qui « répertorie les expressions figées courantes [...] de la langue allemande contemporaine» (Duden-11, préface) :

(1) *etwas mit der Muttermilch aufsaugen* (avoir été élevé avec quelque chose)
(2) *einen über den Durst trinken* (avoir bu un coup de trop)

[1] Communication personnelle de Patrick Hanks, éditeur en chef du NODE.

(3) *Kohldampf schieben* (avoir la dalle)
(4) *Schmu machen* (faire de la gratte)

Afin de déterminer une courbe de croissance nous avons adopté la méthode suivante : le DWDS-E a été fragmenté en 100 échantillons de même taille contenant chacun 10 millions de tokens. Par des procédures d'échantillonnage, la répartition du corpus entier a été respectée à l'intérieur de chaque échantillon. La fréquence de chaque expression idiomatique a été calculée dans chacun des échantillons. On obtient ainsi la croissance d'apparition des expressions figées en prenant l'union des échantillons et en faisant la somme de leurs occurrences. Lorsque l'on approxime ces points de mesure par une fonction, on obtient une courbe qui croît (cf. figure 1 pour les expressions (1) à (4)). Le nombre apparaissant entre parenthèses après chaque expression correspond au nombre d'occurrences dans l'ensemble du corpus.

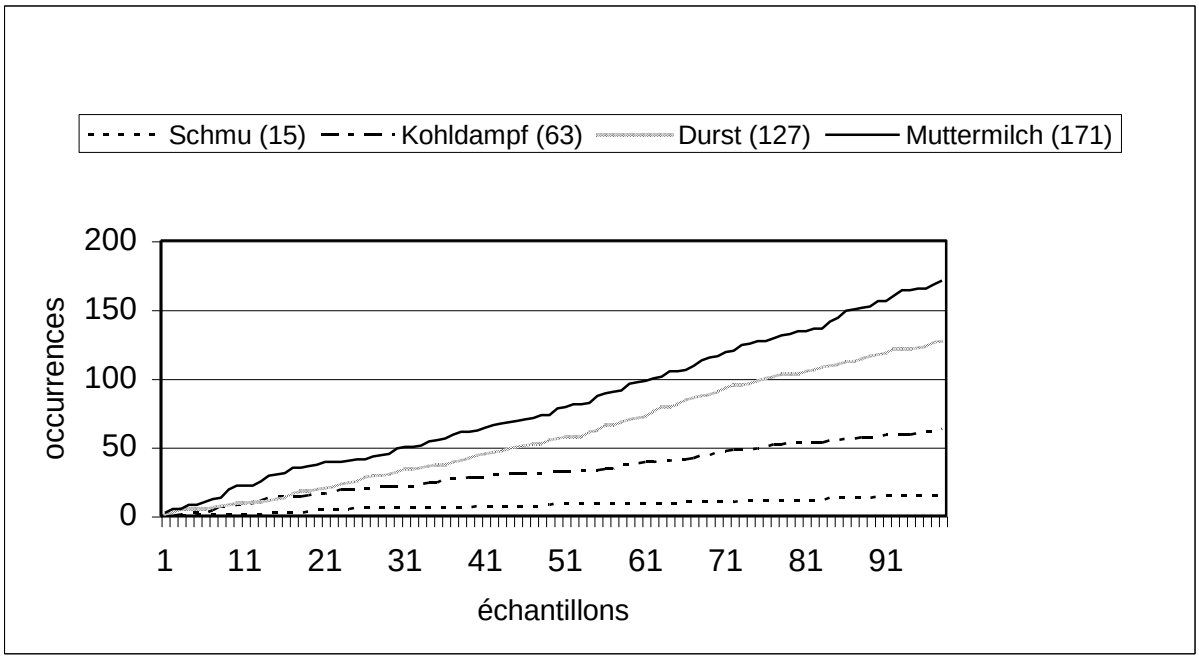

Figure 1 : accroissement des fréquences de quelques expressions figées dans le DWDS-E

Cette croissance régulière confirme que la procédure d'échantillonnage est correcte. Mais elle montre surtout qu'un corpus de la taille de 100 millions de tokens, c'est-à-dire de la taille du BNC, n'est pas suffisant pour servir de base empirique pour des études sur des expressions figées : pour l'expression (1) *Muttermilch*, on aurait une espérance de 17 occurrences, pour (2) *Durst* 13, pour (3) *Kohldampf* 6 et pour (4) *Schmu* 1,5 (100 millions de mots correspondent en effet à 10 échantillons, donc la valeur 10 sur l'abscisse). Si on classe les 46 expressions par leur nombre d'occurrences dans le corpus on obtient les classes suivantes :

i. plus de 100 occurrences : 9 expressions
ii. entre 26 et 100 occurrences : 15 expressions
iii. entre 11 et 25 occurrences : 13 expressions
iv. entre 1 et 10 occurrences : 9 expressions

Les expressions figées très fréquentes (i) auraient pour un corpus de 100 millions de mots (c'est-à-dire de la taille du BNC) en moyenne 20 attestations, les expressions moyennement fréquentes (ii) ne seraient attestées que de 2 à 10 fois. Presque la moitié des expressions (iii et iv) n'apparaîtraient pas ou correspondraient à des hapax. Par ailleurs nous avons compté les

occurrences de ces 46 expressions dans le corpus DWDS, et nous avons effectivement trouvé que 7 des 46 expressions étudiées n'apparaissent aucune fois dans ce corpus de référence de la langue allemande du XX^{e} siècle. Un corpus de la taille du BNC semble donc clairement insuffisant pour servir de base à une description lexicographique des expressions figées. Enfin, la recherche de variantes lexicales ou syntaxiques, qui présente un grand intérêt dans la description des expressions figées, est totalement impossible.

Il ne nous reste plus qu'à conclure que les corpus équilibrés sont trop petits pour pouvoir servir de base à l'élaboration d'un grand dictionnaire monolingue. Ceci vaut aussi bien pour les expressions figées, l'étude de certaines acceptions et les mots rares. D'autres auteurs tirent eux aussi cette même conclusion, tel Hausser (1998 : 5) qui affirme que « les corpus [équilibrés] disponibles ne semblent pas avoir la taille suffisante pour contenir tous les mots décrits dans les grands dictionnaires monolingues. »

3. Corpus opportunistes et très grandes collections de textes

Du fait des lacunes des corpus électroniques équilibrés, les grandes maisons d'édition ainsi que certaines institutions académiques ont commencé à constituer de grandes collections de textes qui dépassent de loin par leur taille le corpus équilibré BNC. Ces institutions ont commencé à créer des corpus opportunistes, c'est-à-dire des corpus où les textes choisis ne sont pas « proportional to their usage in every day language » (Leech 1992). Cela ne signifie pas cependant que les textes soient choisis au hasard, car ne se trouvent dans les corpus que des textes publiés, souvent à cause de leur disponibilité, le plus souvent des articles de journaux sous forme électronique. L'exemple le plus connu est sans doute la Bank of English® qui a été lancée par Collins et l'Université de Birmingham en 1991 et qui contient aussi bien des textes écrits appartenant à des genres différents que des transcriptions de la langue orale. En 2006, ce corpus contenait 524 millions de *tokens*, et il continue de croître[2]. Le DWDS-E déjà cité, avec un milliard de mots ou le « corpus de la langue écrite » de l'Institut de la langue allemande (IDS)[3] avec ses deux milliards de mots constituent deux autres exemples.

Plus récemment, de très grandes collections de textes de plusieurs centaines de millions, voire même de plusieurs milliards de mots, ont été constituées à partir du Web, et ce dans plusieurs langues, notamment l'anglais, l'italien ou l'allemand (Baroni 2006). Grâce aux procédures toujours plus performantes, ces collections de textes sont relativement aisées à établir. Dans le cas de corpus extraits du Web il ne s'agit pas à proprement parler de corpus, car la compilation ne repose pas sur le choix explicite d'une typologie des textes (cf. la définition de corpus en § 1) mais sur la base du lexique présent dans les textes ainsi que sur l'ordre des résultats (« hit score ») fournis par Google (Sharoff 2006).

Ces grandes collections de textes, qu'elles soient basées sur les journaux électroniques ou sur le Web, sont très utiles pour obtenir des attestations pour des mots ou des expressions rares. Dans l'étude qui a comparé le DWDS-E au dictionnaire de la langue contemporaine de la langue allemande, WDG (cf. § 1), nous avons montré qu'il existait des entrées du dictionnaire qui n'étaient pas attestées dans le corpus (Geyken 2004). Toutefois, en comparaison avec le DWDS (corpus équilibré de taille nettement inférieure, avec ses cent millions de mots), les lacunes sont nettement moins importantes. On peut ainsi supposer que le nombre d'entrées de

[2] www.collins.co.uk
[3] www.ids-mannheim.de

dictionnaires qui n'apparaissent pas, en tant que mot-forme, dans un corpus diminue davantage avec des corpus encore plus grands (cf. § 4).

Dans les deux cas, celui des corpus opportunistes et celui des collections extraites du Web, on peut se demander comment le défaut d'« équilibre » par rapport aux types de textes influence la qualité des résultats lexicographiques. Il apparaît que c'est la notion de fréquence qui est touchée. Ce n'est plus donc un critère fiable. On illustrera ce point par deux exemples.

3.1 Le genre grammatical

Les corpus sont souvent utilisés pour évaluer des informations morpho-syntaxiques contenues dans les dictionnaires. En particulier, les informations de fréquence peuvent constituer un indice dans le cas de formes concurrentes. Dans une étude[4], complétée par les résultats du DWDS-E, nous avons comparé la fréquence du genre des anglicismes, tels que *Blackout, Fitness* ou *Toast,* dans quatre grandes collections de textes de langue allemande : le corpus équilibré DWDS, le corpus opportuniste DWDS-E, le corpus opportuniste de l'IDS et une recherche Google. On ne peut comparer ni la taille, ni le contenu de ces collections de textes, mais leur fréquence devrait donner des informations quant à la distribution du genre grammatical des anglicismes.

En ce qui concerne le nom *Blackout,* les grands dictionnaires monolingues allemands contemporains donnent des informations différentes. Les dictionnaires Wahrig[5] et le WDG n'indiquent que le neutre, alors que le Duden-GWB indique le masculin et le neutre. Dans les quatre corpus, on obtient également des résultats différents (cf. figure 2). Ils ne s'accordent que sur le fait que la majorité des documents atteste l'article masculin pour le mot *Blackout.* La prédominance d'une forme ou de l'autre (neutre/masculin) est la moins marquée dans la recherche Google[6] et dans le corpus de l'IDS. Dans ces deux corpus, l'utilisation du neutre est tout à fait courante. La consultation des corpus DWDS et DWDS-E fournirait des résultats différents. Aucun des 14 exemples tirés du corpus équilibré DWDS n'atteste clairement le neutre. Il en va de même pour le DWDS-E, qui ne contient que trois documents attestant le neutre. On a ici une nette préférence pour le masculin. En regardant de plus près les textes qui attestent le neutre, on s'aperçoit qu'il s'agit du langage des jeunes ou du moins d'un registre différent de celui des exemples attestant le masculin comme on le voit dans les exemples suivants :

(1) Bremen (taz) - *Der Streit um das Blackout eines Radio-Bremen-Moderators ist noch nicht zu Ende.*
Bremen (taz) - 'La dispute autour du [das : neutre] blackout d'un modérateur d'une radio de Brême n'est pas encore terminée'.
Bremer Blackouts, in : taz - 12 ½ Jahre taz auf CD-ROM, Berlin : Contrapress-Media-GmbH 1999 [1998]

(2) *Die einzige Erzählung, die ein Happy-End hat, ist die erste, "Bankraub", und auch da darf man sicher sein, daß nur der Blackout es dem Erzähler erläßt, von den Qualen des genossenen Glücks zu berichten.*
'Le seul récit qui se termine bien est le premier, « Hold-up », et là aussi on peut être sûr que seul le [der : masculin] blackout dispense le narrateur de parler des supplices du bonheur joui'.

[4] www2.hu-berlin.de/korpling/lehre/ws-2003/hs-phaenomene-deutsch/Phaenomene-Anglizismen-klein.pdf
[5] www.wissen.de
[6] requête datant du 11-1-2004

Zeitzünder in der Couchecke, in : faz-Buchkritik '98, Frankfurt a.M. : Frankfurter Allgemeine Zeitung GmbH 1998 [1998]

Enfin, la fréquence d'occurrences diffère fortement d'un corpus à l'autre : le corpus de deux milliards de mots de l'IDS ne contient que 54 occurrences, alors que le corpus de un milliard de mots du DWDS-E en contient 86. La comparaison des trois corpus montre que non seulement la quantité mais aussi la qualité des corpus jouent un rôle important dans l'évaluation des résultats.

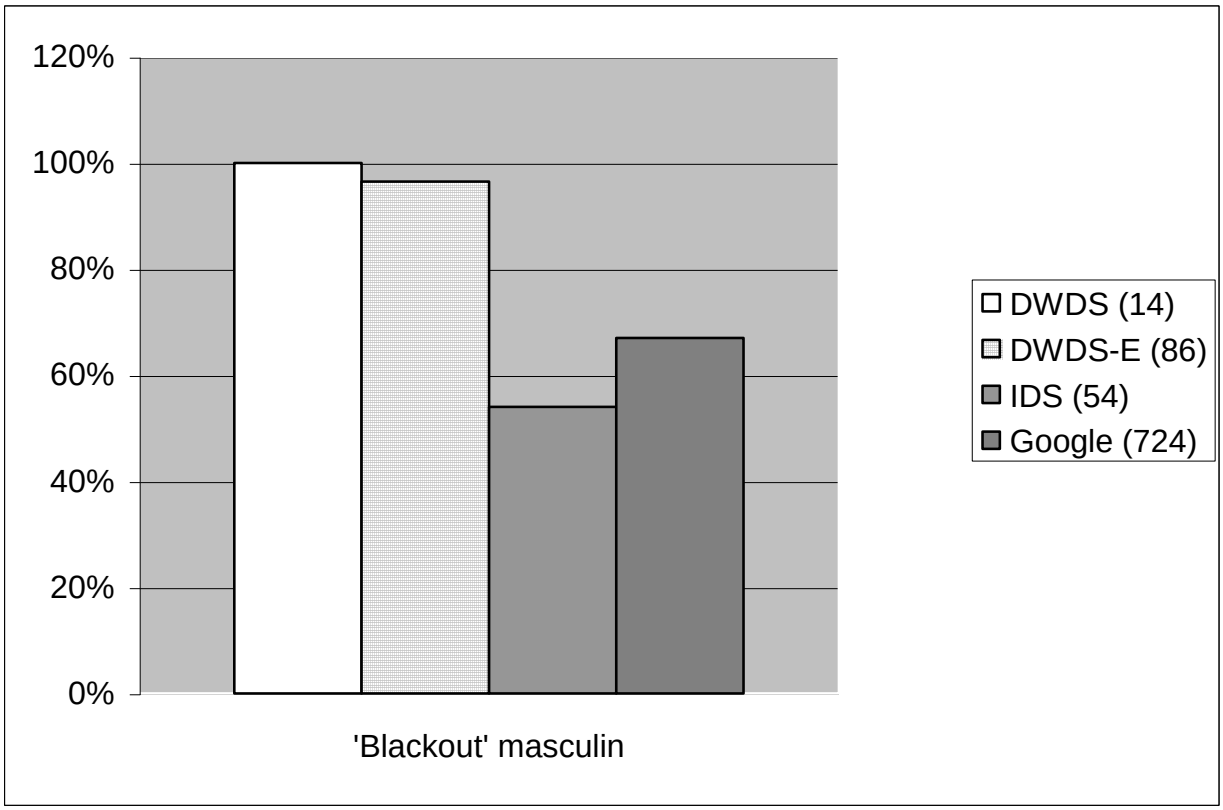

Figure 2 : fréquence du mot *Blackout* dans quatre collections/corpus de textes

3.2 Les archaïsmes

On aimerait également utiliser les corpus pour découvrir les archaïsmes dans les dictionnaires afin de les annoter ou de les éliminer lors de la mise à jour. Comme on l'a vu ci-dessus, il est problématique de déduire de l'absence de données dans un corpus leur absence dans la langue. Néanmoins, dans des projets pratiques, on a parfois tendance à négliger cet argument. Par exemple, le projet LexiView (Heid *et al.* 2000) utilise un corpus de 250 millions de mots constitués à partir de trois sources différentes (les journaux *taz*, *fr* et l'agence de presse *dpa*) pour la mise à jour des mots clés du grand dictionnaire anglais-allemand de Langenscheidt. Même si les auteurs de cet article sont bien conscients que leur corpus est un corpus opportuniste, ils ne résistent pas à la tentation d'utiliser le manque de données pour en déduire qu'il s'agit d'archaïsmes : « […] the use of corpus tools not only showed us words to be taken up into the new Großwörterbuch, it also showed dictionary corpses, old or rare words which miraculously had survived generations of lexicographers and revisions : *Immobilienmagnat* or *immobilisieren* are such words which used to be entries in the Handwörterbuch but had zero-evidence in the corpus. » (Heid 2000 : 192 ff.). Or, la même requête dans le DWDS-E montre que ces deux mots sont encore bien vivants dans la langue : ainsi, pour *immobilisieren,* on

trouve 75 occurrences et pour *Immobilienmagnat* 27, toutes réparties dans des journaux différents (*faz, nzz, sz* et *taz*).

4. Les corpus : une question de taille ?

On est donc amené à se demander si les problèmes ci-dessus ne sont pas dus au fait qu'il n'existe pas de corpus équilibrés de taille suffisante. Un bon nombre de lexicographes et de linguistes travaillant sur corpus sont conscients du problème que constitue le manque de données dans les corpus. Par exemple, Kennedy (1998) affirme que les « corpus studies have shown that a large proportion of the forms or elements of a language occur very rarely in actual use ». Le lexicographe Michael Rundell argumente dans le même sens sans pour autant faire une estimation sur une taille optimale : « What all of this adds up to is a powerful argument for very large amounts data, and it is not yet clear whether, for lexicography at least, we can say what might constitute an optimum size for a corpus » (Rundell 1996).

Mis à part les problèmes pratiques quant à la construction de tels corpus – il serait par exemple très coûteux de produire suffisamment de transcriptions de langage parlé – une manière - certes indirecte - pour estimer une taille optimale consisterait à démontrer que le nombre de types ou mieux de mots-formes converge à partir d'une certaine taille. En linguistique quantitative les questions de la croissance et de la convergence du vocabulaire d'un corpus ont été étudié depuis longtemps (par ex. Habert *et al*. 1997 : 190), mais sur la base de corpus nettement plus petits que le corpus DWDS-E. Nous avons déterminé l'allure du vocabulaire de DWDS-E en deux étapes : premièrement nous avons identifié les mots-formes et les lexèmes dans le corpus à l'aide d'un analyseur morphologique, et deuxièmement nous avons employé des méthodes statistiques pour approximer une fonction de croissance.

Analyse morphologique

Afin d'analyser les 9 millions de types du DWDS-E, nous les avons subdivisés en deux listes : une liste de formes « inintéressantes » pour l'analyse du vocabulaire (liste « négative ») et une liste de formes analysables par une analyse morphologique (liste « positive » des mots-formes (cf. §1). Nous comptons parmi les formes inintéressantes les chiffres, les numéros de téléphone, les dates, des combinaisons particulières comme *20er-Liga*, *ZX-4*, *3/10*, *3/100stel*, *afp/dpa*, etc., ainsi que des noms propres de personnes, de lieux et les noms d'entreprises. Les formes correspondant à des noms propres ont été identifiées par des grammaires locales (Didakowski et al. 2006). Dans la liste des formes « positives », on trouve tous les mots-formes, analysables par un analyseur morphologique. Étant donné la particularité de l'allemand de former des mots composés sans introduire un espace entre les lexèmes, un analyseur morphologique de l'allemand doit incorporer des règles complexes de composition. Ainsi les formes composées telles que *Taschenbuch* 'livre de poche' ou *Landesgruppenvorsitzenden* 'président du groupe d'une région' seront respectivement décomposées en *Tasche* 'poche' et *Buch* 'livre' et *Land*, *Gruppe* 'groupe' et *Vorsitzende* 'président'. Par ailleurs l'analyse morphologique doit être capable de lemmatiser les formes. Par exemple, les mots-formes *Arzt* 'médecin' *Ärzte* [pluriel], *Arztes* [génitif] ou *Ärzten* [datif pluriel] renvoient tous au même lexème *Arzt*.

Pour la lemmatisation du DWDS-E, nous nous sommes servis du système TAGH (Geyken & Hanneforth 2006), un analyseur morphologique basé sur des transducteurs à poids (weighted finite state transducers, WFSA). Ce système est composé d'un dictionnaire de plus de 200 000 racines lexicales. Lors de l'analyse lexicale, plus de mille règles de formation de mots sont utilisées. Elles permettent par exemple de déterminer des adjectifs construits sur base verbale avec le suffixe *-bar*, tels que *verhandelbar* 'négociable', *heilbar* 'guérissable' ; ou bien de

former un nom d'habitant à partir d'un nom de ville tel que *Donaueschinger* à partir de *Donaueschingen* où le morphème *–er* remplace le suffixe *–en* du nom de ville.

La décomposition correcte des mots composés pose différents problèmes, d'une part parce que certains sont sémantiquement opaques, d'autre part à cause de l'ambiguïté potentielle. Par exemple, le mot *Gendarm* 'gendarme' ne doit pas renvoyer aux mots *Gen* 'gène' et *Darm* 'intestin' ou le mot *Eisenhut* 'aconit' ne doit pas renvoyer à *Eisen* 'fer' et *Hut* 'chapeau'. D'autres composés peuvent être découpés de différentes façons. Dans ce cas, le but de l'analyse automatique sera non seulement de bien découper le mot, mais aussi d'associer chaque partie à la racine correcte. Le mot *Telekommunikation* 'télécommunication', par exemple, peut être découpé, lors de l'analyse automatique de quatre façons différentes où le nombre entre parenthèses réfère au poids associé à l'analyse (tableau 4). On utilise ici la notation adoptée par TAGH, notamment '#' pour une délimitation entre deux racines autonomes, '|' pour délimiter un préfixe ou un morphème non autonome et une racine autonome, et '\' pour marquer un élément qui joint deux racines.

- Tele | kommunikation (5)
- Tele | komm # unikat # ion (25)
- Tele | komm# uni# kat # ion, (35)
- Telekom # muni # kat # ion (40)

Tableau 4 : l'analyse morphologique du mot *Telekommunikation*

Dans ce cas, la seule analyse plausible est la première. Et c'est celle qui sera appliquée par l'analyseur TAGH, car le système privilégie le découpage avec un poids minimal : d'une part les poids sont définis en fonction des combinaisons de catégories lexicales ; d'autre part, une combinaison nom-nom est considérée plus probable qu'une combinaison verbe-nom ; on associe ainsi un poids supérieur à la combinaison verbe-nom.

Cette heuristique du poids minimal est suffisante dans la majorité des cas, mais pas dans tous. Par exemple, dans le cas de *Wochenarbeitstag* 'jour de travail en semaine', le découpage peut s'effectuer en Wochen#arbeit\s#tag et en Wochen#arbeit#stag. Les deux analyses contiennent le même nombre de mots découpés. Afin d'éliminer le deuxième découpage, on a besoin d'une nouvelle heuristique : le lexème *Stag*, mot de la marine désignant une corde solide qui permet de fixer et soutenir le mât, apparaît dans le corpus, non seulement très rarement, mais elle n'est jamais partie intégrante d'un composé. Pour cette raison, *Stag* sera exclu de toute analyse de décomposition de mots composés. De cette façon, il ne restera, dans le cas de *Wochenarbeitstag*, que l'analyse correcte.

Il existe également des cas où le découpage du niveau lexical n'est pas évident sans contexte. Par exemple, le mot-forme *Ministern* 'ministres' peut être analysé de deux manières différentes : en tant que pluriel du mot *Minister* 'ministre', et en tant que composé des 2 mots Mini# Stern 'petite étoile'. Le mot *Wachstube* peut être découpé en Wach#stube 'chambre de garde' et Wachs#tube 'tube de cire'. Dans chaque cas, les deux découpages sont plausibles.

D'autres difficultés apparaissent quand le découpage renvoie à différentes racines. Le mot *Tropenholztheke* 'comptoir en bois des tropiques' sera de façon non ambiguë découpé en *Tropen#holz#theke*. Toutefois le mot *Trope* est ambigu et renvoie à la fois à la zone géographique située près de l'Équateur et au pluriel du mot de stylistique *Tropus* 'trope'. L'humain n'aura aucune difficulté à effectuer le découpage de ce composé, alors que la

machine, qui ne possède pas de connaissances sur le monde, ne pourra décider de l'analyse que sur la base d'une plausibilité statistique.

A l'aide de l'analyse morphologique, 6 des 8,9 millions de types du corpus DWDS-E peuvent être analysés. Ces 6 millions de mots-formes peuvent être réduits ultérieurement à 3,9 millions de lexèmes. Grâce aux règles de prétraitement et au filtre des noms propres, environ 1,2 million de types différents constituent la liste « négative ». Il reste ainsi presque 1,7 million de types qui n'ont pu, jusqu'à présent, être catégorisés par l'analyse automatique. Cette liste contient, comme le montre un échantillon de 500 mots que nous avons évalué manuellement, des noms propres, des variantes régionales et dialectales (comme *ick* 'je', *nit* 'ne … pas' ou *wa* 'n'est-ce pas'), des fautes d'orthographe, du matériel en langue étrangère (*mon, the*), des abréviations inusitées (*stellvertr*., *Kammerorch*.), des mots précédant la réforme de l'orthographe de 1902 (*diktiren* 'dicter', *Litteratur* 'littérature', *Antheil* 'part', etc.). Mis à part les mots anciens (il s'agit alors généralement des variantes orthographiques des lexèmes actuels), ces mots-formes ne sont pas des candidats pour un dictionnaire de langue générale et ne doivent donc pas être pris en compte pour le calcul de l'accroissement du vocabulaire.

Approximation d'une fonction d'accroissement du vocabulaire

Après avoir identifié les lexèmes du corpus DWDS-E, nous sommes maintenant en mesure de poser la question de l'accroissement du vocabulaire en comptant l'apparition de nouveaux lexèmes au fur et à mesure que l'on avance dans la lecture du corpus. De quelle façon peut-on mesurer cet accroissement ? La méthode naïve, qui consisterait à compter le nombre de lexèmes d'un corpus de 100 millions de mots, ensuite à compter le nombre de lexèmes du corpus d'un milliard de mots, et de définir l'accroissement comme la différence entre le deuxième et le premier nombre, n'est pas suffisante. En effet, il se peut que certains textes contiennent proportionnellement plus de mots différents que d'autres, alors que l'on s'attend à trouver dans d'autres textes (par exemple les textes littéraires) moins de mots différents (types). Cet argument montre la nécessité d'une démarche statistique bien fondée.

Tout comme précédemment (cf. § 2), nous avons partagé le corpus DWDS-E en 100 échantillons de même taille, soit 10 millions de mots consécutifs, en respectant pour chaque échantillon la répartition de l'ensemble du corpus. Ensuite, pour chaque échantillon, nous avons compté le nombre de lexèmes à l'aide du système morphologique TAGH. L'accroissement résulte de la différence des lexèmes contenus dans les échantillons. Pour cela, on a choisi un procédé itératif semblable à celui de la détermination de la courbe de croissance des expressions figées. En prenant l'union de deux échantillons, on obtient un nouvel ensemble de mots, qui sera de nouveau comparé avec un autre échantillon de mots et formera un ensemble de différences, etc. Nous avons poursuivi ce procédé jusqu'à épuisement de tous les échantillons. Afin de vérifier que la croissance estimée ne dépend pas de la succession des échantillons, nous avons répété le procédé 25 fois en permutant au hasard la succession des échantillons.

La figure 3 rend compte de l'accroissement du vocabulaire mesuré en lexèmes. La courbe correspond à une approximation après 25 itérations. A cause de l'échelle, on voit à peine les variations de points de mesure. Ceci est remarquable puisque cela démontre que les lexèmes se trouvent être répartis très régulièrement dans chaque échantillon. On trouve en moyenne, dans chaque échantillon, 235 000 racines différentes (95 des 100 échantillons contiennent entre 230 000 et 240 000 racines). Seuls 5 échantillons en contiennent un peu moins de 230 000.

Par ailleurs, on observe que l'accroissement devient plus modeste au fur et à mesure qu'on ajoute de nouveaux textes. Après un accroissement fort au début (c'est-à-dire avec les deux premiers échantillons de 120 000 lexèmes), on compte 25 000 lexèmes supplémentaires après 99 échantillons. En d'autres termes, après avoir réuni les 99 échantillons, c'est-à-dire les 990 millions de mots consécutifs, un nouveau lexème apparaît tous les 400 mots.

Ce résultat signifie que la croissance continue même pour de très grands corpus. On pourrait objecter que ce résultat est dû à la particularité de la composition en allemand. Ceci serait vrai si on associait le vocabulaire à l'ensemble des types d'un corpus, car on compte les mots composés de l'allemand comme des types alors que les mots composés du français ou de l'anglais correspondent à des expressions. Si on considère par contre que le vocabulaire d'une langue comprend aussi bien tous les types d'un corpus que les mots composés, le résultat de l'accroissement du vocabulaire semble tout à fait en accord avec des études menées pour le français (Senellart (1996). En effet, nous montrons dans l'exemple suivant que des mots composés « intéressants » apparaissent au fur et à mesure qu'on ajoute des textes pour des composés productifs.

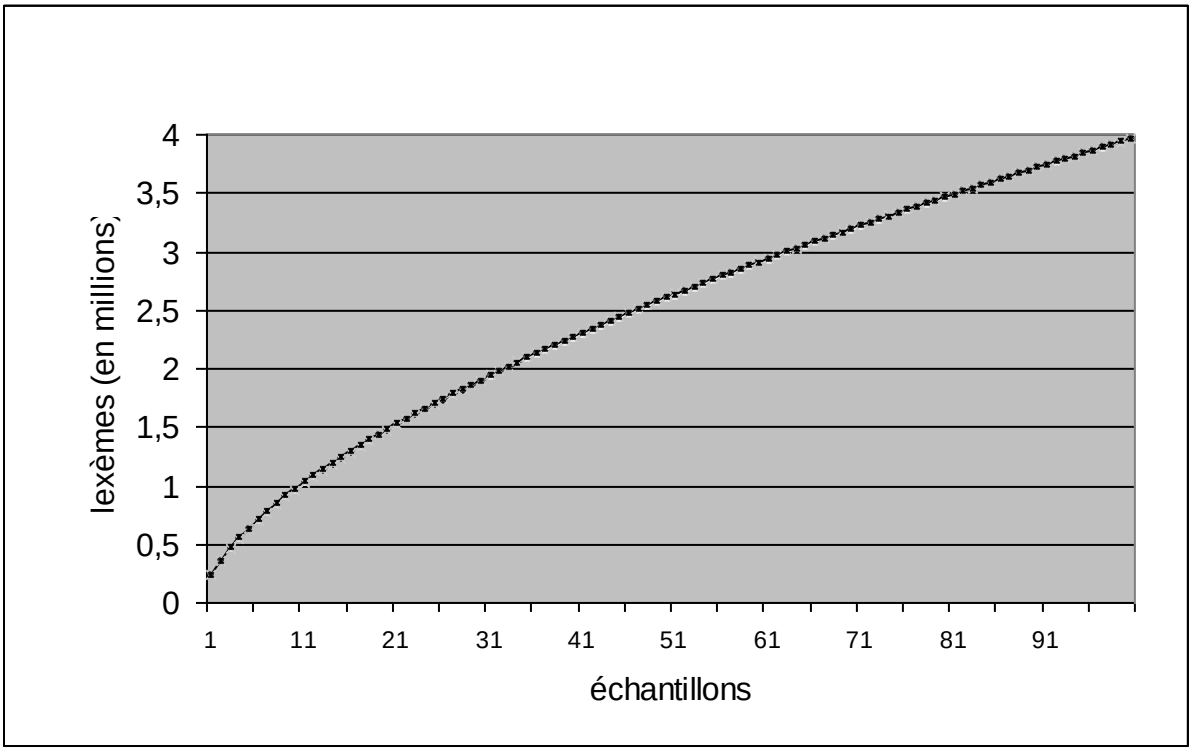

Figure 3 : l'accroissement du vocabulaire mesuré en lexèmes dans le DWDS-E

Exemple : mots composés avec *Selbst* 'soi-même'

Nous avons compté les composés avec *Selbst* 'soi-même' et nous leur avons appliqué le même procédé pour déterminer la courbe de croissance. Les résultats de la figure 4 montrent que le nombre des composés avec *Selbst* continue de croître malgré la grande taille du corpus, un résultat qui est bien en accord avec l'accroissement du vocabulaire décrit précédemment. La richesse des corpus devient encore plus apparente si on compare le nombre des composés en *Selbst* dans le DWDS-E avec les entrées du Duden-GWB. Dans le Duden-GWB, on décrit 244 entrées commençant par *Selbst* par rapport à 10 934 types (7 180 lexèmes) dans le DWDS-E. 307 lexèmes n'apparaissant pas dans le dictionnaire ont plus de 30 occurrences dans le corpus. Parmi les omissions du Duden se trouvent par exemple : *Selbstregulierung* 'autorégulation', *Selbstbedienungsmentalität* 'cupidité' ou *Selbstabfertigung* 'autoexpédition'.

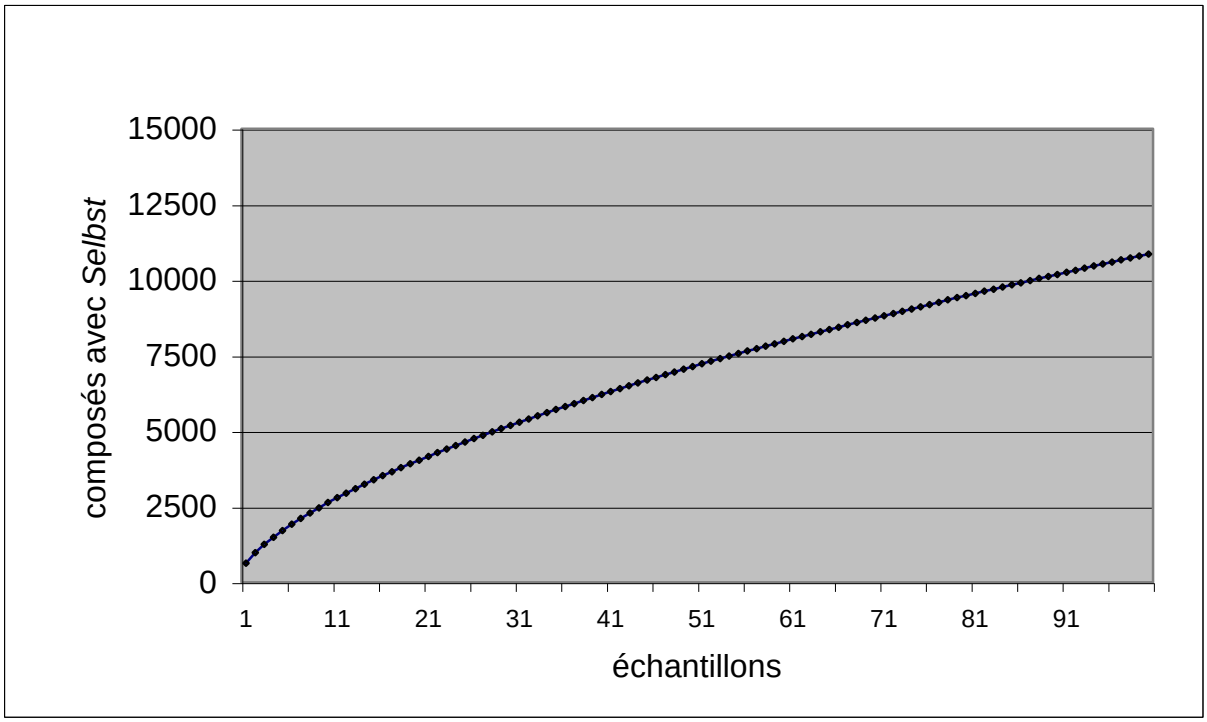

Figure 4 : l'accroissement des composés avec *Selbst* dans le DWDS-E

5. Conclusion

Les grands corpus électroniques constituent désormais, comme on vient de le voir, une aide importante pour la rédaction, la révision et la mise à jour de nombreux dictionnaires. Malgré l'utilité de grands corpus dans la constitution des dictionnaires, un certain nombre de problèmes théoriques persistent, notamment à cause des problèmes méthodologiques liés à leur constitution. Les plus grands corpus équilibrés disponibles actuellement ne contiennent pas suffisamment d'attestations pour servir de base lexicographique unique, notamment pour des phénomènes tels que les expressions figées, les emplois rares. D'autre part, la tentative de surmonter les problèmes du manque d'attestation pour un bon nombre de phénomènes linguistiques par la constitution de corpus opportunistes beaucoup plus grands que les corpus équilibrés suscite d'autres problèmes qui sont liés justement à leur constitution non équilibrée. Nous avons ainsi montré que la fréquence d'un mot-forme pouvait fortement varier d'un corpus opportuniste à l'autre ; la fréquence n'est donc plus un critère fiable, ce qui pose un problème quant à « l'objectivité » des corpus opportunistes.

En résumé, à l'heure actuelle, dans la création et la mise à jour de dictionnaires, les corpus électroniques ne parviennent pas encore à remplacer les observations informelles des équipes de lexicographes ainsi que les fiches traditionnelles. Elles continuent de rester indispensables. Cette remarque n'est pas une objection de principe contre les corpus. Depuis la création du Brown Corpus avec 1 million de mots, en passant par le British National Corpus de 100 millions de mots, on a franchi aujourd'hui le seuil du milliard de mots. Chaque passage à un ordre de grandeur supérieur a fourni de nouvelles pistes de réflexion sur la question de l'usage des corpus en linguistique. Néanmoins les critères qui président à la constitution des corpus ainsi que le problème de leur taille restent des questions qu'il faut continuer de travailler.

Références – Dictionnaires :

[Duden-GWB] : *Duden. Das Große Wörterbuch der Deutschen Sprache in 10 Bänden*. 1999. Mannheim : Dudenverlag.
[Duden-11] : *Duden 11. Idiomatik*, 1999. Mannheim : Dudenverlag.
[DWB] : Grimm Jacob und Wilhelm, *Deutsches Wörterbuch*. 1854-1960. Hirzel : Leipzig.
[NODE] : The *New Oxford Dictionary of English*, 1999. Oxford University Press.
[Littré] : *Dictionnaire de la langue française*, 4 vol., 1863-1873. Littré É. Paris: Librairie Hachette.
[OED]: *Oxford English Dictionary*, second edition, 1989, Clarendon Press : Oxford.
[TLF] : *Trésor de la langue française, Dictionnaire de la langue du XIX^e^ et du XX^e^ siècle (1789-1960)*, 16 vol., 1971-1994. Paris: Klincksieck, vol. 1-10. Paris: Gallimard, vol. 11-16.
[WDG] : Wörterbuch der deutschen Gegenwartssprache, 1961-1977. Akademie-Verlag : Berlin .

Autres Références :

Baroni, M., Kilgarriff, A. (2006). *Large linguistically-processed Web corpora for multiple languages*. Conference Companion of EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics), East Stroudsburg PA : ACL, 87-90.

Biber, D. (1994). *Representativeness in corpus design*. In : Zampolli, Antonio/Calzolari, Nicoletta/Palmer, Martha (eds.) : Current Issues in Computational Linguistics : In Honour of Don Walker (Linguistica Computazionale IX-X). Pisa. Kluwer : Giardini/Dordrecht, 377–407.

Clear, J. (1992). *Corpus sampling*. In : Leitner, G. (éd.). «New directions in English language corpora ». Mouton de Gruyter : Berlin, 21-31.

Didakowski, J., Geyken, A., Hanneforth, T. (2006). *Eigennamenerkennung mit großen lexikalischen Ressourcen*. Proceedings of KONVENS 2006, 3.-7. Oktober 2006, Konstanz, 9-14.

Geyken A. (2004). *Korpora als Korrektiv für einsprachige Wörterbücher* . Zeitschrift für Literatur und Linguistik, Heft 136, 72-100.

Geyken A., Sokirko A., Rehbein I., and Fellbaum Ch. (2004). *What is the Optimal Corpus Size for the Study of Idioms? DGfS-Jahrestagung*, Mayence.

Geyken, A., Hanneforth, T. (2006). *TAGH : A complete morphology for German based on weighted finite state automata*. Finite-State Methods and Natural Language Processing. Lecture Notes in Artificial Intelligence, Band 4002, Springer : Berlin.

Geyken, A. (2007). *The DWDS corpus : A reference corpus for the German language of the 20th century*. In : Fellbaum, C. (éd.) : Collocations and Idioms : Linguistic, lexicographic, and computational aspects. Continuum : London, à paraître.

Gorcy, Gérard (1992). *Le "Trésor de la langue française (TLF)" trente ans après; bilan et perspectives*. Études de linguistique appliquée, n° 85/86, (janvier- juin 1992), 75-88.

Habert, B., Nazarenko, A., Salem, A. (1997). *Les linguistiques de corpus*. Armand Collin/Masson : Paris.

Habert, B. (2000). *Des corpus représentatifs : de quoi, pour quoi, comment ?* Cahiers de l'Université de Perpignan n° 31, 11-58.

Hausser R., (1998). *Häufigkeitsverteilung deutscher Morpheme*. *LDV-Forum*, 15 (1), 6-26.

Heid, U.; Evert, S.; Docherty, V.; Worsch, W. et Wermke, M. (2000), *A data collection for semi-automatic corpus-based updating of dictionaries*. Proceedings of the 9th EURALEX International Congress, 183 - 195.

Jacques, M.-P. (2005). *Pourquoi une linguistique de corpus*. Dans Williams, G. (éd.) 2005. La linguistique de corpus. Presses Universitaires de Rennes, 21-29.
Kennedy G. (1998). *An Introduction to Corpus Linguistics*. Longman : London.
Kilgarriff, A. (2002). *Sketching Words*. In Corréard M.-H. (éd.), Lexicography and Natural Language Processing. Festschrift in Honour of B.T.S. Atkins, 125-138.
Kilgarriff, A. et Grefenstette, G. (2003). *Introduction to the special issue on the web as corpus*. Computational Linguistics 29 :3, 333-348.
Leclère C. (2002). *Organization of the Lexicon-grammar of French Verbs* ». *Lingvisticae Investigationes*, 25 :1, 29-48.
Oostdijk N. (1988). *A Corpus Linguistic Approach to Linguistic Variation*. In Dixon G. (éd.), *Literary and Linguistic Computing*, 3 (1), 12-25.
Polguère, Alain (2003) *Lexicologie et sémantique lexicale. Notions fondamentales*, coll. « Paramètres », Presses de l'Université de Montréal : Montréal.
Rieger B. (1979). *Repräsentativität : Von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung*. In Bergenholtz H., Schaeder B. (éds), «Empirische Textwissenschaft. Aufbau und Auswertung von Textkorpora». Königstein. [Monographien Linguistik und Kommunikationswissenschaft ; 39], 52-70.
Rundell M. (1996). *The corpus of the future, and the future of the corpus* . Special conference on « New Trends in Reference Science ». Exeter.
Rundell M. (2002). *Good Old-fashioned Lexicography : Human Judgement and the Limits of Automation*. In Corréard M.-H. (éd.). « Lexicography and Natural Language Processing ». Festschrift in Honour of B.T.S. Atkins, 138-156.
Senellart J. (1996). *Statistique Prudence : quelques études statistiques avec les noms composés dans le Journal Le Monde*. « Actes des premières journées INTEX. LADL, Université Paris 7.
Sharoff, S. (2006). *Creating general-purpose corpora using automated search engine queries*. In Marco Baroni and Silvia Bernardini, ed., « WaCky! Working papers on the Web as Corpus». Gedit : Bologna.
Sinclair J. (1996a). *Preliminary Recommendations on Corpus Typology*. Rapport technique, EAGLES (Expert Advisory Group on Language Engineering Standards). Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale. Pise.
Sinclair J. (1996b). *The First Cobuild Dictionary* 1987. In Foley, J. (éd.). J.M. Sinclair on Lexis & Lexicography. Unipress : Singapore, 137-153.
Sinclair, J. (1997). *Corpus Evidence in Corpus Description*. In Wichmann A., Fligelstone S., McEnery T., Knowles G. (éds.), «*Teaching and Language Corpora*». Longman : London, 27-39.
Teubert, W., Čermáková, A. (2004). *Directions in corpus linguistics*. In M.A.K. Halliday, W. Teubert, C. Yallop and A. Čermáková (ed). 2004. «Lexicology and Corpus Linguistics». An Introduction. Continuum : London, 113-167.