Corpus Linguistics —
Une modeste contribution à l'histoire des sciences

Carmela Chateau-Smith

Université de Bourgogne, Dijon

… scientific and technological progress has become
one of the most important factors in the
development of human society…[1][2]

Résumé : Au cours du vingtième siècle, les sciences du langage et les sciences de la Terre ont toutes deux connu des changements de paradigme, en grande partie liés à des développements technologiques.

Abstract : During the twentieth century, the sciences of Linguistics and Geology both underwent paradigm shifts. These shifts were to a great extent linked to technological developments.

Résumé : Cette étude explore l'histoire de la mise en place de la linguistique de corpus au cours du vingtième siècle, à partir de 1950, en tant que changement de paradigme dans l'analyse du langage, plus particulièrement appliquée à la langue anglaise. Des parallèles seront évoquées entre le développement de la linguistique de corpus, considéré comme un changement de paradigme impulsé par des avancées technologiques et les bouleversements similaires et synchrones qui ont eu lieu dans le domaine des sciences de la Terre, impulsés de même manière par des technologies nouvelles : il s'agit du passage de la notion d'une Terre figée à la théorie de la dérive des continents et la tectonique des plaques. L'école contextualiste britannique de la linguistique de corpus sera présentée, à travers trois de ses figures clés : le précurseur, John Rupert Firth, le fondateur John McHardy Sinclair, et le visionnaire William E. (Bill) Louw, le premier à avoir compris l'importance du phénomène de la prosodie sémantique.

Abstract : This study explores the history of the development of corpus linguistics as a science, during the twentieth century, from 1950 onwards, seen as a paradigm shift in the analysis of language, with particular application to English. Parallels will be drawn between the development of corpus linguistics, considered as a paradigm shift made possible by technological progress, and the similar and synchronic upheaval which took place in the Earth Sciences, also grounded in the progress of technology, with the change from the notion of a fixed Earth to the theory of continental drift and plate tectonics. The British contextualist school of corpus linguistics will be presented through three key figures in the field : the precursor John Rupert Firth, the founder John McHardy Sinclair and the visionary William E. (Bill) Louw, the first to grasp the importance of the phenomenon of semantic prosody.

Mots-clés : changements de paradigme, linguistique de corpus, histoire et philosophie des sciences, prosodie sémantique.
Key-words : corpus linguistics, history and philosophy of science, paradigm shifts, semantic prosody.

[1] « … le progrès scientifique et technologique est devenu l'un des facteurs les plus importants dans le développement de la société humaine … » Déclaration sur l'utilisation du progrès scientifique et technologique dans l'intérêt de la paix et au profit de l'humanité. <http://www2.ohchr.org/english/law/mankind.htm> Consulté le 29 juillet 2012.
[2] *N. B.* Toutes les traductions proposées dans cet article sont l'œuvre de l'auteur, C. Chateau-Smith.

Introduction

L'une des questions concernant les avancées scientifiques est de savoir si la science fait avancer la technologie ou si c'est la technologie, au contraire, qui fait avancer la science. En science, le modèle du progrès où les avancées scientifiques sont conditionnées par les avancées technologiques est souvent appelé « *instrument-based science* » [la science basée sur les instruments]. D'ailleurs, c'est le terme utilisé dans le dossier de presse de la *National Aeronautics and Space Administration* (NASA, 2012) présentant les recherches à mener après l'atterrissage[3] sur Mars du Rover *Curiosity*[4], où il est précisé que « *The research will use 10 instrument-based science investigations*[5]. » Il est vrai que la création d'un nouvel outil peut parfois mener à des innovations de forme et de fond dans les pratiques scientifiques. L'expression « *instrument-powered science* » [la science dont les instruments sont le moteur] exprimerait encore mieux ce concept de progrès scientifiques mus par des progrès techniques.

De nombreuses sciences se sont développées à partir du moment où les instruments nécessaires à leur évolution technique ont été construits, mais ces deux événements ne se trouvent pas forcément liés de manière directe dans une relation de cause à effet. Ainsi, l'océanographe Timothy J. G. Francis explique que l'hypothèse de F. J. Vine et D. Matthews (1963) ne prendra forme qu'une dizaine d'années après la mise en service des magnétomètres tractés :

> *Oceanography is a technology-led science. When a major new piece of equipment is introduced, big advances in the science follow with a time lag of perhaps ten years—while the scientists learn how to make use of it and to interpret the new information obtained. Thus the introduction of towed magnetometers in the 1950s led to the Vine–Matthews hypothesis and plate tectonics in the 1960s*[6]. (Francis, 1988[7]).

Ce qui est vrai pour l'océanographie l'est aussi pour la linguistique de corpus. De même manière, la collecte de données langagières authentiques, en grande quantité et en format

[3] *N.B.* Le terme anglais « *landing* » ne fait pas référence au nom de la planète Terre.
[4] L'appareil s'est effectivement posé sur la planète Mars à la date prévue, le 6 août 2012.
[5] La recherche fera appel à 10 investigations scientifiques basées sur des instruments.
[6] L'océanographie est une science basée sur la technologie. Quand on introduit une innovation majeure dans l'équipement, de grandes avancées scientifiques suivront avec un décalage dans le temps de peut-être dix ans, pendant que les scientifiques apprennent à s'en servir et à interpréter les nouvelles informations obtenues. Ainsi, l'introduction de magnétomètres tractés dans les années 1950 a conduit à l'hypothèse de Vine et Matthews et à la tectonique des plaques dans les années 1960.
[7] Ce résumé est disponible en ligne : <http://www.sut.org.uk/journal/contents/journal_14.htm>. Consulté le 25 juillet 2012.

exploitable par ordinateur, n'a pu se mettre en place qu'à partir du moment où les outils nécessaires se sont développés. Le développement d'outils permettant d'enregistrer la parole, vers le milieu du vingtième siècle, a facilité la création de corpus de conversations, notamment aux Etats-Unis par Charles C. Fries (1952), et par Randolph Quirk (1960) à Londres, puis par John McHardy Sinclair (1966) à Édimbourg. Ces corpus oraux n'étaient pas accessibles par ordinateur au départ, mais ils permettaient déjà d'étudier la langue orale en action.

En partie à cause de la difficulté de transcription de la parole enregistrée, les premiers corpus électroniques étaient surtout composés de textes écrits. Une tradition d'analyses grammaticales et syntactiques s'est développée autour d'une série de corpus comparables, construits sur le même modèle, les corpus *Brown* (Francis et Kučera, 1964) et *LOB* (Johansson, Leech et Goodluck, 1978), contenant chacun un million de mots, d'une taille permettant d'analyser les mots très fréquents (les mots grammaticaux), mais trop petite pour permettre une analyse exhaustive de la plupart des éléments lexicaux.

Puis viendra l'époque des grands corpus de référence, avec le *British National Corpus* (*BNC* ; BNC Consortium, 1995) et toute la série de corpus proposés par Mark Davies de l'université Brigham Young (*BYU*), notamment le *COCA* (*Corpus of Contemporary American* ; Davies, 2008-) et le *COHA* (*Corpus of Historical American* ; Davies, 2010-). Seront également abordées la notion de *Web-as-Corpus*[8] développée entre autres par Marco Baroni (Baroni et Bernardini, 2006), et la création de corpus à partir de documents disponibles en ligne, avec l'outil *SketchEngine* d'Adam Kilgarriff et son équipe (Kilgarriff *et al.*, 2004).

La deuxième partie de cette étude regardera plus en détail les travaux de John McHardy Sinclair et l'école contextualiste britannique avec la mise en place des notions fondamentales pour une approche inductive de la linguistique de corpus. L'évolution de la notion de « *collocation* » sera retracée, tout d'abord dans la période qui a précédée le développement des corpus informatisés, à travers les recherches de H. E. Palmer et de J. R. Firth. Les notions de « *concordance* » et de « *context* » seront également mises en lumière, ainsi que la quête pour identifier les unités étendues de sens (« *extended units of meaning* »). La dernière partie présentera une analyse de l'expression « *silly ass* », qui a été utilisée par J. R. Firth comme un exemple représentatif du phénomène de collocation.

[8] La Toile en tant que corpus.

1. Le développement des corpus informatisés

Comme Thomas S. Kuhn l'a précisé dès 1962, dans *The Structure of Scientific Revolutions*[9], l'histoire d'une science est réécrite à chaque changement de paradigme. La date de la création du premier corpus informatisé en anglais fait partie d'une telle histoire, de même que l'origine du nom de la discipline.

Retraçant les origines de la linguistique de corpus, Jacqueline Léon présente une analyse pertinente de certains textes de Geoffrey Leech (1991, 1992 et 2002) qu'elle conclut ainsi : « *By ignoring the strong empiricist British filiation inherited from Firth's work, Corpus Linguistics has been deprived of a real precursor*[10]. » (Léon, 2005 : 47).

1.1. Créer une filiation : Quirk et Francis, vus par Leech

Pour quelles raisons Geoffrey Leech aurait-il laissé dans l'ombre l'aspect firthien de l'histoire de la linguistique de corpus ? En fait, par ses publications en 1991 et 1992, Leech cherchait à établir une filiation en ligne directe depuis les travaux de Randolph Quirk et de W. Nelson Francis, en opposition à la forme de linguistique prônée par Noam Chomsky, sans trop se préoccuper des branches collatérales. Il est donc compréhensible qu'en 1991 il ne mentionne que brièvement les travaux les plus récents de J. McH. Sinclair car, pendant les années 1980, l'équipe de Sinclair se consacrait non pas à la syntaxe, mais à l'étude du lexique et à la lexicographie (Sinclair, 1987).

De même manière, lors d'une étude bibliographique retraçant l'histoire des listes de fréquence de mots en anglais avant 1944, qui comporte trois pages de tableaux récapitulatifs, le psychologue américain Terry Bontrager[11] n'a fait aucune mention de la tradition britannique de Harold E. Palmer, Albert S. Hornby et Michael P. West (Bontrager, 1991). Étant donné que l'objectif de Bontrager était de retracer les origines des listes de mots utilisées pour les tests standards d'évaluation aux Etats-Unis, ce positionnement est tout à fait cohérent.

[9] Ce livre a lui-même provoqué une révolution dans le domaine de l'histoire et philosophie des sciences.
[10] En ignorant la forte filiation empiriste britannique héritée de travaux de Firth, la linguistique de corpus a été privée d'un véritable précurseur.
[11] « *I do not recall the 'British school'. It is likely that I was unaware of it when I wrote the paper.* » (Bontrager, communication personnelle, 29 août 2011). Je ne me rappelle pas 'l'école britannique'. Il est probable que je n'en avais pas connaissance quand j'ai écrit l'article.

Dans le cas de Bontrager, comme dans le cas de Leech, pour comprendre leurs choix, il convient d'analyser le contexte de culture et de situation dans lequel s'inscrivent leurs travaux. Le contexte de culture correspond à leurs communautés scientifiques respectives et au paradigme au sein duquel s'inscrivent les travaux de leurs communautés. Le contexte de situation est celui de l'époque et des événements ponctuels qui, ensemble, fournissent un cadre aux recherches menées.

Le contexte de situation pour Geoffrey Leech en 1991 et en 1992 était le suivant. Dès 1969, à l'université de Lancaster, Leech avait proposé de créer un corpus informatisé (qui deviendrait par la suite le corpus *LOB* (*Lancaster/Oslo–Bergen*), compilé selon le modèle du corpus *Brown* (Leech, 2009 : 7). Ces deux corpus, ainsi que le *London–Lund* (anglais parlé, basé sur le *Survey of English Usage*), ont été proposés sur CD-ROM en 1991 et en 1992, sous le titre *International Computer Archive of Modern and Medieval English* (*ICAME*)[12].

Il est donc tout à fait naturel que dans les deux textes publiés à cette époque, Leech ait cherché à démontrer la cohérence de cet ensemble de corpus. Plus tard, dans son article de 2009, Leech fait enfin référence aux premiers travaux de John McHardy Sinclair, mais seulement en note de fin et avec une prosodie sémantique qui reste néanmoins plutôt négative, car il ne parle pas de « *corpus* », seulement de « *collection* », réduisant ainsi l'impact des travaux de Sinclair.

> *In the 1960s John Sinclair (at first with Michael Halliday at Edinburgh, and then at Birmingham after John's move there) had put together a collection of over 100,000 words of transcribed spoken English.*[13] (Leech, 2009 : 19).

Cette même démarche a pu être observée lors des débats quant à la validité de la théorie de la genèse des continents proposée par Alfred Wegener dès 1912. Nombre de ses détracteurs ont utilisé le terme « *hypothesis* » plutôt que « *theory* », comme par exemple dans cette citation, où Charles Schuchert exprime clairement l'enjeu d'un tel changement de paradigme en géologie : « *If we are to believe Wegener's hypothesis we must forget everything which has been learned in the last 70 years and start all over again*[14]. » (Chamberlin, 1928 : 87) [Si l'on croit à l'hypothèse de Wegener, nous devrions oublier tout ce qui a été appris au cours des 70 dernières années et tout reprendre à zéro.]

[12] Accessibles <http://www.hd.uib.no/icame.html> et <http://icame.uib.no/newcd.htm>. Consultés le 25 juillet 2012.

[13] Dans les années 1960 John Sinclair (d'abord avec Michael Halliday à Edimbourg, puis à Birmingham après que John soit parti là-bas) avait rassemblé une collection de plus de 100 000 mots d'anglais parlé transcrit.

1.2. Nommer la discipline : les linguistiques de corpus ?

Si l'origine du nom de la discipline n'est pas connue avec précision, Leech indique toutefois que « *The term corpus linguistics made only occasional appearances until the publication of a book of that title edited by Aarts and Meijs* (*1984*). »[15] (Leech, 1992 : 105). Nancy Belmore a ensuite glané quelques renseignements supplémentaires par le biais d'une question posée sur la liste de diffusion *Corpora*, le 9 février 1998.[16] Jan Aarts lui a répondu qu'il avait utilisé le terme dès 1980, mais ne le trouvait pas très approprié, car : « *it is an odd discipline that is called by the name of its major research tool and data source.* »[17] [C'est une discipline étrange que l'on appelle par le nom de son principal outil de recherche et de sa source de données.] Geoffrey Leech également trouvait à redire, allant même jusqu'à proposer un nouveau nom (et peut-être une nouvelle approche) :

> *When we talk about corpus linguistics today, of course we assume that the corpus is machine-readable, and is to be investigated by means of computers. So in fact the branch of linguistics we are discussing at this Symposium should strictly be labelled "computer corpus linguistics" to distinguish it from the corpus linguistics of the pre-computer age.*[18] (Leech, 1992 : 106).

La linguistique de corpus a-t-elle existé en tant que discipline avant que l'accès à l'ordinateur ne se soit généralisé ? Tout dépend d'abord de ce qui est entendu par l'expression « linguistique de corpus ». Comme en témoignent les écrits des différents acteurs, il existe plusieurs formes de linguistique de corpus (Williams, 2010). Dans la définition de l'équipe de Randolph Quirk : « *Corpus Linguistics is the study of naturally-occurring language structure and use by first collecting samples of spoken or written language and second, analysing these samples*[19]. » Peter H. Fries, dans sa présentation des travaux de son père, Charles C. Fries, suggère que ce dernier a tout à fait sa place dans l'histoire de la linguistique de corpus, et propose la filiation suivante : « …*corpus linguistics is a reassertion of older traditions in the study of language that were current before the rise of formalist approaches*[20]. » (P. H. Fries, 2010 : 90). Charles C. Fries a travaillé sur un corpus d'enregistrements de

[15] Le terme linguistique de corpus n'a fait que des apparitions ponctuelles jusqu'à la publication d'un livre portant ce titre sous la direction d'Aarts et Meijs (1984).
[16] « *Who coined the term 'corpus linguistics'?* » Accessible <http://www.hd.uib.no/corpora/1998-1/0083.html>. Consulté le 25 juillet 2012.
[17] Liste de diffusion *Corpora*, 6 juillet 1998. Accessible <http://nora.hd.uib.no/corpora/1998-3/0006.html>. Consulté le 25 juillet 2012.
[18] Lorsque nous parlons de linguistique de corpus aujourd'hui, bien sûr, nous supposons que le corpus est accessible par machine et sera étudié par ordinateur. Donc, en fait, la branche de linguistique dont nous parlons en ce colloque devrait plutôt porter le nom de « linguistique de corpus par ordinateur » pour la distinguer de la linguistique de corpus de l'ère pré-informatique. (Trad. CC).
[19] La Linguistique de Corpus est l'étude de la structure linguistique naturelle et de sa mise en emploi, d'abord par la collecte d'échantillons de langue parlée ou écrite et ensuite par l'analyse de ces échantillons. Accessible <http://www.ucl.ac.uk/english-usage/about/index.htm>. Consulté le 7 septembre 2011.
[20] … la linguistique de corpus est une réaffirmation des anciennes traditions dans l'étude du langage qui avaient cours avant la montée des approches formalistes.

conversations téléphoniques pour produire une analyse publiée en 1952 sous le titre *The Structure of English*. Randolph Quirk reconnaît l'influence considérable des travaux de Fries, allant jusqu'à préciser :

> *Fries had of course already done innovative work on unedited manuscript English (soldiers' letters, for example). But now the new electronic recording had enabled him to do even more innovative work on unedited spoken English, and whatever its obvious deficiencies his book on The Structure of English (1952) gave me a huge buzz. From then on, I've never been without a tape recorder—and never above using a hidden mike*[21]. (R. Quirk, lors d'un entretien avec K. Brown en février 2001).

1.3. Qui a produit le premier corpus informatisé ?

Pour ce qui concerne les corpus en langue anglaise, l'histoire de la discipline atteste qu'il y avait au moins trois prétendants au titre de premier corpus informatisé. Chronologiquement, les travaux de Quirk précédaient ceux de Sinclair et ceux de Francis :

> *In a general sense, the prehistory of ICAME must include the foundations of English corpus linguistics as laid by Randolph Quirk, who established his Survey of English Usage at University College London in 1959–60, and by W. Nelson Francis, whose brainchild was the Brown Corpus, compiled at Brown University in 1961–64.*[22] Leech (2009 : 5)

Mais tout dépend de l'adjectif utilisé. D'après le site du dictionnaire du même nom, c'est le projet *COBUILD* (*Collins Birmingham University International Language Database*) qui a vu naitre le premier vrai corpus électronique : « *Professor Sinclair personally oversaw the creation of this very first electronic corpus*[23]. » [Le Professeur Sinclair a personnellement supervisé la création de ce tout premier corpus électronique.]

Comment départager ces trois candidats ? Tout d'abord, s'il est vrai que la description de la collecte des données démontre que le *Survey of English Usage*[24] a produit un véritable corpus d'un

[21] Fries avait bien sûr déjà mené des recherches innovantes à partir de manuscrits non révisés en anglais (lettres de soldats, par exemple). Mais l'avènement de l'enregistrement électronique lui a permis des recherches encore plus innovatrices sur de l'anglais parlé non filtré et, malgré quelques lacunes, son livre *The Structure of English* (1952) a été une révélation pour moi. Dès lors, je n'ai jamais été sans magnétophone – et l'utilisation d'un micro caché ne m'a jamais répugné. Consulté le 7 septembre 2011. Accessible <http://www.ucl.ac.uk/english-usage/about/quirk.htm>.

[22] Au sens large, la préhistoire de *ICAME* doit inclure les fondements de la linguistique de corpus en anglais établis par Randolph Quirk, qui avait implanté son *Enquête sur l'usage en anglais* à University College, Londres en 1959–1960, et par W. Nelson Francis, qui a conçu le Corpus *Brown*, compilé à l'université Brown de 1961 à 1964.

[23] <http://www.mycobuild.com/about-john-sinclair.aspx>. Consulté le 7 septembre 2011.

[24] De nombreuses informations concernant l'histoire de ce projet sont présentées sur le site de l'UCL. Accessible <http://www.ucl.ac.uk/english-usage/about/history.htm>. Consulté le 7 septembre 2011.

million de mots d'anglais naturel[25], en revanche, son aspect informatisé n'était pas assuré au départ.

> *The first corpus projects predated cheap computer power and mass storage. The original Survey corpus was first recorded on reel-to-reel Revox tape recorders, transcribed by hand, and then typed up, stored and annotated on paper cards*[26].

Ce n'est qu'une vingtaine d'années plus tard que ce corpus a pu être modifié et transcrit de manière à le rendre enfin accessible par ordinateur. Ce retard est en grande partie dû à la proportion de langue orale (50%) contenue dans le corpus.

Pour résumer l'histoire de la linguistique de corpus, il s'avère que le premier corpus d'anglais parlé authentique et naturel était celui produit par Charles C. Fries à l'aide de cinquante heures de conversations téléphoniques enregistrées de 1946 à 1948, avec 300 participants, pour un total de 250 000 mots (Fries, 1952). Le premier corpus accessible par ordinateur était le corpus *Brown*, composé de 500 échantillons de 2 000 mots extraits uniquement de textes écrits, publiés en 1961 (Francis et Kučera, 1964). La première collection d'anglais parlé produite par l'équipe de Sinclair a démarré en 1963 et un échantillon de 135 000 mots de conversation spontanée en anglais, avec environ 30 participants, a été transcrit à partir d'une base de données enregistrées bien plus importante, d'un million de mots environ (Sinclair, 1970, dans Krishnamurthy, 2004 : 19). Ce serait donc le premier vrai corpus électronique d'anglais parlé, même si le terme « *corpus* » n'a pas été employé à l'époque. Le corpus produit par Randolph Quirk ne serait que le deuxième corpus (*Survey of English Usage*, à partir de 1959) contenant une grande quantité d'anglais parlé (500 000 mots) à avoir été transcrit pour l'ordinateur (*London-Lund* corpus, Svartik et Quirk, 1980). Dans tous les cas, même s'il est difficile d'attribuer avec certitude le titre de « premier linguiste de corpus », il est néanmoins certain que le changement de paradigme en linguistique a commencé vers la fin des années cinquante.

---

[25] <http://www.ucl.ac.uk/english-usage/about/index.htm>. *From its inception in 1959, the Survey collected samples of naturally-occurring language for the purposes of description and analysis*. Depuis ses débuts en 1959, l'Enquête a recueilli des échantillons de langue produite dans des conditions naturelles à des fins de description et d'analyse. Consulté le 7 septembre 2011.

[26] Les premiers projets de corpus ont démarré avant la généralisation d'ordinateurs puissants et bon marché, ayant une forte capacité de stockage de masse. Le corpus de l'Enquête originale a été enregistré sur bobine avec des magnétophones Revox, transcrit à la main, puis dactylographié, annoté et stocké sur carte. <http://www.ucl.ac.uk/english-usage/about/index.htm>. Consulté le 7 septembre 2011.

1.4. Un million de mots par ordinateur : les corpus *Brown* et *LOB*

L'ensemble généralement identifié sous le nom de *Brown Corpus*[27] a été créé entre 1962 et 1964 ; il est composé de 500 échantillons de textes en anglais publiés en 1961, chacun d'une longueur d'environ 2.000 mots, pour un total de 1.014.312 mots. Les échantillons, prélevés de manière aléatoire, étaient sélectionnés pour correspondre à une série de quinze catégories définies par une conférence de linguistes à l'université de Brown en février 1963.[28]

> *The Brown corpus was not only the first electronic corpus, but the first corpus which aimed to sample a given language domain in a systematic way, so that there should be clear justification for claiming that the end-product was a representative cross-section of its domain—in this case, the domain of written American English published in a particular year.*[29] (Sampson et McCarthy, 2005 : 27)

L'objectif de Winthrop Nelson Francis et Henry Kučera était de créer un corpus d'anglais, représentatif de la langue utilisée à l'écrit aux États-Unis à l'époque où les données ont été prélevées. Le choix de n'utiliser que des textes publiés servait à garantir une certaine qualité de langue : « …*the almost pedantic accuracy the early corpus linguists strove for at the time* » […la précision presque pédante que les premiers linguistes de corpus s'efforçaient d'atteindre à l'époque] d'après John Sinclair (Krishnamurthy, 2004 : xviii). Un corpus de textes écrits et publiés sera forcément plus proche d'un usage normatif, presque prescriptif, alors qu'un corpus d'anglais de conversations libres aura un aspect nettement plus descriptif. Ce ne sont pas uniquement des considérations d'ordre pratique qui ont motivé Francis et Kučera, car Quirk et Sinclair étaient tous deux en train de créer des collections d'anglais parlé à cette même époque, comme Fries l'avait fait avant eux. Pour décrire son corpus, Francis emploie le terme « *standard* », qui peut avoir la connotation « standardisé », mais également la notion de « norme », voire de qualité à rechercher.

Cette préoccupation avec la norme pose souvent problème dans l'enseignement des langues : si la langue écrite et publiée comporte en générale peu d'erreurs, la langue orale est tout autre et le désir de ne produire que des phrases parfaitement formées et organisées conduira souvent l'apprenant à se taire. En réponse à un questionnaire sur leur capacité à s'exprimer en anglais, près de 70% des chercheurs français interrogés considéraient que leur formation initiale

[27] Le nom en entier était au départ : *A Standard Sample of Present-Day Edited American English, for Use with Digital Computers*

[28] La liste de catégories est accessible en ligne à l'adresse <http://icame.uib.no/brown/bcm.html>.

[29] Le corpus Brown n'était pas seulement le premier corpus électronique, mais aussi le premier corpus qui visait à échantillonner un domaine linguistique donné de manière systématique, de sorte qu'il y ait une justification claire pour affirmer que le produit final était un échantillon représentatif de son domaine – dans ce cas précis, le domaine de l'anglais américain écrit, publié au cours d'une année donnée.

ne les avait pas bien préparés à communiquer à l'oral en anglais (Banks, 1999 : 7). Pourtant, il suffit d'étudier un corpus d'anglais parlé en conditions réelles pour démontrer à quel point même les locuteurs natifs ayant un bon niveau d'étude se répètent, se trompent et se reprennent, lorsqu'ils parlent de manière libre et non préparée.[30]

Toutefois, la création d'un corpus libre de droits dans le cadre de recherches linguistiques est un travail de longue haleine et ce premier corpus établi de manière raisonnée a servi de modèle à toute une série de corpus comparables. La liste des rubriques pour le corpus *Brown* a été établie par un petit groupe de linguistes chevronnés : John B. Carroll, W. Nelson Francis, Philip B. Gove, Henry Kučera, Patricia O'Connor et Randolph Quirk.

Après avoir assisté à cette conférence aux États-Unis, Quirk a poursuivi à Londres la création d'un corpus de même taille (un million de mots). Néanmoins, il convient de préciser que le *Survey of English Usage* est composé de moins d'échantillons que le corpus *Brown*, mais de taille plus grande (200 échantillons de 5 000 mots), avec 50% d'oral et 50% d'écrit. La partie orale de ce corpus contient des conversations et des monologues enregistrés, souvent sans que les participants en aient été avertis, entre 1953 et 1987. Cette partie a été informatisée par l'équipe de Jan Svartik, à l'université de Lund, et la version finale de ce corpus d'anglais oral, le corpus *London-Lund*, contient 100 ensembles de 5 000 mots, soit 500 000 mots.

Un corpus sur le modèle *Brown*, composé des textes écrits en anglais britannique, publiés en 1961, a été élaboré au départ à l'université de Lancaster, par Geoffrey Leech, à partir de 1973-4 avec la participation de Stig Johansson, qui a ensuite obtenu un poste à l'université d'Oslo, puis avec l'aide technique de Knut Hofland, de l'université de Bergen, à partir de 1977 (Leech et Johansson, 2009). Ce corpus est plus connu sous le nom de corpus *LOB* (*Lancaster/Oslo–Bergen*).[31]

Ces trois corpus, le *Brown*, le *London-Lund* et le *LOB*, ont été réunis pour former l'ensemble *ICAME* : *International Computer Archive of Modern English*. Il s'agit d'un ensemble de trois corpus compatibles plutôt que d'un seul corpus, car la représentativité n'existe qu'à l'intérieur de chaque corpus. Dans l'esprit des concepteurs, il s'agissait de corpus de référence, c'est-à-dire des corpus représentatifs non pas d'un type de langage spécifique (les « *restricted languages* » prônés par J. R. Firth), mais au contraire d'une langue courante et générale, apte à servir de modèle authentique des usages contemporains, pour l'anglais américain écrit, pour l'anglais britannique écrit et pour l'anglais oral (surtout britannique). Toutefois, si le corpus *Brown* a pu être réalisé

[30] *cf. A Handbook of Spoken Grammar*, (2011) par Kenneth Paterson, Caroline Caygill, et Rebecca Sewell.
[31] <http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM>. Consulté le 9 septembre 2011.

suffisamment rapidement pour servir de modèle d'un usage contemporain (textes publiés en 1961, corpus rendu accessible à partir de 1964), la première version du corpus *LOB* n'était disponible qu'à partir de 1978, soit dix-sept ans après la parution des textes qu'il contient. Néanmoins, de nombreuses études, notamment concernant la syntaxe et la grammaire, ont été menés à partir de ces données.

D'autres corpus ont été construits à partir de ce même modèle, afin de pouvoir mener des recherches comparatives et diachroniques. En 1991, l'équipe de l'université de Freiburg, dirigé par Christian Mair, a entrepris la compilation du corpus *F-LOB*, composé de textes en anglais britannique publiés en 1991 puis, à partir de 1992, la même équipe a produit le corpus *Frown*, composé de textes en anglais américain publiés en 1992. L'équipe de Lancaster a produit le corpus *BLOB-1931*, sous la direction de Geoffrey Leech et de Paul Rayson, avec des textes en anglais britannique de 1931 et également le *BE06*, sous la direction de Paul Baker, avec des textes en anglais britannique de 2003 à 2008, dont la majorité a été publiée en 2006. Cet ensemble de corpus compatibles, sur un même modèle, permet des recherches comparatives, diachroniques mais surtout grammaticales et syntactiques (Leech, 2011) car, en conformité avec la loi de Zipf (1935), les mots lexicaux sont les plus nombreux en tant que classe, mais les moins fréquents individuellement. Toutefois, certaines comparaisons qualitatives restent envisageables, même pour les mots lexicaux.

1.5. Les grands corpus de référence

Si un corpus d'un million de mots permet déjà de nombreuses analyses par ordinateur, qui ne seraient pas forcément envisageables sans cet outil, qu'en est-il de corpus plus grands ? L'équipe de Lancaster a continué à produire des corpus compatibles, construits sur un même modèle, de même taille, afin de pouvoir mener des recherches comparatives. D'autres équipes ont choisi d'autres options, notamment à partir du moment où la puissance de calcul et de stockage des ordinateurs a rendu possible la création de corpus de référence, de très grande taille.

1.5.1 *Le British National Corpus* (*BNC*)

Le premier des très grand corpus et le plus connu des corpus de référence pour l'anglais est le *British National Corpus* ou *BNC*. Ce corpus contient presque cent millions de mots,[32] soit cent fois la taille du corpus *Brown*, avec en plus une proportion non-négligeable de documents oraux transcrits (environ 10 %, soit dix millions de mots). Le corpus est composé de 4049 textes

[32] Le *BNC* contient très précisément 96.986.707 mots, d'après le site qui lui est consacré. Accessible <http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=numbers>. Consulté le 10 septembre 2011.

ou échantillons dont la longueur ne dépasse pas 45 000 mots, afin de ne pas déséquilibrer le corpus. Le travail de compilation a commencé en 1991 et une première version du corpus est parue en 1994.[33]

Un aussi grand corpus ne peut être consulté sans passer par une interface qui permet d'organiser les résultats ; le *BNC* est un corpus étiqueté, comme les corpus de la famille *Brown*, ce qui permet des requêtes très ciblées, mais en même temps empêche certaines autres requêtes d'aboutir. De plus, l'étiquetage du *BNC* a été produit de manière automatique et, de ce fait, contient parfois des anomalies, voire des erreurs : « *the tagging provided by many corpora can be at the same time a boon, an obstacle and an object of interest in its own right.*[34] » (Denison, 2007).

Le *BNC* est un corpus représentatif de la langue anglaise dans la deuxième moitié du vingtième siècle. C'est un monument dans l'histoire de la linguistique de corpus, mais de toute évidence il ne pourra plus être utilisé pour servir de référence à l'anglais du vingt-et-unième siècle, sauf de manière diachronique.

1.5.2 L'interface *VIEW* : Mark Davies

Plusieurs très grands corpus sont maintenant accessibles en ligne. Pour des raisons de droits des auteurs, l'accès à certains de ces corpus est payant et le tarif est parfois prohibitif pour un chercheur isolé. Il existe pourtant toute une série de corpus disponibles gratuitement en ligne à des fins de recherches linguistiques. Il s'agit de la série de corpus produits par l'équipe de Mark Davies à l'université de Brigham Young (*BYU*).[35] Pour ce qui concerne l'anglais, il existe trois bases de données : le *Corpus of Contemporary American English* (*COCA*), 425 millions de mots, de 1990 à 2011 ; le *Corpus of Historical American English* (*COHA*), 400 millions de mots, de 1810 à 2009 ; le *TIME Magazine Corpus of American English,* 100 millions de mots, de 1923 à 2006 (Davies, 2007–).

Neil Millar a récemment présenté une étude diachronique de l'utilisation de modaux dans le corpus *TIME* (2009), dont les résultats contredisaient quelque peu des études basées sur les corpus de la famille *Brown*. Geoffrey Leech (2011) a publié une réponse qui confirme les résultats qu'il avait déjà obtenus en 2003. L'un des problèmes de la comparabilité des études sur corpus est qu'il est essentiel de s'assurer que la comparaison est fondée sur des éléments mesurés de la même manière. L'utilisation d'une même interface devrait faciliter ce type de recherches.

[33] <http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=numbers>. Consulté le 11 septembre 2011.
[34] L'étiquetage fourni avec de nombreux corpus peut être à la fois une aubaine, un obstacle et un objet d'intérêt en tant que tel.
[35] <http://corpus.byu.edu/>. Consulté le 11 septembre 2011

Pour faciliter les comparaisons de l'anglais britannique avec ces très grands corpus d'anglais américain, le *BNC* est également disponible par le biais de la même interface, qui permet plusieurs types de requêtes, y compris la co-sélection, la recherche de collocats et la fréquence par sous-corpus et même par période chronologique. En revanche, pour des raisons de droit d'auteur, il n'est pas possible d'accéder aux textes eux-mêmes, seulement à de courts extraits.

1.5.3. Internet et corpus : Kilgarriff et Baroni

L'avènement d'internet a vu naître la notion de « *Web as Corpus* » [La Toile en tant que Corpus] et le développement d'outils pour permettre la création de corpus à partir des pages internet (Kilgarriff et Grefenstette, 2003 ; Baroni et Bernardini, 2006). Aujourd'hui, il serait absurde de nier l'importance d'internet pour la création de corpus, mais il serait tout aussi incohérent d'imaginer qu'il soit possible de créer un corpus simplement en téléchargeant de nombreuses pages de documents sans tri. Des outils tels que *WebBootCaT* permettent de créer un corpus à partir d'internet en utilisant des mots clés, en général par triplets (Baroni et al, 2009), mais cette première récolte nécessite encore beaucoup de travail avant de pouvoir prétendre au nom de corpus.

1.6. Au-delà de la technologie, une véritable science du langage

Après avoir retracé l'histoire de la linguistique de corpus au cours du vingtième siècle, depuis les premières collections de textes jusqu'aux grands corpus informatisés, il est possible de préciser quelques principes essentiels qui se dégagent de ce parcours.

Pour s'assurer de la valeur scientifique de l'étude et de son objectif linguistique, le travail de recherche doit être mené de façon rigoureusement documentée, que ce soit dans la collecte de données, ou dans la composition du corpus, ou enfin dans la configuration des outils employés pour les analyses, afin de garantir un maximum de reproductibilité à l'étude.

Enfin, il existe plusieurs manières de constituer un corpus et plusieurs types de recherches peuvent être menés à partir de ces corpus. La section suivante va retracer le développement de la linguistique de corpus contextualiste et son orientation axée de plus en plus sur le lexique plutôt que sur la grammaire.

2. Au-delà de la grammaire, l'étude du lexique

La deuxième partie de cette étude regardera plus en détail les travaux de John McHardy Sinclair et l'école contextualiste britannique avec la mise en place des notions fondamentales pour la linguistique de corpus et l'approche inductive.

D'abord, l'évolution de la notion de « *collocation* » sera retracée, notamment à travers les recherches de H. E. Palmer et de J. R. Firth. Les notions de « *concordance* » et de « *context* » seront également mises en lumière, ainsi que la quête pour identifier les unités étendues de sens (« *extended units of meaning* »). La dernière partie présentera une analyse de l'expression « *silly ass* », qui a été utilisée par J. R. Firth comme un exemple représentatif du phénomène de collocation.

2.1. La notion de collocation, avant les corpus informatisés

Cette notion, fondamentale pour la linguistique de corpus, a déjà été mentionnée dans l'introduction de cet article, mais c'est dans cette section que l'histoire de son développement sera présentée. Les travaux s'intéressant à ce phénomène, avant l'avènement des corpus informatisés, sont surtout ceux de H. E. Palmer et ceux de J. R. Firth.

2.1.1. La collocation, vue par H. E. Palmer

La notion de collocation a d'abord été développée dans les années 1930 par Harold E. Palmer (Smith, 1999). Pour ce dernier, parmi les 20 000 mots les plus courants en anglais, seulement mille d'entre eux posaient de réels problèmes, sauf pour ce qui concerne la prononciation (H. E. Palmer, 1955 [1938] : iii). Il leur a donc consacré *A Grammar of English Words*, livre qui se situe à la frontière entre grammaire et dictionnaire, qu'il décrit ainsi : « … *a sort of no-man's land in which reside the great majority of those points that perplex those to whom English is a foreign language* » (Palmer, 1955 [1938] : v). [… une sorte de zone neutre où résident la grande majorité de ces points qui embarrassent ceux pour qui l'anglais est une langue étrangère.]

La solution proposée par Palmer dans son livre est exprimée dans le sous-titre de la section suivante : « *Richness and Abundance of Examples* » [Richesse et abondance d'exemples]. Pour certains mots, toute une série d'exemples est proposée, d'autres contiennent seulement les informations permettant de distinguer entre les différents emplois du mot en question. Voici à

titre d'exemple, la présentation du mot « *earth* », qu'il écrit sans majuscule (Palmer, 1955 [1938] : 50).

> EARTH
> earth [ə:θ], earths [ə:θs], *n.*
> 1. = the terrestrial globe
> How far is the earth from the sun?
> **¶ on earth**
> the greatest on earth.
> 2. = Soil, ground.
> *a. Uncountable.* Bury it in the earth.
> *b. Countable* [ə□□s]
> different earths = sorts of earth [soil].

Il convient de noter que seuls les mots « *terrestrial globe* » et « *soil* » ne figurent pas dans ce dictionnaire de grammaire. Pour le linguiste, il est évident que la difficulté d'un mot pour un locuteur non-natif dépend énormément de la langue maternelle de ce locuteur. Certains mots anglais d'origine latine, considérés difficiles pour des locuteurs natifs, sont parfaitement transparents pour des locuteurs dont la langue maternelle est également d'origine latine, comme les francophones, par exemple. Palmer travaillait au Japon et le japonais contient peu de mots apparentés avec l'anglais, sauf par le biais des emprunts, qui seraient au nombre de 1.500 environ, et fonctionneraient surtout comme des groupes indissociables (Nagasawa, 1958 : 53). Palmer devait avoir parfaitement conscience de ce problème, car il précise, au sujet des définitions utilisées dans son livre :

> *These are intended for the use of the teacher or for the student whose recognition vocabulary is already fairly extensive, for in many cases the definition or paraphrase contains words much rarer than the word to be defined.*[36] (Palmer, 1955 [1938] : ix-x).

Palmer propose deux éléments permettant de mieux cerner la notion de collocation et de différencier les collocations des autres locutions. La première définition concerne le fonctionnement dans la langue que l'apprenant cherche à acquérir, alors que la deuxième signale leur fonctionnement lors d'une traduction vers la langue maternelle de l'apprenant. Pour lui, cette notion comporte un aspect pratique, voire pragmatique, pour l'apprenant, car une collocation est «*… a succession of two or more words that may best be learnt as if it were a single word »* (Palmer 1955

[36] Ils sont destinés à l'usage de l'enseignant ou l'étudiant dont le vocabulaire de reconnaissance est déjà assez étendu, car dans de nombreux cas la définition ou paraphrase contient des mots beaucoup plus rares que le mot à définir.

[1938] : x) […une suite de deux mots ou plus qui serait plus facilement apprise en la considérant comme un seul mot.]

La différence entre une collocation et une locution est présentée ainsi :

> *While collocations are comparable in meaning and function to ordinary single 'words' (and indeed are often translated by single words in the student's mother-tongue), phrases are more in the nature of conversational formulas, sayings, proverbs, etc.*[37] (Palmer 1955 [1938] : xi)

S'il explique en détail la différence entre ces deux éléments, Palmer choisit toutefois de ne pas utiliser le terme « *idiom* » et justifie ainsi son choix :

> *What are usually called 'idioms' are generally nothing other than (a) collocations, (b) phrases and sayings, (c) rarer semantic varieties of words and collocations, (d) peculiar construction patterns and, in short, any word or form of wording that is likely to puzzle a foreign student*[38]. (Palmer, 1955 [1938] : xii).

Il s'agit de bien plus qu'un simple problème de terminologie. Ce qui est en jeu est la conception fondamentale de la langue. Les mots-clés sont *collocation* et *pattern*. Il ne s'agit pas d'universels linguistiques. Palmer propose des solutions pour aider l'apprenant à gérer l'une des spécificités de l'anglais, plus particulièrement spécifique à ses mille mots les plus fréquents, en commençant par faire la liste de ces difficultés.

> *These words are "difficult" for various reasons—reasons that are not apparent to those to whom English is the mother-tongue:*
> *(1) each word may belong to two or more "parts of speech"*;*
> *(2) each word may have two or more meanings and "stretches of meaning," in some cases the number is very considerable;*
> *(3) each word may enter into two or more "sentence-patterns" occupying its own particular place in the sentence (among the verbs alone nearly 30 distinct "patterns" are to be found);*
> *(4) each word may have several inflected forms and derivatives, many of them being irregular in form and meaning;*
> *(5) each word may enter into a large number of* collocations *and* phrases *(successions of two or more words the meaning of which can hardly be deduced from a knowledge of their component words); some of these, again, may each have two or more meanings and stretches of meaning;*

[37] Alors que les collocations sont comparables par le sens et par la fonction à des « mots » ordinaires simples (et d'ailleurs sont souvent traduits par un seul mot dans la langue maternelle de l'étudiant), les locutions sont plus de l'ordre des formules de conversation, des dictons, proverbes, etc.

[38] Ce qu'on appelle habituellement des « idiomes » ne sont généralement rien d'autre que (a) des collocations, (b) des expressions et dictons, (c) des variétés sémantiques rares de mots et de collocations, (d) des modes de construction particulière et, en bref, tout mot ou forme d'expression susceptible d'intriguer un étudiant étranger.

> *(6) each word may be a component part of one or more "compounds" (or "compound words") the meaning of which can hardly be deduced from a knowledge of the component words;* (Palmer, 1955 [1938] : iv).

> Ces mots sont « difficiles » pour diverses raisons, peu perceptibles pour ceux qui ont pour langue maternelle l'anglais :
> (1) chaque mot peut appartenir à plusieurs « parties du discours »* ;
> (2) chaque mot peut avoir deux ou plusieurs significations et « des étendues de sens », dans certains cas, le nombre en est considérable ;
> (3) chaque mot peut faire partie de plusieurs « modèles » ayant chacun sa propre place dans la phrase (rien que pour les verbes on peut trouver près de 30 « modèles » distincts) ;
> (4) chaque mot peut avoir plusieurs formes fléchies et dérivées, beaucoup d'entre elles sont irrégulières en forme et en signification ;
> (5) chaque mot peut entrer dans un grand nombre de *collocations* et d'*expressions* (successions de deux ou plusieurs mots dont le sens peut difficilement être déduit simplement en connaissant les mots qui les composent), certains d'entre eux, encore, peuvent avoir chacun deux ou plusieurs significations et étendues de sens ;
> (6) chaque mot peut faire partie d'un ou plusieurs « composés » (ou « mots composés ») dont le sens peut difficilement être déduit par la connaissance des mots dont il est composé ;

Chaque catégorie est suivie d'un renvoi à une série d'exemples. Pour le renvoi * en (1), il précise : « *Thus the word* since *is a preposition, an adverb and a conjunction.* » (Palmer, 1955 [1938] : iv) [Ainsi le mot *since* est une préposition, un adverbe et une conjonction.]

Pour Palmer, il ne s'agit pas de l'ensemble des mots en anglais, seulement de certains mots parmi les plus fréquents. Ce principe s'apparente à la notion de « *delexicalisation* » [la délexicalisation], et est également en cohérence avec la loi de Zipf et la notion du moindre effort (Zipf, 1949). Pour Sinclair, dont les travaux s'inscrivent dans la lignée de ceux de Palmer, ce phénomène était le fruit de l'usure. Dans l'hommage rendu à Sinclair dans le journal *Euralex*, Patrick Hanks précise : « *He described the semantic lightness of frequent words as 'the blue jeans principle': the more you use them and wash them, the more the colour washes out.*[39] » [Il a décrit la légèreté sémantique des mots fréquents comme « le principe du jean » : plus vous l'utilisez et plus vous le lavez, plus la

[39] Hommage lors du décès de Sinclair. Consultée le 15 septembre 2011. Accessible <http://www.euralex.org/elx_newsletters/Bogaards%20-%202007%20-%20SUMMER%202007%20EURALEX%20NEWSLETTER.pdf>.

couleur se délave.] La « couleur » du mot est sa signification d'origine ; la collocation est l'environnement linguistique du mot qui permet de constater l'apparition de nouveaux sens. Pour les mots, comme pour les vêtements en jean, le même phénomène se produit : plus ils sont utilisés, plus ils deviennent souples.

Toutefois, la collocation n'est pas une propriété réservée uniquement aux mots fréquents, contrairement à ce que laisse supposer l'étude de Palmer. Combien de mots ont des collocations ? Pour approfondir cette question, il convient maintenant de regarder comment ce phénomène a été traité dans les travaux de J. R. Firth.

2.1.2 La notion de collocation chez Firth

Pour J. R. Firth, il existait pour les sons une forme de « *contextual distribution* » [distribution contextuelle]. Les voyelles pouvaient avoir diverses fonctions majeures : situationnelle, lexicale, morphologique et phonaesthétique[40] (Firth, 1957[1935] : 37). Il restait plutôt hostile à certaines explications trop complexes :

> *If sounds are described, classified, and explained by this statistical contextual technique, most contemporary theories of elision, coalescence, and assimilation will be seen to be confusing and, what is much more to the point, entirely unnecessary*[41]. (Firth, 1957[1935] : 37).

Ainsi, des analyses qui se contentent de décrire les données de manière statistique suffisent à expliquer les phénomènes constatés, sans avoir besoin de déborder du cadre strictement linguistique. Les principes prônés par Firth peuvent s'appliquer de manière équivalente aussi bien aux sons qu'au sens :

> *Nevertheless a pragmatic functionalism seems to me to lead to much clearer definition, and to the statement and explanation of facts, without having to postulate a whole body of doctrine in an elaborate mental structure such as is derived from de Saussure*[42]. (Firth, 1957[1935] : 36)

La collocation est une forme de distribution des mots et peut être très utile lors de différents types d'analyse, notamment lors des analyses statistiques. La citation suivante est

[40] Cette fonction est un lien entre le son et le sens et s'applique aussi à certains groupes de consonnes, comme la 'liquidité' de « *slippy* » et « *sloppy* ». (Firth, 1957[1935] : 39).

[41] Si les sons sont décrits, classés et expliqués par cette technique statistique contextuelle, la plupart des théories contemporaines de l'élision, la coalescence, et l'assimilation se révéleront prêter à confusion et, ce qui est beaucoup plus important, tout à fait inutiles.

[42] Néanmoins un fonctionnalisme pragmatique me semble conduire à une définition beaucoup plus claire, et à la déclaration et l'explication des faits, sans avoir à postuler tout un corps de doctrine dans une structure mentale complexe telle que celle dérivée de Saussure.

donnée en entier, bien qu'elle soit très longue, car elle fournit une base assez précise pour l'analyse des collocations :

> *Just as phonetic, phonological, and grammatical forms well established and habitual in any close social group provide a basis for the mutual expectancies of words and sentences at those levels, so also the study of the usual collocations of a particular literary form or genre or of a particular author makes possible a clearly defined and precisely stated contribution to what I have termed the spectrum of descriptive linguistics, which handles and states meaning by dispersing it in a range of techniques working at a series of levels.*
>
> *The statement of meaning by collocation and various collocabilities does not involve the definition of word-meaning by means of further sentences in shifted terms. Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words.*[43] (Firth, 1957 [1951] : 193-4)

La collocation pour Firth se situe donc au niveau des relations constatées entre les mots. Ces relations s'établissent à l'intérieur d'un groupe social étroit ou d'une forme particulière de littérature. Il conviendrait de rajouter à ces restrictions le cadre temporel, qui peut avoir énormément d'importance, notamment pour la langue orale. L'expression « *mutual expectancies* » [attirances mutuelles ou attentes réciproques] suggère que les mots s'attirent de manière réciproque autant pour des raisons de sonorité que pour des raisons d'ordre grammatical ; la collocation est une relation de cooccurrence qui intervient au niveau sémantique, mais *a postiori*.

2.2 Sinclair, le rapport *OSTI* et l'étude du lexique

Dès le début de sa carrière, à l'université d'Edimbourg, John McHardy Sinclair s'est intéressé à l'étude de la langue orale. La collecte d'une grande quantité de données orales, enregistrées à Edimbourg et à Londres, a servi de base pour un projet mené pour le *UK Government Office for Scientific and Technical Information* (*OSTI*), dont le rapport final, *English Lexical Studies,* terminé en janvier 1970, n'a été publié qu'en 2004. Dans l'entretien avec Wolfgang Teubert qui sert d'introduction à cette édition tardive du rapport *OSTI*, Sinclair explique que

[43] Tout comme les formes phonétiques, phonologiques, et grammaticales, bien établies et habituelles dans tout groupe social étroit, fournissent une base pour les attirances mutuelles des mots et des phrases à ces niveaux, de même l'étude des collocations habituelles d'une forme littéraire particulière ou d'un genre ou d'un auteur particulier permettent une contribution bien définie et précisément formulée à ce que j'ai appelé le spectre de la linguistique descriptive, qui manipule et exprime le sens en le dispersant à travers une gamme de techniques fonctionnant à différents niveaux.

La déclaration de la signification par collocation et par collocabilités différentes n'implique pas la définition du sens des mots au moyen d'autres phrases avec des termes décalés. La signification par collocation est une abstraction au niveau syntagmatique et n'est pas directement concernée par l'approche conceptuelle ou idéationnelle de la signification des mots.

l'objectif prioritaire était « *research into the lexis of English* » [recherches sur le lexique de l'anglais] et que leur approche était innovante car elle s'intéressait à la langue orale et à la collocation (Krishnamurthy, 2004 : xviii).

Bien avant la publication du rapport OSTI, dans un recueil en l'honneur de J. R. Firth, Sinclair expliquait déjà comment il fallait entreprendre l'étude du lexique, en précisant : « *The two interpenetrating ways of looking at language form are* ***grammar*** *and* ***lexis****.* » (Sinclair, 1966 : 411). [Les deux façons de regarder la forme linguistique qui s'interpénètrent sont la **grammaire** et le **lexique**.] Il mettait aussi en garde contre les difficultés de la tâche d'analyse lexicale, liées à la spécificité des éléments lexicaux.

> *At the present time, lexical statements look very much weaker than statements made using the precise and uncompromising machinery of grammar. There is a much less elaborate framework to lexical descriptions and much less certainty in the statements.*[44] (Sinclair, 1966 : 411)

Il est évident que l'étude du lexique posait plus de problèmes que l'étude des éléments grammaticaux, tant qu'il n'existait pas d'ordinateurs suffisamment puissants pour pouvoir effectuer des calculs sur une très grande quantité de données. Il est facile d'oublier les progrès réalisés dans ce domaine depuis les années soixante.

2.2.1. Arguments concernant la lemmatisation et les anti-dictionnaires

Comme la loi de Zipf a démontré, les éléments grammaticaux sont toujours les premiers mots du classement et ils sont considérablement plus fréquents que les éléments lexicaux, car la fréquence d'un mot est inversement proportionnelle à son rang (Zipf, 1935). L'une des premières découvertes de l'équipe de Sinclair étaient que ces mots grammaticaux possèdent aussi leurs propres collocations, tout comme les mots lexicaux, ce qui plaide contre l'utilisation d' anti-dictionnaires (« *stop-lists* »).

C'est justement parce que la classe des éléments lexicaux contient beaucoup plus d'éléments que celle des éléments grammaticaux qu'il est difficile d'en collecter un échantillon suffisamment grand pour permettre l'analyse exhaustive d'un élément donné. La lemmatisation était considérée comme une manière d'augmenter le nombre d'occurrences des mots lexicaux, en regroupant les mots d'une même famille ensemble, mais les résultats obtenus pour le rapport *OSTI* ont démontré que les différentes formes d'un même lemme n'avaient que rarement les

[44] A l'heure actuelle, les descriptions du lexique ont l'air d'être bien plus faibles que les déclarations faites à partir des mécanismes précis et sans concession de la grammaire. Il existe un cadre beaucoup moins élaboré pour des descriptions lexicales et il y a beaucoup moins de certitude dans ces déclarations.

mêmes collocations, confirmant ainsi le bien-fondé des procédés suivis par R. C. Eldridge, dès 1911. C'est pour cela que Sinclair préconise de ne pas imposer la lemmatisation de manière irréversible ; pour obtenir des résultats à une échelle plus fine et plus précise, il faut pouvoir mener des analyses sur chaque forme lexicale indépendamment.

De toute manière, la loi de Zipf implique aussi que la description des éléments rares risque de poser moins de problème que celle des éléments fréquents, car « *different meanings of a word will tend to be equal to the square root of its relative frequency.* » (Zipf, 1945 : 255). [le nombre de significations différentes d'un mot aura tendance à être égal à la racine carrée de sa fréquence relative]. Ainsi, un mot lexical peu fréquent aura peu de sens différents, ce qui implique également qu'il se trouvera souvent dans les mêmes contextes, ou du moins dans des contextes similaires, car les différents sens des homophones sont généralement révélés par leur contexte.

2.2.2. L'étiquetage du corpus : *To tag or not to tag?*

Ce changement d'optique a progressivement donné de plus en plus d'importance aux mots lexicaux mais les a souvent associés à la grammaire, avec des modalités portant le nom de « *pattern grammar* » [la grammaire paradigmatique] développé par l'équipe *COBUILD* (Francis, Hunston et Manning, 1996), puis ce qu'il convient d'appeler « *local grammar* » [la grammaire locale] proche de ce que Halliday a appelé « *lexicogrammar* » [lexicogrammaire] (Halliday, 1961 ; Halliday et Matthiessen, 2004). Mais pour Halliday la grammaire reste prioritaire et il exprime le rapport entre les deux éléments de la manière suivante : « *He* [the grammarian] *would like to turn the whole of linguistic form into grammar, hoping to show that lexis can be defined as 'most delicate grammar'.* (Halliday, 1961 : 267). [Lui [le grammairien] aimerait convertir l'ensemble de la forme linguistique en grammaire, avec l'espoir de montrer que le lexique peut être défini comme « la grammaire la plus délicate ».].

Il est vrai que, pour beaucoup de linguistes, il est très difficile d'abandonner la notion de grammaire. Lors de sa participation au projet *COBUILD*, le lexicographe John Williams a réalisé un travail de recherche[45] en sciences sociales sur la manière dont la rédaction du livre *Collins COBUILD Grammar Patterns 1: Verbs*, publié en 1996, a été négociée. Il propose le commentaire suivant pour l'un des entretiens menés avec cette équipe de trois personnes (Gill Francis, Susan Hunston et Elizabeth Manning) :

[45] Il s'agit d'une thèse de Master, soumise en 1998, dont le titre est *Grammatical constructions and social construction: a case study in the practice of linguistic 'science'* [Constructions grammaticales et construction sociale: étude de cas dans la pratique de la «science» de la linguistique].

> *What is striking* [.../...] *is that the grammarians appeared to be on the cusp of a major paradigm shift (Kuhn 1962, 1970) - the abandonment of familiar concepts such as 'subject' and 'object'. Yet, for pedagogical and commercial reasons, this development was nipped in the bud by senior figures within COBUILD - including the very scholar whose pioneering empirical methods had contributed to the incipient paradigm shift*[46].

Comme les témoignages récoltés par John Williams le confirment, il est extrêmement difficile de mener des recherches de manière totalement inductive, même si c'était l'approche prônée par l'équipe de Birmingham, lors de la création du dictionnaire *COBUILD*. En effet, il existe deux manières d'aborder la linguistique de corpus : « *corpus-based and corpus-driven approaches* » [les approches hypothético-déductive et inductive] (Tognini-Bonelli, 2001 : 17). La première approche « *uses corpus evidence mainly as a repository of examples to expound test or exemplify given theoretical statements.* » (Tognini-Bonelli, 2001 : 10) [se sert des éléments de preuve fournis par le corpus principalement comme référentiel d'exemples pour exposer, tester ou illustrer des propositions théoriques préétablies.], alors que la deuxième approche : « *builds up the theory step by step* in the presence of *the evidence.* » (Tognini-Bonelli, 2001 : 17) [construit la théorie pas à pas, face aux éléments de preuve.]. Dans le premier cas, le corpus permet de vérifier des hypothèses concernant le fonctionnement de la langue, et dans le deuxième cas, c'est à partir de l'analyse du corpus qu'émergent les hypothèses. Toutefois, il est assez difficile pour le linguiste de prendre la position d'un observateur neutre : comment se libérer complètement de tout *a priori* concernant la langue ? Bronislaw Malinowski (1923) avait déjà soulevé ce problème lors de ses travaux sur le kiriwina, en soulignant l'importance d'analyser une langue de l'intérieur, sans lui imposer une structure externe.

Pour l'instant, l'étiquetage contribue encore à cette division fondamentale dans la manière d'aborder le corpus, car l'étiquetage est basé sur une vision préexistante des catégories linguistiques, un découpage en catégories parfois problématique quand le corpus n'est pas pris comme point de départ.[47] Mais les choses sont appelées à changer. D'après David Denison, il existe d'autres possibilités pour l'utilisation des étiqueteurs à l'avenir :

---

[46] Ce qui nous frappe ici [.../...] est que les grammairiens semblaient être au seuil d'un changement de paradigme majeur (Kuhn 1962, 1970) - l'abandon des concepts familiers tels que « sujet » et « objet ». Pourtant, pour des raisons pédagogiques et commerciales, ce développement a été tué dans l'œuf par leurs supérieurs au sein de *COBUILD* - y compris par le chercheur même qui avait contribué, par ses méthodes empiriques pionnières, à ce changement de paradigme naissant.

[47] Communication personnelle d'Adam Kilgarriff, le 13 juin 2012, à propos d'un futur outil lexicographique pour le français : « *I'm not sure if we'll include the grammatical categories as part of the output, it might be that they are both error-strewn and not needed.* » [Je ne sais pas si nous allons inclure les catégories grammaticales lors de la restitution, il se pourrait qu'elles soient à la fois parsemées d'erreurs et non-nécessaires.]

> *As a final question, could we actually* ***reveal*** *language change by tagging procedures, rather than merely playing catch-up after the event?*
> *Perhaps we can, but only if -*
> *Diachronic linguists produce a typology of systematic category changes. -Synchronic linguists build such possibilities into the lexicons and taggers that they use.*[48] (Denison, 2007)

Un étiqueteur qui part du corpus pour déceler les catégories à utiliser serait un outil précieux pour une approche vraiment inductive de la linguistique de corpus, mais comme l'une des lexicographes de l'équipe *COBUILD* l'avait précisé, au cours de la série d'entretiens avec John Williams : « *If the data, if the thing itself was so empirically objective, then it wouldn't have needed us to do it. A computer could have done it*[49]. » Malheureusement, sans verser dans le néo-luddisme, les avancées technologiques conduisent souvent à remplacer l'humain par la machine et la tentation est grande de ne pas tout abandonner à l'informatique, afin de conserver son poste le plus longtemps possible.

2.2.3. L'empan optimal, cinq à gauche et quatre à droite

L'une des avancées significatives apportées par l'informatique est qu'elle a permis de mettre en œuvre la recherche de collocats de manière systématique. En linguistique de corpus, les collocats ne se trouvent pas forcément juxtaposés, et l'écart entre deux collocats est très variable. C'est à force d'expérimentation sur le corpus[50] que le meilleur empan a pu être décelé. Au-delà d'un certain seuil, le rapport collocatif n'est plus significatif, et ce rapport doit être calculé pour chaque langue. Dans son entretien avec Teubert, Sinclair précise, au sujet de l'empan optimal :

> *We recalculated it a few years ago on the basis of a much larger corpus of English and came to almost the same result, finding that five words to the left and four words to the right might result in a slight improvement of semantic relevance.*[51] (Krishnamurthy, 2004 : xix)

C'est pour tenir compte de ce nouveau calcul que l'empan de cinq mots à gauche et quatre à droite a été utilisé pour les analyses présentées ici. Il reste néanmoins toujours possible

[48] Une dernière question, pourrions-nous vraiment révéler des changements de langue par des procédures d'étiquetage, plutôt que de ne parvenir à les rattraper qu'après coup ? Peut-être que nous le pourrons, mais seulement si
- les linguistes diachroniciens produisent une typologie des changements de catégorie systématiques
- les linguistes synchroniciens intègrent de telles possibilités dans les lexiques et étiqueteurs qu'ils utilisent.

[49] Si les données, si la chose elle-même avait été si empiriquement objective, alors il n'y aurait pas eu besoin de nous pour faire ce travail. Un ordinateur aurait pu le faire.

[50] Le chapitre 3 du rapport OSTI, écrit par Robert Daley en 1972, présente en détail le raisonnement statistique qui a permis d'optimiser l'empan. (Krishnamurthy, 2004 : 34-56)

[51] Nous l'avons recalculé il y a quelques années sur la base d'un corpus d'anglais beaucoup plus vaste et avons obtenu pratiquement le même résultat, constatant qu'un empan de cinq mots à gauche et quatre à droite pourrait entraîner une légère amélioration de la pertinence sémantique.

d'employer l'expression consacrée, « *nine-word-window of collocative power* » (Louw, 2000) [la fenêtre de neuf mots, puissance collocative], à condition de préciser qu'il s'agit d'une fenêtre de neuf mots autour du nœud central, cinq mots à gauche et quatre à droite.

2.2.4. Collocations ascendantes et descendantes

L'analyse informatisée d'un corpus permet d'obtenir rapidement la liste des mots qu'il contient, par ordre de fréquence. Chaque mot aura donc une certaine fréquence, et les mots grammaticaux, invariables, seront les plus fréquents. Ainsi, le mot le plus fréquent dans la plupart des corpus représentatifs de la langue anglaise dans son ensemble est l'article défini « *the* »,[52] mais dans le cas du français, la lemmatisation aura un impact non-négligeable. Dans un corpus non-lemmatisé, le mot le plus fréquent serait le mot « de »,[53] mais dans un corpus lemmatisé ce serait l'article défini « le, la, les ».[54]

La collocation est un rapport entre deux mots qui ne sont pas forcément contigus. La collocation ascendante décrit le rapport entre un mot moins fréquent et un mot plus fréquent et révèle plutôt la colligation (une association avec une catégorie grammaticale). Pour la collocation descendante, le rapport va du mot le plus fréquent vers celui qui est moins fréquent et fournit « *a semantic analysis of a word* » [une analyse sémantique d'un mot] (Sinclair, 1991 : 115).

L'ordre des mots peut avoir de l'importance dans la collocation, notamment pour les expressions figées. Par exemple, en anglais il est possible de trouver « *kith and kin* » ou « *kin* » seul, mais il serait tout à fait inhabituel de trouver « *kith* » sans « *kin* », sauf dans une discussion de ce type, où le mot est cité, mais pas utilisé, ou alors dans un jeu de mots, comme dans cette publicité évoquée par Michael Stubbs : « *no more expensive to call your kith in Sydney than your kin in Southampton* ». Dans cet exemple, « *kith* » se trouve en position G5 par rapport à « *kin* » et serait donc récupérable par un programme de collocation avec un empan de G5, D4, tel celui préconisé par Sinclair en 2004.

La linguistique de corpus s'intéresse en priorité à ce qui est fréquent, plutôt qu'à l'exemple rare ou même unique, le cas du mot qui ne se retrouve qu'une fois, ou *hapax legomenon*, car la force

---

[52] Une discussion à propos de ce phénomène a été lancée par Mike Scott (le créateur de Wordsmith Tools) sur la liste de diffusion *Corpora,* le 13 septembre 2011, suite à un article de James Pennebaker dans *New Scientist* qui cite un corpus récent où le pronom « *I* » est le mot le plus fréquent (édition du 3 septembre 2011, p. 45). Il s'agit probablement d'un corpus oral d'un type spécifique.

[53] Données du Corpus français de l'université de Leipzig, consulté le 8 mars 2012.<http://wortschatz.uni-leipzig.de/ws_fra/index.php>.

[54] Liste lemmatisée des mots fréquents en français, par Etienne Brunet. Consulté le 8 mars 2012. <http://eduscol.education.fr/cid47916/liste-des-mots-classee-par-frequence-decroissante.html>.

du corpus est de présenter de nombreux exemples du même phénomène. Pourtant, les hapax restent toujours les plus nombreux en tant que classe, et les éliminer complètement de l'analyse pourrait conduire à une perte d'information. Pour décrire exhaustivement l'usage d'un mot peu fréquent, il est parfois utile d'aller plus loin que l'étude des collocats et de regarder plus en détail les concordances, afin de déceler des regroupements potentiels parmi les hapax.

2.3. Les concordances et comment les interpréter

Une concordance est une série d'exemples centrés sur un mot sélectionné, le nœud (« *node* »). En général, le concordancier est un programme informatique qui permet de spécifier l'étendue de texte à présenter de chaque côté du nœud. Dans le livre *Reading Concordances*, paru en 2003, Sinclair proposait un parcours d'apprentissage pour faciliter l'analyse des concordances. Sur le site d'accompagnement de son livre,[55] le programme informatique qu'il conseillait, l'outil *ConcApp,* créé par Chris Greaves,[56] avait été choisi en premier lieu pour sa simplicité d'utilisation. La linguistique de corpus dans la tradition contextualiste britannique nécessite la consultation des lignes de concordances, car il s'agit de déceler des régularités ou « *patterns* ». D'autres formes de linguistique assistée par ordinateur utilisent beaucoup plus de statistiques, parfois au détriment d'une véritable analyse linguistique ; il est assez révélateur que le type de programme employé, *R* ou *PAST* par exemple, peut s'appliquer à bien d'autres objets que des données linguistiques. Le titre même du livre de Sinclair, *Reading Concordances*, incite à considérer la concordance comme une forme de texte, qu'il convient d'observer et de lire en détail, afin de pouvoir analyser les faits de langue présents en nombre.

Un concordancier simple permet de trier les lignes de concordance, par le contexte gauche ou droite, et peut également mettre en valeur non seulement le nœud, qui est généralement présenté en position centrale, mais aussi les collocats, par un affichage en gras, ou souligné, ou en couleur. Ce tri des concordances permet de trouver des ressemblances parmi les collocats et ainsi de regrouper les hapax par affinité ou par similitude. Cependant, il est nécessaire d'avoir quelques connaissances dans la langue du corpus pour pouvoir être capable d'apercevoir les regroupements sémantiques possibles. Ceci introduit forcément une certaine variabilité dans l'analyse, d'où l'importance de bien préciser la démarche employée, afin de rendre l'étude reproductible, d'un chercheur à l'autre, d'un outil à l'autre et également d'un corpus à un autre.

[55] L'adresse du site internet d'accompagnement au livre est <http://www.twc.it/rc/readings.htm>.
[56] Le logiciel (devenu payant) est téléchargeable sur le site <http://www.edict.biz/pub/concapp/>.

C'est au prix du respect de ces règles de base que la linguistique de corpus peut être considérée comme une véritable science du langage.

2.4. L'expression « *silly ass* » : un exemple de collocation ?

L'une des expressions utilisées par J. R. Firth pour exemplifier la collocation est « *silly ass* ». Il voulait ainsi démontrer, par le biais d'une expression typique de son époque, que l'une des utilisations du mot « *ass* » [âne] en anglais parlé contemporain sert à indiquer la bêtise humaine, souvent en référence à un homme jeune[57]. Pour cette raison, lorsque « *ass* » [âne] se trouve en collocation avec « *silly* » [idiot], l'expression s'applique à un être humain et ne fait plus référence à l'animal.

En fait, sans le dire, Firth va au-delà de la simple collocation, car trois des quatre exemples qu'il fournit ne contiennent que le mot « *ass* » (Firth, 1957 [1951] : 195). La collocabilité avec « *silly* » peut donc être implicite, devenant ainsi un exemple de prosodie sémantique, expression employée pour la première fois par William E. (Bill) Louw en 1993, mais choisie par lui et J. McH. Sinclair dès 1988, en hommage à J. R. Firth.

La prosodie sémantique est le terme employé pour décrire le phénomène de « mauvaises fréquentations » : un mot ou expression neutre, qui se retrouve constamment dans des contextes négatifs, deviendrait négatif ou « coupable par association ». Par exemple, le verbe « causer », qui pourrait paraître plus neutre que « provoquer », par exemple, se retrouve souvent dans des contextes négatifs et se colore imperceptiblement de négativité. La définition dans le *Trésor de la langue française informatisé* (*TLFi*)[58] tient compte de cette négativité par association : « CAUSER, verbe transitif : Être à l'origine de, avoir pour effet quelque chose (l'effet désigne généralement un dommage). » Il existe aussi des mots ayant une prosodie sémantique positive, mais ils sont bien moins fréquents. Un exemple pourrait être le verbe « construire », qui se trouve souvent dans des contextes offrant la vision d'un futur agréable, avec des collocats tels que « projet », « avenir » et « meilleur ». Son quasi-synonyme « bâtir » se trouve dans des contextes plus concrets, à une exception près : on bâtit des châteaux en Espagne. Cette dernière expression indique surtout

[57] Le personnage comique créé par Pelham Grenville (P. G.) Wodehouse, l'aristocrate plutôt falot, Bertie Wooster, a été décrit en ces termes par A. N. Wilson (2008), Consulté le 28 juin 2012 : <http://www.telegraph.co.uk/comment/columnists/anwilson/3555169/Was-Bertie-Wooster-a-silly-ass-or-a-wise-man.html>.

[58] <http://atilf.atilf.fr/tlf.htm>. Dictionnaire bâti sur une base de données spécifique, surtout des textes littéraires, écrits en français, datant des XIX^e^ et XX^e^ siècles (70 millions de mots) ainsi que des documents scientifiques ou techniques (20 millions de mots).

l'absurdité des projets ainsi conçus, confirmant par défaut la prosodie sémantique favorable du verbe « construire ».

Dans un texte de 1996, Sinclair a développé la notion d'une « étendue de sens » (*extended unit of meaning*), qui comprend la collocation, la colligation, la préférence sémantique et la prosodie sémantique. La partie la plus concrète lors de l'analyse est la collocation, car il s'agit simplement et mécaniquement de la fréquence de cooccurrence de deux mots, selon un empan délimité (en principe, cinq mots à gauche et quatre à droite). La colligation est une relation avec une catégorie grammaticale, donc un élément plus abstrait, une préposition, par exemple. La préférence sémantique est la sélection d'un champ sémantique, tel le fait de voir ou d'exprimer quelque chose. Enfin, l'élément le plus abstrait, révélé surtout par l'examen des concordances, est la prosodie sémantique, que Sinclair définit ainsi :

> *A semantic prosody (Louw [1993]) is attitudinal, and on the pragmatic side of the semantics/pragmatics continuum. It is thus capable of a wide range of realisation, because in pragmatic expressions the normal semantic values of the words are not necessarily relevant. But once noticed among the variety of expression, it is immediately clear that the semantic prosody has a leading role to play in the integration of an item with its surroundings. It expresses something close to the "function" of the item – it shows how the rest of the item is to be interpreted functionally. Without it, the string of words just "means" – it is not put to use in a viable communication*[59] (Sinclair, 1996 : 87-88).

Sinclair prétend que la prosodie sémantique joue un rôle clé dans la communication, et qu'elle serait l'élément à partir duquel les différents choix d'expression découleraient. Si tel est le cas, il convient de préciser que l'expression choisie pour servir d'exemple de « *modern colloquial English* » par Firth en 1951 (1957 [1951] : 194) paraît aujourd'hui quelque peu démodée (Taylor, 2012 : 110[60]), indiquant que la prosodie sémantique, comme la langue, évolue avec le temps. De tels changements peuvent être révélés par des corpus diachroniques, où la date du texte est indiquée.

[59] Une prosodie sémantique (Louw [1993]) révèle une attitude et se place vers l'axe pragmatique du continuum entre sémantique et pragmatique. Elle est donc apte à se réaliser de nombreuses manières, car dans les expressions pragmatiques, la valeur sémantique normale des mots n'a pas forcément cours. Mais une fois décelée parmi les expressions utilisées, il est clair que la prosodie sémantique joue un rôle clé dans l'intégration d'un élément à son environnement. Elle exprime en quelque sorte la « fonction » de l'élément – elle indique comment interpréter la totalité de l'élément. Sans elle, la chaine de mots a du sens, mais ne sert pas véritablement à communiquer.

[60] L'analyse présentée par John R. Taylor prend pour point de départ un autre texte de Firth où le même exemple est utilisé.

Dans le *British National Corpus* (*BNC*)[61], conçu pour représenter l'anglais britannique de la deuxième moitié du vingtième siècle, le premier collocat de « *silly* » est « *thing* » et le mot « *ass* » n'arrive qu'en 28e place, avec seulement sept occurrences, présentées dans le tableau suivant (Table 1).

1 FNT W_fict_prose "I want you to tell me I'm a **silly ass** for letting it happen." "Et voila.
2 FNT W_fict_prose "Et voila. You're a **silly ass** for letting it happen." He paused.
3 FNU W_fict_prose I might take my father's advice and go into business." "**Silly ass**," she said.
4 HH9 W_fict_prose don't bother," she said. "Don't be a **silly ass**. Get in. We're holding up the
5 J0X W_fict_poetry how messy if We get crushy with love and play the **silly ass**! Behind the
6 GTE W_biography creating and recreating the role of the **silly ass** forever working his way
7 HRF W_biography brilliantly played by Michael Crawford -- is the prototype **silly ass**.

Table 1 « *silly ass* » dans le *BNC*

Chaque fois, il s'agit du phrasème « *silly ass* ». Un examen détaillé des sources des occurrences a été mené, d'abord dans les métadonnées fournies avec le *BNC*, puis par une recherche ciblée pour chaque texte. Cette méthodologie est conforme à celle préconisée par Sinclair (Ghadessy *et al.*, 2001) pour un petit nombre de données, c'est-à-dire l'analyse par intervention humaine précoce (*EHI* ou *Early Human Intervention*).

Cette manière de procéder permet d'établir que, pour les quatre extraits de romans (les occurrences 1 à 4), l'auteur est toujours une femme, l'expression est toujours utilisée dans un dialogue entre deux personnages et que, même si les dates de publication vont de 1989 à 1992, le cadre temporel de chaque roman se situe bien avant, en 1934 pour le texte de Lisa Appignanesi (occurrences 1 et 2), en 1950 pour celui de Beryl Bainbridge (occurrence 3) et dans les années 1960 pour le texte de Nina Bawden (occurrence 4).

Les deux occurrences dans le texte d'Appignanesi se suivent, car le deuxième personnage reprend exactement l'expression employée par le premier, en changeant seulement de pronom personnel. Le personnage qui parle en premier est une jeune femme qui se prénomme Violette, qui semble être d'origine française, et qui emploie l'expression pour parler d'elle-même. C'est inhabituel d'employer cette expression pour parler d'une femme ; c'est d'ailleurs le seul cas de ce type dans le corpus. Cette inversion de ce qui est attendu témoigne d'une certaine altérité, révélant l'origine linguistique du personnage, voire peut-être même de l'auteur, car Lisa Appignanesi est d'origine polonaise et a grandi en France et à Montréal.

[61] Avec l'interface de Mark Davies, *BYU-BNC* <http://corpus2.byu.edu/bnc/>. Consulté le 28 juin 2012.

Dans l'extrait du poème (occurrence 5), Herbert Lomas utilise le terme pour décrire le jeu de l'amour, qui peut abêtir, même s'il ne rend pas complètement fou. Le terme s'applique donc au couple. Il faut noter aussi l'enchaînement d'adjectifs qui se terminent en « *y* » : « *messy* » [en désordre], « *crushy* » [en pincer pour quelqu'un] et « *silly* » [sot], rappelant l'aspect phonésthétique attribué notamment aux voyelles par Firth.

Les deux autres occurrences, du domaine de la biographie, décrivent deux acteurs connus surtout pour leur interprétation de personnages falots, au théâtre et au cinéma. Là aussi, le cadre temporel a de l'importance : si Ralph Lynn (occurrence 6) était surtout connu pendant les années 1930, le film[62] dans lequel Michael Crawford joue le rôle du Lieutenant Ernest Goodbody (occurrence 7) a été tourné en 1967, mais l'action se situe au cours de la deuxième guerre mondiale, et son personnage est déjà anachronique et décalé par rapport à cette époque.

La liste des cinq premiers collocats du mot « *ass* » dans le *BNC* (« *your ; law ; kick ; pain ; ox* ») permet de trier parmi les différents sens potentiels (« *meaning potentials* ») de ce mot[63]. Tout d'abord, lorsque « *ass* » est en collocation avec le cinquième collocat « *ox* », il s'agit toujours du sens animalier, le premier sens du mot, ici dans le contexte de la crèche de Noël, avec l'âne et le bœuf. Les seize occurrences où « *ass* » est en collocation avec son deuxième collocat « *law* » sont présentées dans le tableau suivant (Table 2).

[62] Il s'agit du film de Richard Lester « *How I won the war* », avec Michael Crawford et John Lennon.
[63] Le sixième collocat est « *silly* », toujours dans l'expression « *silly ass* » (Table 1).

1	list of shops, not goods, as at present, at present the **law** is an **ass**, it's based on goods, not the
2	was it that said, The **Law** 's an **Ass**. Mm. one thing, times No. It was the Beadle in Oliver
3	Colonel Stanley interrupted testily. "I understand. The **law**'s an **ass**. Dickens wrote that. He
4	she silenced me. "I know, I know. The **law**'s an **ass**. Never mind, what it comes to is that
5	"The **law** is in love with the past." "The **law** is an **ass**," said Willoughby. "It's foolish not to
6	refereeing has to be consistent. Otherwise the letter of the **law** is an **ass**. Nottingham Forest:
7	judged "irrational" by the High Court (who said the **law** was an **ass**?). But I can't help feeling
8	choice and Goram when all else fails which has so far made an **ass** of the **law** of averages.
9	feel safer, the Attorney-General told TODAY. "The **law** did sometimes make an **ass** of itself"
10	faced their full quota of overs. But because the **law** **as**[64] an **ass**, overs were arbitrarily taken
11	the **law** can not act to correct it. The phrase that the **law** is an **ass**, because it falls to cover the
12	the lawyers' bluff, they might find that the **law** is not quite the **ass** it sometimes appears.
13	serve as food for the argument that in so doing the **law** becomes an **ass**. What possible sense
14	impression being given is that while the **law** may be an **ass**, the profession of psychology is
15	If the **law** thought she could be so, then the **law** must be an **ass**. So the family overstayed its
16	(A.P.) Herbert, who delighted in finding cases which proved the **law** an **ass**, wrote a highly

Table 2 Occurrences de « *law* » et « *ass* » dans le *British National Corpus*[65]

Dans ce tableau, les occurrences 2 et 3 indiquent que l'origine de l'expression est le roman de Charles Dickens, *Oliver Twist*, mais il est difficile de savoir si la stupidité dont il est question s'applique à l'animal ou à l'homme. Une analyse de ce roman avec *WConcord* confirme que c'est bien le personnage de Bumble, le bedeau, qui prononce cette phrase, mais avec un défaut de prononciation, l'absence de la forme de liaison de l'article indéfini « *an* » (forme utilisée lorsque le mot qui le suit commence par un son vocalique), ainsi : « *the law is a ass--a idiot.* » [la loi est un nâne – un nidiot]. Comme Bumble utilise cette même expression, avec la même erreur de prononciation, pour décrire tout chat qui n'apprécierait pas Madame. Corney, il semble logique de supposer que c'est l'animal métaphorique qui est représenté par la phrase, avec sa stupidité d'âne bâté, trop lourdement chargé pour pouvoir avancer rapidement. Ce collocat témoignerait donc de la transition entre l'âne réel, peu représenté dans le corpus, qui devient l'animal mythique ou emblématique avec la crèche de Noël, puis l'âne des fables, qui incarne la stupidité.

Les trois autres collocats, « *your* », « *kick* » et « *pain* », sont associés au sens corporel humain de « fesses ». Le premier collocat (37 + 3 occurrences) est toujours lié au sens corporel lorsque l'expression est « *your ass* ». En revanche, pour les trois occurrences qui n'ont pas ce sens, les deux mots sont séparés, et dans deux de ces cas, le mot « *ass* » arrive avant le mot « *your* »,

[64] Ligne 10 – Il existe une faute de frappe ou erreur dans le texte d'origine, soulignée ici et écrite en gras « ***as*** » à la place de « *is* ».
[65] Accessible par l'interface *BYU-BNC* < http://corpus2.byu.edu/bnc/>. Consulté le 28 juin 2012.

indiquant qu'il s'agit de bruit, ou de faux positifs, où les mots ne sont pas à analyser en collocation. Le troisième cas est plus intéressant, car il s'agit d'une maladresse de traduction et elle est citée comme contre-exemple : « *Visitors to Thailand would do well to avoid the donkey rides, where a notice reads:"Would you like to ride on* ***your*** *own* ***ass****?"* » Une telle erreur de collocation, s'il ne s'agit pas d'ironie ou d'absence de sincérité (Louw, 1993), témoigne souvent d'une maîtrise imparfaite d'une langue étrangère (Louw et Chateau, 2010).

Conclusion

Après avoir retracé l'histoire de la linguistique de corpus dans la deuxième moitié du vingtième siècle, cette étude a également présenté quelques notions fondamentales pour la pratique de cette science dans la tradition contextualiste britannique, dans la lignée de J. R. Firth et de John McHardy Sinclair. Quelques termes de base ont été explicités et une étude de cas centrée sur l'expression « *silly ass* » a permis de démontrer la manière de procéder et la démarche à suivre.

Bibliographie

Aarts, Jan et Willem Meijs. (éds.). 1984. *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research.* Amsterdam: Rodopi.

Banks, David. 1999. « Becoming part of the network: French scientists and the use of English at conferences », *ASp* 23-26. <http://asp.revues.org/2442>. Consulté le 13 octobre 2012.

Baroni, Marco and Bernardini, Silvia (éds.) 2006. *Wacky! Working papers on the Web as Corpus.* Bologna: GEDIT.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta. 2009. « The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora ». *Language Resources and Evaluation* 43/3. 209-226.

BNC Consortium. 1995.*The British National Corpus*, version 1 (BNC 1.0). Distribué par *Oxford University Computing Services* pour le *BNC Consortium*. Consulté le 25 juillet 2012. Accessible <http://www.natcorp.ox.ac.uk/>.

Bontrager, Terry. 1991. « The Development of Word Frequency Lists prior to the 1944 Thorndike Lorge List ». *Reading Psychology* 12/2. 91–116.

Chamberlin, Rollin T. 1928. « Some of the objections to Wegener's theory ». In van Waterschoot van der Gracht, Willem A. J. M. (éd.). *Theory of Continental Drift: A Symposium.* Tulsa, OK : American Association of Petroleum Geologists ; Londres : Thomas Murby & Co. 83–87.

Davies, Mark. 2004–. *BYU-BNC.* (Basé sur le *British National Corpus* compilé par Oxford University Press). Disponible <http://corpus.byu.edu/bnc/>. Consulté le 27 juillet 2012.

Davies, Mark. 2007–. *TIME Magazine Corpus: 100 million words, 1920s–2000s*. Disponible <http://corpus.byu.edu/time/>. Consulté le 27 juillet 2012.

Davies, Mark. 2008–. *The Corpus of Contemporary American English: 450 million words, 1990–present.* Disponible <http://corpus.byu.edu/coca/>. Consulté le 27 juillet 2012.

Davies, Mark. 2010–. *The Corpus of Historical American English*: *400 million words, 1810–2009*. Disponible <http://corpus.byu.edu/coha/>. Consulté le 27 juillet 2012.

Denison, David. 2007. « Playing tag with category boundaries ». *VARIENG e-Series* 1. Accessible < http://www.helsinki.fi/varieng/journal/volumes/01/denison/>. Consulté le 27 juillet 2012.

Eldridge, R.C. 1911. *Six thousand common English words, their comparative frequency and what can be done with them.* Buffalo : Clement Press.

Firth, John R. 1957 [1935]. « The Use and Distribution of Certain English Sounds ». *Papers in Linguistics 1934-1951*. Londres : Oxford University Press. 34-46.

Firth, John R. 1957 [1951]. « Modes of Meaning ». *Papers in Linguistics 1934-1951*. Londres : Oxford University Press. 190-215.

Francis, Gill, Susan Hunston et Elizabeth Manning. 1996. *Collins COBUILD Grammar Patterns 1: Verbs.* Birmingham : Harper Collins.

Francis, Timothy J. G. 1988. « European Collaboration in Marine Geoscience an Informal Discussion Paper ». *Underwater Technology* 14/1. Consulté le 25 juillet 2012. Accessible <http://www.sut.org.uk/journal/contents/journal_14.htm>.

Francis, W. Nelson et Henry Ku era. 1964. *Brown Corpus Manual: Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.* Accessible <http://icame.uib.no/brown/bcm.html>. Consulté le 21 août 2011.

Fries, Charles C. 1952. *The structure of English: an introduction to the construction of English sentences.* New York, NY : Harcourt Brace.

Fries, Peter H. 2010. « Charles C. Fries, linguistics and corpus linguistics ». *ICAME Journal* 34. 89–121.

Ghadessy, Mohsen, Alex Henry et Robert L. Roseberry. (éds.). 2001. Small Corpus Studies and ELT: Theory and Practice. Amsterdam et Philadelphie : John Benjamins.

Halliday, Michael A.K. 1961. « Categories of the theory of grammar ». *Word* 17. 241–92.

Halliday Michael A.K. et Christian M. Matthiessen. 2004 [1985]. *Introduction to Functional Grammar*. 3e édition. Londres : Edward Arnold.

Hanks, Patrick. 2007. « John Sinclair, (1933-2007) ». In Bogaards, Paul (dir.). *EURALEX Newsletter: Summer 2007. International Journal of Lexicography* 20/2. 209-215.

Johansson, Stig, Geoffrey N. Leech et Helen Goodluck. 1978. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, For Use with Digital Computers*. Consulté le 21 août, 2011. Accessible <http://khnt.hit.uib.no/icame/manuals/lob/index.htm>.

Kilgarriff, Adam et Gregory Grefenstette (éds.). 2003. « Introduction to the Special Issue on Web as Corpus ». *Computational Linguistics* 29/3. 333–347.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, et David Tugwell. 2004. « The Sketch Engine ». *Proceedings EURALEX 2004, Lorient, France*. 105–116.

Krishnamurthy, Ramesh (éd.). 2004. *English Collocation Studies: The OSTI Report*. Londres : Continuum.

Kuhn, Thomas S. 1970 [1962]. *The Structure of Scientific Revolutions*, 2e édition, Chicago, MA : Chicago University Press.

Leech, Geoffrey. 1992. « Corpora and theories of linguistic performance ». In Svartvik, Jan (éds.). *Directions in Corpus Linguistics, Proceedings of Nobel Symposium, 4-8 August 1991*. Berlin et New York, NY : Mouton de Gruyter. 105–122.

Leech, Geoffrey et Stig Johansson. 2009. « The coming of ICAME ». *ICAME Journal* 33. 5–20.

Leech, Geoffrey. 2011. « The modals ARE declining: Reply to Neil Millar's 2009 "Modal verbs in TIME: Frequency changes 1923–2006" ». *International Journal of Corpus Linguistics* 14/2. 191–220. *International Journal of Corpus Linguistics* 16/4. 547v564.

Léon, Jacqueline. 2005. « Claimed and Unclaimed Sources of Corpus Linguistics ». *Henry Sweet Society Bulletin* 44. 36–50.

Louw, William E. (Bill) 1993. « Irony in the text or insincerity in the writer? – the diagnostic potential of semantic prosodies ». In Baker, Mona, Gill Francis et Elena Tognini-Bonelli (dir.). 1993. *Text and Technology*, Amsterdam : John Benjamins. 157–176.

Louw, William E. (Bill) et Carmela Chateau. 2010. « Semantic prosody for the 21st Century: Are prosodies smoothed in academic contexts? A contextual prosodic theoretical perspective ». *JADT 2010 : 10e Journées d'Analyse statistique des Données Textuelles du 9 au 11 juin 2010*. Rome. Accessible <http://www.ledonline.it/ledonline/index.html?/ledonline/jadt-2010.html>. Consulté le 7 octobre, 2012.

Louw, William E. (Bill) 2000. « Contextual Prosodic Theory: Bringing Semantic Prosodies To Life ». Heffer, Chris et Helen Sauntson (dir.). 2000. In *Words in Context, A tribute to John Sinclair On his*

Retirement. CD-ROM. <http://www.revue-texto.net/docannexe/file/124/louw_prosodie.pdf>. Consulté le 22 août 2011.

Malinowski, Bronislaw. 1923. « The problem of meaning in primitive languages. » In Ogden Charles K. et I.A. Richards. 1923. *The Meaning of Meaning*. London : Routledge et Kegan Paul. 296–336.

Millar, Neil 2009. « Modal verbs in TIME: Frequency changes 1923–2006 ». *International Journal of Corpus Linguistics* 14/2. 191–220.

Nagasawa, Jiro. 1958. « A Study of English-Japanese Cognates ». *Language Learning* 8/1-2. 53–102.

NASA. Mars Science Laboratory Landing. Dossier de presse, juillet 2012. Accessible <http://www.nasa.gov/news/media/presskits/index.html>. Consulté le 23 juillet 2012.

Palmer, Harold E. 1955 [1938]. *A Grammar of English Words: One Thousand English Words and their Pronunciation, together with Information concerning the Several Meanings of Each Word, its Inflections and Derivatives, and the Collocations and Phrases into which it Enters*. Londres : Longmans, Green.

Paterson, Kenneth, Caroline Caygill, et Rebecca Sewell. 2011. *A Handbook of Spoken Grammar. Strategies for speaking natural English*. Londres : Delta Publishing.

Quirk, Randolph. 1960. « Towards a description of English Usage ». *Transactions of the Philological Society*. 40–61.

Sampson, Geoffrey et Diana McCarthy (éds.). 2005. *Corpus Linguistics: Readings in a Widening Discipline*. Londres : Continuum.

Sinclair, John McH. 1966. « Beginning the study of lexis ». In Bazell, Charles E., J. C. Catford, M. A. K. Halliday et R. H. Robins (dir.). 1966. *In Memory of J. R. Firth*. Londres : Longman. 410–430.340

Sinclair, John McH. (dir.). 1987. *Looking Up. An Account of the COBUILD Project in Lexical Computing*. Londres et Glasgow : Collins.

Sinclair, John McH. 1991. *Corpus, Concordance*, *Collocation*. Oxford : Oxford University Press.

Sinclair, John McH. 1996. « The search for units of meaning ». *Textus: English Studies in Italy* 9. 75–106.

Sinclair, John McH. 2003. *Reading Concordances*. Londres : Pearson Longman. Site <http://www.twc.it/rc/readings.htm>.

Smith, Richard C. 1999. *The Writings of Harold E. Palmer, An Overview*. Tokyo : Hon-No-Tomosha. <http://www2.warwick.ac.uk/fac/soc/al/research/collect/elt_archive/halloffame/palmer/archive/palmer_1924.pdf>. Consulté le 11 septembre 2011.

Svartvik, Jan et Randolph Quirk. (éds). 1980. *A Corpus of English Conversation*. Lund Studies in English, 56. Lund: Liber/Gleerups.

Taylor, John R. 2012. *The Mental Corpus: How language is represented in the mind.* Oxford : Oxford University Press.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam : John Benjamins.

Vine, F. J. et D. H. Matthews. 1963. « Magnetic Anomalies over Oceanic Ridges ». *Nature*. 199/4897. 947–949.

Williams, Geoffrey. 2010. « Many rooms with corpora ». *International Journal of Corpus Linguistics* 15/3. 400-408.

Williams, John. 1998. « *Grammatical constructions and social construction: a case study in the practice of linguistic 'science'* ». Dissertation de diplôme de Master en Sciences sociales de l'université de Birmingham, soutenue le 28 août 1998.

Zipf, George K. 1935. *The Psychobiology of Language: An introduction to dynamic philology*. Boston, MA : Houghton-Mifflin. Consulté le 20 août 2011. Accessible <http://babel.hathitrust.org/cgi/pt?id=mdp.39015008729983>.

Zipf, George K. 1945. « The meaning-frequency relationship of words ». *Journal of General Psychology* 33. 251–256.

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA : Addison-Wesley.