LA LINGUISTIQUE ET LE CORPUS : UNE AFFAIRE PRÉPOSITIONNELLE

Geoffrey WILLIAMS
Université de Bretagne Sud, Lorient

SOMMAIRE

Résumé : *La linguistique de corpus a été très largement développée comme discipline dans le monde anglo-saxon. Ce paradigme de recherche est sorti de la linguistique appliquée à partir de deux grandes traditions ; l'enseignement de l'anglais comme langue seconde et une approche contextualiste de la linguistique, approche associée à Firth. Dans cette communication, je montre comment les deux traditions se sont fusionnées avec le projet COBUILD. Je décris l'évolution de la discipline comme paradigme de recherche développé autour des corpus soigneusement constitués et utilisant une analyse inductive. Dans la conclusion, je plaide pour la reconnaissance de la linguistique de corpus autonome par opposition à la linguistique sur corpus qui implique d'autres disciplines telles que la sociolinguistique ou le TAL.*

1. Introduction

En anglais, la situation est simple, '*corpus linguistics'* est un mot composé formé de deux substantifs, dont l'un va limiter le champ de référence de l'autre. La linguistique est une discipline, le mot corpus décrit l'objet. Le reste est sujet à interprétation, la puissance de l'anglais est dans l'ambiguïté, une ambiguïté que nous n'essaierons pas de lever dans l'immédiat.

Pour le français, la situation est plus complexe puisque nous ne pouvons pas simplement juxtaposer deux mots, il faut les lier avec une préposition, et le choix d'une préposition implique une interprétation. Faut-il mettre *de* ou *des*, ou peut-être *sur* ?

Avant même de choisir la préposition, nous rencontrons une difficulté supplémentaire : le mot '*linguistics'* en anglais semble être un pluriel, mais comme '*physics'* ou '*mathematics'*, il est en réalité invariable. Par conséquent, 'de' signifiera la présence d'une discipline unique, 'des', que plusieurs disciplines – au lieu de plusieurs approches de la même discipline – sont en jeu. 'Sur' est une interprétation supplémentaire impliquant que d'autre domaines de la linguistique peuvent utiliser les corpus sans faire de la linguistique de corpus *per se*, ce qui soulève la question de la nature des corpus.

Le but de cet article est d'essayer de démêler les différentes interprétations de '*corpus linguistics*' en décrivant l'origine anglo-saxonne de la discipline, le contexte de recherche de la discipline de son origine à nos jours. En comparant les différentes interprétations françaises du terme nous essaierons non d'imposer une définition, mais de clarifier la situation entre les différentes approches de cette discipline.

2. Les origines de la linguistique de corpus dans la tradition anglo-saxonne

Il est toujours trop facile d'essayer de trouver un inventeur, comme s'il suffisait de crier *eurêka* et de trouver des merveilles. Cette tendance est exaspérée par des velléités patriotiques : la théorie

de Darwin a provoqué un intérêt pour les écrits de Lamarck, décriés auparavant, la parenté de la photographie est disputée entre un Daguerre et un Fox Talbot, et ainsi de suite.
En ce qui concerne la linguistique de corpus, la question de l'antériorité des corpus se dispute entre plusieurs corpus électroniques, FRANTEXT, Brown, OSTI… En réalité, chaque ensemble textuel a été créé en reconnaissant des possibilités offertes par l'informatique naissante afin de résoudre des problématiques différentes. Il est donc inutile de chercher l'antériorité d'un tel ou un tel, d'autant plus que, pendant cette période antérieure au courrier électronique, les chercheurs travaillaient en relative isolation. Le plus important sera de voir pourquoi et comment des tendances actuelles ont évolué afin que des chercheurs de nos jours puissent échanger des informations en comprenant l'autre.
Dans la tradition anglo-saxonne de la linguistique de corpus, la lexicographie, l'enseignement et les corpus sont intimement liés. La tendance contextualiste est le fruit de l'interaction entre les trois éléments de base.

2.1. La situation avant 1945

2.1.1. La lexicographie

On peut ainsi dater le développement de la linguistique de corpus à 1755 avec le dictionnaire de Johnson, le premier dictionnaire basé sur un 'corpus' sous la forme de fiches de travail accompagnées de citations. Une telle affirmation est peut-être un peu osée, mais pas totalement infondée puisqu'avec Johnson débute une tradition lexicographique plus normative que prescriptive mais basée sur des textes authentiques, bien que limitée à des textes 'nobles' de la littérature. La tradition lexicographique instaurée par Johnson est à la base du *Oxford English Dictionary*. Plus récemment, la tradition lexicographique d'Oxford a donné naissance à une autre forme de dictionnaire, le dictionnaire pour apprenant, avec le *Oxford Advanced Learner's Dictionary*, issu du *Learner's Dictionary of Current English* de Hornby publié en 1948 (Cowie). Ces dictionnaires pour apprenants sont toujours basés sur des fiches, mais avec des exemples tirés de la langue générale. Le ton a changé en 1987 avec la publication du *COBUILD Advanced Learner's English Dictionary*, basé sur un grand corpus de référence. Dorénavant tous les dictionnaires pour apprenants seront basés sur corpus, et les corpus seront de plus en plus utilisés pour l'élaboration de dictionnaires monolingue et bilingue des éditeurs britanniques, et maintenant dans beaucoup d'autres pays. Ce qui a poussé à ces deux révolutions, celle de Hornby, puis celle de l'équipe de COBUILD, est l'enseignement de l'anglais comme langue seconde.

2.1.2. L'enseignement de l'anglais comme langue seconde

Grâce à l'Empire Britannique, l'anglais était devenu dans la période suivant la première guerre mondiale une langue dominante dans les affaires. Il fallait par conséquent que les gens apprennent l'anglais (pas nécessairement celui de la langue de Shakespeare) d'une manière plus pragmatique pour le travail. Les bases pour un enseignement de la langue fondé sur une linguistique appliquée avaient déjà été jetées avec la publication de Sweet *Practical study of Languages* en 1899 (Howatt 1984), développé à partir d'un article publié en 1884. Dans l'approche de Sweet, le lexique et la phraséologie étaient centraux, mais il fallait que le lexique soit structuré et que les phrases soient un lien entre le texte et la grammaire, autrement dit, un certain contexte était nécessaire pour apprendre. Les phrases ne seront pas inventées, mais authentiques, l'autre credo du contextualisme.
L'enseignement des langues s'est beaucoup développé dans la période avant la première guerre mondiale, mais en ce qui concerne la linguistique de corpus, la période la plus importante date de l'entre-deux-guerres avec les travaux de Palmer au Japon. Cette période a vu un intérêt intense pour des vocabulaires essentiels pour apprenants, mais également les premiers travaux sur la collocation en anglais.
Pendant ses années au Japon, Palmer a publié extensivement sur la théorie et la pratique de l'enseignement de l'anglais comme langue seconde (Howatt *op. cit.*). Palmer s'est beaucoup investi dans l'étude du lexique, dans le but de créer un vocabulaire contrôlé pour l'apprentissage, deux rapports ayant été publiés sur ce thème. Il a aussi collaboré avec West, l'auteur du *General Service List*, liste de mots à la base de nombreuses méthodes d'apprentissage. C'est précisément cet intérêt pour un vocabulaire restreint au service des apprenants qui a donné naissance à un dictionnaire de langue générale pour apprenants, le *Learner's Dictionary of Current English* de Hornby.

L'autre aspect des travaux sur le vocabulaire de Palmer est son rapport sur les collocations, *Second Interim Report on English Collocations* (Palmer 1933). L'étude des collocations était une suite logique à des rapports sur le vocabulaire montrant qu'au delà des mots simples, il y avait ce que Palmer a appelé des « *comings-together-of-words* », des rassemblements de mots (*ibid*. p.1). Après une discussion des classifications possibles, Palmer décide de les appeler 'collocations', réutilisant un terme vague ayant déjà été employé par Sweet. D'après Palmer, il sera nécessaire de définir ce que l'apprenant doit apprendre comme combinaisons ; les combinaisons figées et sémi-figées. La suite est une classification des collocations par parties de discours. Ces collocations sont trouvées dans des textes authentiques, mais par le biais de l'intuition du linguiste. Le rapport est souvent cité, mais n'a jamais été largement publié. La tradition collocationnelle de Palmer a beaucoup influencé la phraséologie, tradition qui a cependant largement ignoré les possibilités offertes par les corpus jusqu'assez récemment. L'analyse des collocations en corpus est issue d'une autre tradition de recherche, le contextualisme de Firth.

2.1.3. Firth et le contextualisme

Firth est souvent vu comme le père de la collocation, même si ses écrits sont postérieurs à ceux de Palmer. Il est probable que nous ayons ici une des coïncidences historiques de découvertes quasi-simultanées. Il est possible que les travaux de Firth soient aussi plus largement lus en raison de sa position de Professeur de linguistique à Londres et de la large diffusion de ses écrits par ses étudiants. Les écrits de Firth sont beaucoup plus énigmatiques que ceux de Palmer, sans la démonstration pratique que nous donne le *Interim Report*. La phrase célèbre de Firth « you shall know a word from the company it keeps » montre que le point de vue est différent de celui de Palmer. Pour celui-ci, il s'agissait d'unités polylexicales à découvrir, à mettre dans un dictionnaire et à transmettre aux apprenants, mais « the company words keep » est une approche autre, où la nécessité d'avoir des ensembles bien formés est moins importante que la notion d'associativité. La différence se trouve dans une approche textuelle, par opposition à une approche lexicographique, de la collocation. La textualité est centrale aux thèses de Firth qui ont développé les notions de contexte de culture et contexte de situation de Malinowski.

Anthropologue de renom, Malinowski reste très connu pour ses travaux sur les habitants des îles Trobriand. Il a reconnu très tôt l'importance de prendre en compte les aspects culturels dans la compréhension de la langue, le sens ne pouvant pas être évalué en dehors du contexte de situation.

> *Without some imperative stimulus of the moment, there can be no spoken statement. In each case, therefore, utterance and situation are bound up inextricably with each other and the context of situation is indispensable for the understanding of the words* (Malinoswki 1924 : 307).

Ces deux notions de base ont été reprises et développées par Firth, qui a travaillé également à l'Université de Londres, pour élaborer une théorie linguistique ; le contextualisme. La linguistique de Firth était un rejet de l'approche mentaliste. Selon lui (1935 : 19)

> *I do not therefore follow Ogden and Richards in regarding meaning as relations in a hidden mental process, but chiefly as situational relations in a context of situation and in that kind of language which disturbs the air and other people's ears, as modes of behaviour in relation to the other elements in the context of situation*

Firth était néanmoins un homme de son époque, ses sources sont authentiques, mais largement littéraires. Firth est resté aussi un théoricien du langage, le contextualisme ayant surtout été développé par ses étudiants, notamment Halliday et Sinclair.

2.2. Le contextualisme d'après guerre

Dans le développement du contextualisme, deux disciples de Firth sont à noter : Halliday et Sinclair. Halliday est à l'origine de la grammaire systémique et fonctionnelle, une grammaire descriptive très employée dans la linguistique de corpus contextualiste puisque complète, mais neutre. Si Halliday a surtout développé l'aspect grammatical, c'est Sinclair qui sera à l'origine de la partie lexicale et donc 'l'inventeur' de l'analyse de corpus contextualiste.

Une publication majeure dans le développement du contextualisme est parue en 1966, « In Memory of J. R. Firth » (Bazell *et al.*). Cette collection d'articles est à la fois une rétrospective sur les travaux de Firth, mort en 1960, et un programme pour le futur. Ainsi, des linguistes comme Jakobson et Lyons vont commenter l'apport de Firth, tandis que les articles de Halliday « Lexis as a linguistic level » et Sinclair « Beginning the study of lexis » annoncent les recherches qui vont mener à la grammaire systémique et fonctionnelle et à la linguistique de corpus contextualiste.

2.2.1. Halliday et la grammaire systémique et fonctionnelle

La théorie de Halliday a été annoncée dans son article de 1961 sur la catégorisation dans la grammaire. C'est une grammaire descriptive, textuelle et fermement basée sur le contexte. Ainsi, dans l'introduction de son œuvre majeure « An Introduction to Functional Grammar » (1994), il déclare que

> Just as each text has its environment, the 'context of situation' in Malinowski's terms, so the overall language system has its environment, Malinowski's 'context of culture'. The context of culture determines the nature of the code. As a language is manifested through its texts, a culture is manifested through its situations; so by attending to text-in-situation a child construes the code, and by using the code to interpret text he construes the culture. (1985 : xxxi)

Dans sa grammaire, l'analyse est essentiellement descendante, du texte à la phrase, de la phrase aux mots. Cependant, dans une théorie de lexico-grammaire, il y a forcement interaction entre la grammaire et la lexis. Ainsi il insiste :

> A text is a semantic unit, not a grammatical one. But meanings are realized through wordings; and without a theory of wordings -- that is, a grammar -- there is no way of making one's interpretation of the meaning of a text. (*ibid.* : xvii)

Dans son texte de 1966 annonçant le programme de recherche lexicale dans la grammaire, Halliday insiste sur le fait que la lexis est partie intégrante de la grammaire et constitue la partie la plus délicate, au sens de la plus fine, 'one-member classes' (1966 :149). Le fait que la lexis entre dans une classe unique ne veut pas dire que les mots sont relégués à une simple liste en marge de la grammaire. La grammaire de Halliday est systémique et multi-niveaux, il y a forcément une interaction entre tous les constituants qui forment le texte, et entre le texte et son environnement. Ainsi, la cohésion textuelle tient un rôle essentiel dans la grammaire (Halliday & Hasan 1971). Une partie de la notion de cohésion est basée sur la collocation, l'interaction entre mots. Tandis que Halliday utilise l'interaction collocationnelle dans le texte, Hoey l'a amenée plus loin dans le corpus (Hoey 1991, 2005).

En tant que grammaire descriptive, la grammaire systémique et fonctionnelle occupe une place de choix dans l'étude des corpus. Cependant, c'est largement une grammaire textuelle, l'aspect lexical ayant été traité par l'autre disciple de Firth, John Sinclair.

2.2.2. Sinclair et le rapport OSTI

Dans le titre même de son article en mémoire de Firth (1966), Sinclair a noté que nous n'étions qu'au début d'une étude contextualiste du lexique. Il a rapidement trouvé que l'outil informatique pouvait offrir un moyen d'aller plus loin. Ainsi il était amené à créer un corpus électronique. Le résultat de ces études sur corpus était un rapport publié en 1970, rapport qui a jeté les bases de la linguistique de corpus contextualiste, bien que peu diffusé à l'époque et publié seulement très récemment (Sinclair *et al.* 1970, 2004).

Le débat sur qui a créé le premier corpus électronique est largement stérile. Le mouvement vers une analyse des textes avec des outils informatiques était inévitable : il était dans l'air du temps, mais avec des objectifs différents. Comme l'a montré Léon (2005), l'arrivée de la théorie générative n'a eu aucun effet sur le développement de la linguistique de corpus contextualiste, qui a continué à évoluer dans le contexte de la linguistique appliquée.

Les premiers corpus ont été construits pour des raisons très différentes ; le TLF était largement littéraire, le Brown était également un corpus d'écrit, mais basé sur des échantillons et le *Survey of English Usage*, créé pour des recherches sur la syntaxe était largement inspiré par la tradition Firthienne mais n'a été numérisé que très tardivement. L'objectif du corpus OSTI était par contre d'explorer la lexis dans le paradigme contextualiste en faisant un corpus initialement basé sur l'oral. Le projet a démarré en 1963 (Teubert 2004). L'assemblage du corpus a commencé à l'Université d'Edimbourg et a été complété à l'Université de Birmingham. À l'époque, le fait d'avoir un ordinateur dédié à un projet linguistique était quelque chose d'extraordinaire dans un monde où uniquement les élites des sciences dures y avaient accès (Sinclair, communication personnelle).

Le rapport OSTI, officiellement *The Report to the Office for Scientific and Technical Information (OSTI) on the Lexis Research project for the period January 1967 – September 1969* était le résultat des travaux sur le corpus construit à Edimbourg et exploité à Birmingham. Outre la problématique de la création d'un corpus, le rapport est un véritable programme de recherche contextualiste, où les collocations s'avèrent centrales à l'approche. La notion de collocation significative a déjà été introduite par Sinclair (1966), mais ici la notion est explorée en relation avec des données issues du corpus. C'est dans ce rapport que les termes clés, comme *empan* et

fenêtre, sont introduits et justifiés. Déjà la notion du principe d'idiome commence à apparaître.
Bizarrement, le rapport OSTI a été oublié par la suite, de la même manière que Palmer (1933) est souvent cité, mais n'est pas disponible. Néanmoins, l'approche élaborée dans le rapport OSTI a servi de base pour un projet encore plus ambitieux, le projet COBUILD.

2.2.3. L'école de Birmingham à COBUILD

COBUILD était une collaboration entre l'Université de Birmingham et les dictionnaires Collins. L'objectif était de construire un grand corpus de référence pour l'anglais et de l'utiliser pour la création d'un dictionnaire pour apprenants basé uniquement sur une analyse de corpus. C'est effectivement avec le projet COBUILD que nous trouvons unifiées les deux traditions d'étude de la collocation : la tradition de Palmer a été fructifiée dans *l'Oxford Advanced Learner's Dictionary,* et la tradition contextualiste s'est développée séparément. Avec le COBUILD, nous avons enfin un dictionnaire où la collocation trouve sa juste place, mais au lieu d'être basés sur l'intuition d'un lexicographe, les collocations et les sens doivent être justifiés par les données du corpus. Dans l'école de Birmingham, le rêve de Firth de voir la linguistique et la lexicographie unifiées a également été réalisé.
Le projet COBUILD était plus qu'un dictionnaire et un corpus. La création et l'exploitation du corpus ont été décrites par les membres de l'équipe (Sinclair *et al.* 1987). Mais de nombreuses autres applications sont issues de ce projet : des grammaires, des méthodes d'apprentissage, des études linguistiques... Les autres éditeurs de dictionnaires d'apprentissage ont été obligés de suivre, c'est ainsi que le British National Corpus a été créé par un consortium. Le BNC est un corpus annoté et balisé, donc avec une valeur ajoutée importante. Le BNC a fixé de nouvelles normes d'excellence dans la création de corpus, mais est également figé dans le temps, alors que le corpus COBUILD a continué d'évoluer, pour devenir l'actuel *Bank of English*.
Tandis que le corpus COBUILD était extrêmement important en taille pour son époque, d'autres corpus plus petits ont également été créés pour les besoins des études dans les langues de spécialité au sein de l'école de Birmingham.
Ce que nous appelons l'école de Birmingham a commencé dans les années soixante autour de Sinclair et Coulthard. L'école était concernée par les applications dans l'enseignement de la linguistique appliquée. Ainsi nous trouvons la tradition, personnifiée par Palmer, de la recherche appliquée. L'analyse de discours, surtout le discours scientifique, dans le but d'enseigner les langues de spécialité était centrale. Le texte de Barbier (1962) sur les caractéristiques des articles de recherche était le début des analyses sur le genre de Swales (1990). Tandis que Swales et d'autres travaillaient sur l'analyse des textes scientifiques, Roe (1977) travaillait sur un corpus scientifique jetant les bases pour les nombreuses études sur l'anglais de spécialité de l'Université d'Aston.

3. L'ère actuelle

La suite du développement de la linguistique de corpus est liée à la démocratisation des outils informatiques et des ressources électroniques. D'abord l'avènement des clones PC, en commençant avec l'Amstrad, et les Mac-Apple a rendu l'outil disponible à un plus grand nombre. En même temps nous avons vu l'arrivée des concordanciers comme Microconcord (Scott & Tribble) et ATA (Aston Text Analyser de Roe) pour DOS et Conc pour Mac. Il faut souligner que le but n'est pas le développement des outils, mais l'emploi des outils pour regarder les mots en contexte à travers le mot-clé en contexte, KWIC. En linguistique de corpus contextualiste, l'outil informatique n'est qu'une loupe pour mieux voir. L'intérêt se trouve dans le détail : pouvoir généraliser est important, mais non pas formaliser. Ce que nous observons est un réseau de choix, suivant le principe d'idiome (Sinclair 1991). À ce stade, il n'y avait que deux moyens pour obtenir des données : les entrer manuellement, ou utiliser un scanner, un outil encore rare. Il est possible qu'à cette époque les critères de création de corpus aient été mieux suivis : quand les documents sont difficiles à obtenir, on fait plus attention au choix des textes.
L'avènement de Windows a encore simplifié les choses, d'autant plus qu'Internet est rapidement arrivé avec un choix de plus en plus important de documents. Les premiers concordanciers travaillaient uniquement sur du texte ASCII, pour traiter le html, puis le sgml : il a fallu faire évoluer les outils. Ainsi, Microconcord s'est mué en WordSmith Tools (Scott – www.lexically.net) et Conc en MonoConc (Barlow – www.athelstan.com), dorénavant disponible pour Windows. Puis, plus tard le BNC est devenu disponible sur CD-ROM, accompagné de SARA, qui est maintenant

devenue XAIRA, outil pouvant traiter tout corpus en XML, même très basique.

3.1. Qui fait quoi ?

On peut distinguer cinq grands centres de linguistique de corpus, l'Université de Birmingham avec l'équipe de Sinclair, et maintenant Teubert, son successeur dans la chaire de Harper Collins, l'Université d'Aston à Birmingham avec Roe, l'Université de Liverpool autour de Hoey et Scott. Et puis il y a le centre de Lancaster, beaucoup plus TAL dans son approche fondée sur les travaux de Leech, et Oxford, maison mère de la TEI en Europe. Il y a évidement d'autres centres qui se créent avec le mouvement des chercheurs.

Les trois premiers restent plus contextualistes avec un minimum d'intervention sur le corpus, puisque Sinclair défend l'idée de zéro annotation (Sinclair 2005). Le but reste largement l'enseignement des langues, surtout les langues de spécialité, et le développement de la lexicographie. L'autre école se tourne vers des approches plus larges dans la création d'outils d'annotation et les applications typiquement TAL. Cependant, il ne faut pas une histoire de chapelles avec des écoles distinctes. Il y a simplement un continuum avec un glissement vers le TAL dans un sens, et vers d'autres disciplines de la linguistique appliquée dans l'autre.

La linguistique de corpus, *corpus linguistics*, s'est taillée une place de choix dans la linguistique appliquée. La meilleure introduction à l'approche contextualiste reste le livre de Sinclair (1991) *Corpus, Concordance, Collocation*. La différence entre l'approche contextualiste inductive, *corpus-driven*, et d'autres méthodologies est décrite par Tognini-Bonelli, travaillant dans le cadre de l'école de Sinclair. Pour une introduction à la discipline, il faut lire Kennedy (1998), ou Hunston (2002) pour les applications en linguistique appliquée.

3.2. Que font-ils ?

La linguistique de corpus est une linguistique appliquée, la théorie est issue de la pratique, et non l'inverse. La langue est atteinte à travers la parole (Tognini-Bonelli 2001) et n'a pas d'existence propre en dehors du contexte. Ainsi, la linguistique de corpus se trouve en poursuivant la tradition établie par Palmer dans l'enseignement de l'anglais comme langue seconde à des non-spécialistes. Des études sur des corpus scientifiques visent à analyser des problèmes phraséologiques dans l'écrit scientifique (Gledhill 2000) ou la création de dictionnaires d'aide à la rédaction (Williams 2002a). Ces deux derniers étaient des étudiants de Roe, lui-même issu de l'école de Birmingham et élève de Sinclair. L'analyse des corpus scientifiques, soit comme étude linguistique, soit comme aide à la rédaction, est un thème récurrent dans la linguistique de corpus contextualiste (Tognini-Bonelli & Del Lungo Camiciotti 2005). Toujours dans l'enseignement, d'autres travaillent pour faire entrer le concordancier dans la salle de classe (Sinclair (éd.) 2004, Gavioli 2005).

Les applications de la linguistique de corpus sont nombreuses (Hunston 2002), et incluent la linguistique légiste, domaine développé par Coulthard (1994). D'autres études concernent la terminologie (Pearson 1998) ou la traduction (Kenny 2001).

Dans les domaines plus linguistiques, Hunston & Francis (2000) ont mené des études sur des grammaires locales utilisant le corpus COBUILD. Williams (1998, 2002b) a exploré les réseaux thématiques dans un corpus spécialisé et utilise la collocation comme outil de catégorisation. Les patrons thématiques et les mots-clés sont le sujet de nombreuses études (Scott & Tribble 2006). L'analyse de discours sur corpus est un autre domaine important (Stubbs 1996, Partington *et al.* 2004).

Cette liste est loin d'être exhaustive. Le paradigme contextualiste en linguistique de corpus est employé partout dans le monde, sur l'anglais et d'autres langues. Je n'ai pas non plus parlé de l'autre grande tradition de linguistique de corpus représentée par l'ICAME. Les approches sont nombreuses, mais l'objet d'étude reste un corpus constitué selon des critères linguistiques (Sinclair 2005). L'objet est le corpus, les outils informatiques ne sont que des outils pour mieux voir dans le corpus, les objectifs sont toujours une meilleure compréhension du langage parlé par les êtres humains pour les êtres humains, c'est-à-dire la communication.

4. Conclusion

En guise de conclusion, il est temps de faire un petit rappel. Cet article n'entre pas dans la rubrique histoire de la linguistique. Je ne retrace pas des origines pour faire de l'histoire, mais pour expliquer des paradigmes de recherche actuels. Ce n'est pas non plus pour prouver qu'un

paradigme est meilleur qu'un autre, mais que les paradigmes existent, et qu'il faut les regarder et les comprendre afin de créer des échanges et d'avancer dans la recherche sur le sable mouvant que constitue le langage.

La linguistique de corpus est largement issue du monde anglo-saxon, et en anglais le mot linguistique est invariable, c'est une seule et unique discipline avec une multitude de facettes. Parmi ces facettes se trouve la linguistique de corpus : par le jeu de la collocation si chère à Firth, le mot corpus a pris un sens particulier. Il s'agit d'un ensemble de textes soigneusement choisis pour les besoins de la recherche linguistique et qui cherche à représenter une partie de la langue en action. Dans ce sens l'environnement de la langue, avec tous les aspects sociolinguistiques, doit être pris en compte, c'est-à-dire, le contexte culturel et le contexte situationnel. Pour un linguiste de corpus contextualiste il n'est nullement besoin de mettre ces paramètres dans une définition de corpus, c'est un acquis, cela va de soi depuis Malinowski. Dire que le sens du mot corpus est plus restreint en linguistique de corpus n'est pas dire qu'il ne peut pas y avoir d'autres types de corpus, simplement que l'association des mots linguistique et corpus a créé des attentes plus restreintes. Les autres corpus, juridique, littéraire existent, et on peut en faire des études linguistiques : ainsi il existe une linguistique *sur corpus* à côté de la linguistique *de corpus* où la constitution du corpus est en soi une partie essentielle de l'étude.

La ou les linguistiques, je ne vois pas la nécessité d'éclater une discipline sur une simple particule. Le TAL n'est pas la linguistique de corpus, la pragmatique ou la sociolinguistique non plus, chacune a son propre but. Cependant, ils peuvent utiliser les corpus, mais nous sommes de retour sur la linguistique de corpus.

Si la linguistique de corpus existe comme discipline autonome, où se trouvent les frontières avec d'autres disciplines ? Là, je retourne la question : avons-nous vraiment besoin de frontières quand toutes nos propres études sur le langage prouvent que les frontières n'existent pas ? La linguistique de corpus, comme d'autres disciplines de la linguistique, rentre parfaitement dans la notion de prototype, avec un nœud central et une périphérie qui glissera subtilement vers d'autres disciplines dans un continuum. Les catégories n'existent pas en soi, nous les créons pour mieux saisir la complexité. Parler des linguistiques de corpus c'est noyer le poisson, si tout le monde le fait, personne ne le fait, et tout le monde est perdant. La linguistique de corpus existe, elle est récente et sa méthodologie et son épistémologie se forgent. Pour la forger, il faut simplement la reconnaître.

BIBIOGRAPHIE

BARBER, C.L. 1962. Some Measurable Characteristics of Modern Scientific Prose, in J. Swales *Episodes in ESP*, Hemel Hempstead, Pergamon Press, pp. 3-14.

BAZELL, C. E, CATFORD, J. C., HALLIDAY, M. A. K., ROBINS, R. H. (éds.) 1966. *In Memory of JR FIRTH*, London, Longman.

COULTHARD, M. 1994. On the use of corpora in the analysis of forensic texts, *Forensic Linguistics*, 1, pp. 27-44.

FIRTH, J.R. 1935. *The Semantic of Linguistic Science*, in J.R. Firth, *Papers in Linguistics* 1934-1951, Oxford, OUP. 1948.

GAVIOLI, L. 2005. *Exploring corpora for ESP Learning*, Amsterdam, John Benjamins.

GLEDHILL, C. J. 2000. *Collocations in science writing*, Tübingen, Gunter Narr Verlag.

HALLIDAY, M. A. K. 1961. Categories of the Theory of Grammar, *Word*, 17.3, pp. 241-92.

HALLIDAY, M. A. K. 1966. Lexis as a linguistic level, in C. E. Bazell *et al.*, *In Memory of JR FIRTH*, pp. 148-162.

HALLIDAY, M.A.K., HASAN, R. 1976. *Cohesion in English*, London, Longman.

HOEY, M. 1991. *Patterns of Lexis in Text*, Oxford, Oxford University Press.

HOEY, M. 2005. *Lexical Priming: A New Theory of Words and Language*, London, Routledge.

HOWATT, A.P.R. 1984. *A History of English Language Teaching*, Oxford, OUP.

HUNSTON, S., FRANCIS, G. 2000. *Pattern Grammar: A corpus-driven approach to the Lexical Grammar of English*, Amsterdam et Philadelphie, John Benjamins.

HUNSTON, S. 2002. *Corpora in Applied Linguistics*, Cambridge, CUP.

KENNEDY, G.1998. *An introduction to corpus linguistics,* London & New York, Longman.

KENNY, D. 2001. *Lexis and Creativity in Translation*, Manchester, St Jerome Publishing.

LÉON, J. 2005. Claimed and unclaimed sources of *Corpus Linguistics, Henry Sweet Society Bulletin*, N°44, pp. 36-50.
MALINOWSKI, B. 1923. The problem of meaning in primitive languages. Supplement to CK. Ogden and I.A. Richards, *The Meaning of Meaning*, pp. 296-336.
MALINOWSKI, B. 1935. Coral Islands and their Magic, vol 2. The language of Magic and gardening, London, George Allen and Unwin Ltd.
OGDEN, C.K., RICHARDS, I.A. 1923. *The Meaning of Meaning*, London, Routledge and Kegan Paul.
PALMER, H. E. 1933. *Second Interim Report on English Collocations*, Tokyo, Kaitakusha.
PARTINGTON, A., MORLEY, J., HAARMAN, L. (éds) 2004. *Corpora and Discourse : Proceedings of CamConf 2002 Università degli Studi di Camerino, Centro Linguistico d'Ateneo Sept 27th-29th 2002*, Bern, Berlin, Bruxelles, Frankfurt/M., New York, Oxford, Wien, Peter Lang.
PEARSON, J. 1998. *Terms in Context*, John Benjamins.
ROE P. 1977. *Scientific Text*, ELR University of Birmingham.
SINCLAIR, J. McH. 2005. Corpus and Text: Basic Principles, in M. Wynne (éd.), *Developing Linguistic Corpora: A Guide to Good Practice*, pp. 1-16.
SINCLAIR, J. McH. 1991. *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.
SINCLAIR, J. McH., JONES, S., DALEY, R. 2004. *English Collocation Studies: The OSTI Report*, Londres - New York, Continuum.
SINCLAIR, J. McH. (éd.) 1987. *Looking Up: an account of the COBUILD Project in Lexical Computing*, London, Collins.
SINCLAIR, J. McH. (éd.) 2004. How to use corpora in language teaching, Amsterdam, John Benjamins.
SINCLAIR, J. McH. 1966. Beginning the study of lexis, in C. E. Bazell *et al.*, *In Memory of JR FIRTH*, pp. 410-430.
SINCLAIR, J. McH. *et al.* 1970. *English Lexical Studies: Report to OSTI on Project C/LP/08,* Department of English, University of Birmingham.
SWALES, J. M. 1990. *Genre Analysis*, Cambridge, Cambridge University Press.
TOGNINI-BONELLI, E., DEL LUNGO CAMICIOTTI, G. (éds.) 2005. *Strategies in academic discourse*, Amsterdam, John Benjamins.
TOGNINI-BONELLI, E. 2001. *Corpus Linguistics at Work*, Amsterdam, John Benjamins.
TUTIN, A., GROSSMAN, F. 2003. *Les collocations : analyse et traitement*, Amsterdam, de Werelt.
WILLIAMS, G. 1998. Collocational Networks : Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles, *International Journal of Corpus Linguistics*, Vol. 3/1, pp. 151-171.
WILLIAMS, G. 2002a. Corpus-driven lexicography and the specialised dictionary: headword extraction for the Parasitic Plant Research Dictionary, in A. Braasch, C. Povlsen (éds), *Proceedings of the 10th EURALEX International Congress*, Copenhague, CSK, pp. 859-864.
WILLIAMS, G. 2002b. In search of representativity in specialised corpora: categorisation through collocation, *International Journal of Corpus Linguistics*, Vol. 7/1, pp. 43-64.
WILLIAMS, G. 2003. Les collocations et l'école contextualiste britannique, in A. Tutin et F. Grossman, *Les collocations : analyse et traitement*, Amsterdam, de Werelt, pp. 33-44.
WYNNE, M (éd.). 2005. *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford, AHDS.