Constitution d'un corpus oral deFLE : enjeux théoriques et méthodologiques

Najib Arbach

THÈSE / UNIVERSITÉ RENNES 2
sous le sceau de l'Université européenne de Bretagne
pour obtenir le titre de
DOCTEUR DE L'UNIVERSITÉ RENNES 2
Discipline : Linguistique
École doctorale - Arts, Lettres, Langues

présentée par
Najib Arbach

EA 3874 LIDILE
UFR Langues

Constitution d'un corpus oral de FLE

Enjeux théoriques et méthodologiques

Thèse soutenue le 6 février 2015
devant le jury composé de :

Marie-Claude LE BOT
Professeure, Université Rennes 2 / *Directrice de thèse*

Paul CAPPEAU
Professeur, Université de Poitiers / *Rapporteur*

Dominique LEGALLOIS
Maître de Conférences-HDR, Université de Caen / *Rapporteur*

Élisabeth RICHARD
Maître de Conférences-HDR, Université Rennes 2 / *Examinatrice*

RÉSUMÉ

Les méthodologies de constitution de corpus linguistiques ont été amplement étudiées, mais sont moins abondantes quand il s'agit de corpus oraux ; ces méthodologies sont encore plus rares en ce qui concerne l'interlangue orale. Le projet CIL (Corpus Inter Langue), en cours de finalisation à l'Université Rennes 2 et sous la supervision de l'équipe d'accueil LIDILE (EA 3874), vise à la constitution d'un corpus de productions écrites et orales d'apprenants en FLE et ALE. Cette thèse concerne le corpus oral de FLE du projet global (CIL-FLE). Partant du constat que l'intérêt des linguistes pour la langue orale a systématiquement été en retard par rapport à celui porté à la langue écrite, nous nous intéressons dans un premier temps à l'étude de l'oralité dans différents domaines de la linguistique d'un point de vue historique et épistémologique. Le second chapitre est consacré à la linguistique de corpus de manière générale et au corpus en tant qu'objet linguistique en particulier. En ce qui concerne la linguistique de corpus, nous tentons de présenter les différentes méthodologies auxquelles les linguistes ont recours lorsqu'il s'agit de consulter des données : introspection, élicitation ou consultation de données authentiques. Le concept de corpus est ensuite analysé selon un ensemble de critères définitoires que nous étudions en détail, afin de proposer une définition du corpus linguistique. Le troisième et dernier chapitre est la mise en application des constats théoriques dans la constitution du corpus CIL-FLE : nous détaillons les constituants du corpus, les protocoles de collecte et d'archivage. C'est au protocole de transcription que nous nous intéressons en particulier, en insistant sur les difficultés de la transcription de l'interlangue. Le corpus CIL-FLE, qui représente environ 105000 mots, représente le fruit de ce travail et sera ainsi détaillé.

Mots-clés : corpus, corpus FLE, corpus oral, corpus d'apprenants, linguistique de corpus, parole spontanée, transcription, transcription de l'interlangue.

Abstract

The need to design linguistic corpora to support research in linguistics has triggered the development of numerous studies exploring various approaches and methodologies regarding good practices for written corpus building. Fewer studies are available when it comes to spoken data and those that concern the interlanguage of learners are even rarer. The CIL project (Corpus Inter Langue), under completion at the University of Rennes2 and supervised by a research team specialising in the fields of linguistics and pedagogy (LIDILE), aims at building a large corpus of written and spoken productions in EFL and in FFL. This phd dissertation mainly focuses on the FFL (French as a Foreign Language) corpus (CIL-FLE).

The first chapter of the thesis is dedicated to the study of oral speech as a linguistic object from both a historical and an epistemological perspective. The second chapter tackles the question of corpus linguistics generally speaking as well as the concept/ notion of corpus as a linguistic object. Regarding corpus linguistics, we will review and explore the diverse approaches and methods that are used so as to carry out research enquiries: introspection, elicitation or consultation of authentic data. The concept of corpus is then analysed according to/following a series of criteria which we will closely examine in order to propose a definition of the linguistic corpus. The third and last chapter will implement the former theoretical findings through the description of the CIL corpus design. Thus, corpus constituents, transcription and archiving protocols will be described in detail. We are particularly interested in the transcription protocol and we will insist on the difficulties encountered when attempting to transcribe learners 'data. Finally, the CIL-FLE corpus, which contains approximately 105 000 words and was developed all along this phd, will be described.

Key words: corpora, FFL, oral speech, learner corpora, corpus linguistics, spontaneous speech, transcription, interlanguage transcription.

Remerciements

Je remercie Madame Marie-Claude Le Bot pour l'encadrement de cette thèse. Ses conseils et son soutien m'ont été précieux et je lui suis reconnaissant de m'avoir dirigé. Les lacunes et défauts qui subsistent sont de ma responsabilité.

Je remercie les rapporteurs Monsieur Paul Cappeau et Monsieur Dominique Legallois d'avoir accepté d'évaluer ce travail. Je remercie également Madame Élisabeth Richard pour sa présence lors de ma soutenance et pour toute son aide.

Je remercie Madame Martine Schuwer qui a mis en place le projet CIL. Je remercie Saandia Ali pour tous nos travaux communs, ainsi que l'ensemble des membres de LIDILE.

Je remercie mes professeurs à l'Université de Damas, notamment Monsieur Nabil Zreik pour son soutien et son aide. Je remercie le ministère syrien de l'Enseignement supérieur de m'avoir permis de venir poursuivre mes études en France.

Je remercie tous ceux qui ont contribué, de près ou de loin, à la constitution du corpus CIL, nos étudiants en Master 1 et bien entendu les apprenants qui ont accepté d'être enregistrés.

J'exprime ma gratitude à ma famille, Hanna, Madeleine, Soulafa, Layla, Jad, Rimi, Samira et Najib, et je remercie mes amis pour leur soutien.

Table des matières

Table des tableaux et des figures

Tableaux

Figures

Introduction générale

En tant que locuteur de français non natif et enseignant de Français Langue Étrangère (FLE), nous avons souvent constaté une difficulté qui va au-delà de l'évaluation des apprenants : celle de spécifier leur interlangue, d'en cerner les caractéristiques et les divergences avec les normes pour pouvoir décrire leur système et d'en comprendre les fonctionnements. Cette tâche est encore plus ardue si la question est posée sur l'interlangue orale des apprenants. Le point de départ de notre thèse était donc de nous intéresser à cette interlangue orale d'apprenants en FLE.

Nous n'irons pas jusqu'aux extrêmes de la linguistique de corpus, tel Teubert (2005 : 1) qui affirme que « le corpus est considéré la ressource par défaut pour presque tout linguiste », ou Murison-Bowie (1996 : 182), pour lequel « rien de véritablement significatif ne peut être fait sans un corpus » ; néanmoins, avec une problématique telle l'interlangue orale d'apprenants, force est d'admettre que la méthodologie à envisager laissait peu de place à l'introspection. C'est donc la constitution d'un corpus oral d'apprenants en FLE qui a été l'un de nos premiers objectifs. Très vite, nous avons constaté que la constitution de ce corpus n'était pas une étape préliminaire, qui serait une procédure préparatoire à l'étude réelle de l'interlangue. Le corpus n'est pas un ensemble d'outils neutres, mais le produit d'une réflexion et de positionnements linguistiques. Constituer un corpus, c'est déjà offrir une vision du langage, et la constitution d'un corpus d'apprenants en FLE selon des critères scientifiques est devenue le dessein de cette thèse. La rareté des données orales en FLE dans le monde, leur quasi-inexistence en France ont été autant de facteurs encourageants à cette entreprise. En outre, notre inscription en thèse en novembre 2008 a coïncidé avec le lancement du projet CIL[1] au sein de l'équipe LIDILE[2]. CIL est un projet de constitution d'un corpus oral et écrit d'apprenants en FLE et ALE. Nous avons participé aux différentes étapes du projet et avons plus particulièrement pris la responsabilité de la partie orale des apprenants en FLE. C'est dans le cadre de ce projet que la présente étude a donc été effectuée.

Nous avons également constaté que la linguistique de corpus était une notion à la fois ancienne et jeune, en cours de construction. Elle est ancienne car l'empirisme dont fait preuve

[1] CIL : Corpus Inter Langue.

[2] LIDILE : Linguistique et didactique des langues. Équipe d'accueil 3874.

un linguiste en faisant appel aux données n'est pas une nouveauté ; mais elle est toutefois jeune car la linguistique des corpus est tributaire des technologies et ce sont le magnétophone pour les corpus oraux, puis l'ordinateur qui ont permis la naissance des corpus modernes articulés autour de critères définitoires rigoureux.

La linguistique de corpus est également un champ d'activité pluridisciplinaire pouvant être sollicité quel que soit le domaine linguistique. Si nous nous y sommes intéressé en nous penchant sur l'interlangue orale d'apprenants, d'autres l'ont fait en étudiant l'acquisition du langage chez l'enfant ou encore les spécificités du parler parisien contemporain. La polyvalence des corpus a naturellement entraîné une chaîne de problématiques parallèles à résoudre, et les questions que nous nous sommes posées sont les suivantes : comment et pourquoi le linguiste s'est-il intéressé à la langue parlée, et comment ceci a-t-il donné naissance aux corpus oraux ? Quels sont les domaines concernés et quelle est la situation actuelle en ce qui concerne l'accès aux données ? Quelles sont les approches théoriques aux corpus, et que signifie la linguistique de corpus ? Quels sont les critères définitoires d'un corpus et comment le différencient-ils d'une masse quelconque de données ? Quelles sont les étapes de constitution d'un corpus et quelles sont les mesures à prendre avant, pendant et après la constitution ? Si ce travail n'est évidemment pas le premier à aborder ces questions, il vient cependant combler quelques lacunes. Les manuels de référence en linguistique de corpus, ainsi que les études épistémologiques consacrées aux corpus et les guides de constitution (Sinclair, 1991 ; Aijmer & Altenberg, 1992 ; Habert *et al.*, 1997 ; Kennedy, 1998 ; McEnery & Wilson, 2001 ; Meyer, 2002 ; McEnery & Hardie, 2011) n'accordent que peu de place aux corpus oraux dans leurs ouvrages, et les liens entre corpus et études de la langue orale restent flous car aucune étude ne s'est exclusivement intéressée à cet aspect. Au niveau des guides consacrés aux outils de la linguistique de corpus, tels les logiciels, les concordanciers ou les outils d'annotation, des études et des manuels sont régulièrement publiés (Delais-Roussarie, 2002 ; Bazillon, 2011 ; notamment pour l'oral) ou mis à jour, et nous ne présenterons donc pas une revue des logiciels de transcription ou d'annotation. D'autre part, si nous convenons que la difficulté majeure de la constitution d'un corpus oral est sa transcription, les études sur la transcription de l'oral ne manquent pas non plus (Bilger *et al.*, 1997 ; Bilger, 2008). Toutefois, ces travaux concernent généralement la transcription de la langue parlée des natifs ; nous nous proposons ici de nous intéresser à la transcription de l'interlangue orale, qui soulève de nombreuses difficultés que le transcripteur de la langue standard ne rencontre pas.

Ces questions que nous allons traiter visent à répondre à notre problématique principale : comment constituer un corpus oral d'apprenants en FLE qui soit conforme aux normes en vigueur, afin d'en assurer à la fois l'exploitabilité, la pertinence et la mise à disposition de la communauté linguistique. Dans le même temps, un corpus oral d'apprenants en FLE, fût-il un corpus dont les données peuvent être qualifiées « de spécialisation » (Sinclair, 2004), reste un corpus qui partage les caractéristiques communes à tous les corpus linguistiques. Nous avons donc choisi de discuter de ces caractéristiques en général, afin de mieux constituer notre corpus. Nous nous concentrerons cependant sur trois critères : l'oralité des données, l'enseignement des langues et les corpus francophones.

Plan de thèse

En premier lieu, il ne nous semble pas possible de proposer une méthodologie et une réflexion sur la constitution d'un corpus oral sans aborder l'histoire et l'épistémologie des domaines de la langue parlée, du corpus et de sa linguistique ; nous considérons que cette histoire et cette épistémologie ne sont pas dissociables de la science elle-même et il nous paraît donc nécessaire de commencer par une revue historique des méthodologies, des grands corpus collectés dans l'histoire et de la relation dialectale entre méthodologies et théories linguistiques. Cette revue, que nous espérons épistémologique et non simplement historique, sera faite dans le premier chapitre, dans lequel nous commencerons par l'examen des liens tissés entre les différentes théories linguistiques et les technologies d'une part, et la langue orale d'autre part, nous conclurons cette première section par la présentation des sciences du TAL.

Nous enchaînerons par la présentation des grands champs de la linguistique et de l'évolution des bases de données au sein de ces champs en corpus modernes. C'est pourquoi nous tenterons, chaque fois que ceci sera possible, de terminer la présentation par le corpus de référence en la matière.

Nous discuterons donc 1) de l'acquisition du langage chez l'enfant, domaine où la langue écrite n'a pas sa place 2) des chantiers construits par les lexicographes 3) de la description généraliste de la langue orale par le biais de corpus 4) de l'apport méthodologie de la sociolinguistique à la linguistique de corpus et 5) des liens entretenus entre corpus et enseignement des langues.

Le second chapitre sera consacré aux différentes théories de la linguistique de corpus, ainsi qu'au corpus lui-même. Le premier point que nous discuterons est la traditionnelle opposition faite entre description du langage par introspection et description du langage empirique sur base de corpus. Nous nous intéresserons aux fondements théoriques de l'analyse linguistique par le biais de données, afin de cerner les spécificités linguistiques d'une démarche empirique par rapport à la démarche rationnelle dans l'analyse des faits linguistiques. Nous enchaînerons pour ce faire par une discussion des positionnements des génératives, dont les critiques illustrent les limites de la linguistique de corpus. Cette première approche nous permettra de comprendre les enjeux théoriques de la linguistique de corpus et ce qui en découle dans la constitution des corpus au niveau méthodologique.

Le corpus est un objet complexe, non défini, voire indéfinissable tellement ses constituants sont riches et variés. Il peut être écrit, oral, représentatif d'un langage spécial ou être de référence. Les corpus peuvent être également annotés ou pas, ou multimodaux, et les possibilités d'annotations sont nombreuses. Les corpus peuvent être comparables, parallèles, multilingues ou unilingues et à l'heure d'Internet, se pose la question d'y voir un corpus ou pas. Nous retracerons donc succinctement la généalogie des corpus en tant qu'objet scientifique en effectuant une revue de la littérature qui visera à lister les points communs qu'ont proposés les auteurs en tant qu'éléments constitutifs des corpus qui sont la collecte des données, leur numérisation, la structuration des corpus et leur représentativité, ainsi que l'annotation, la documentation et la standardisation des données d'un corpus. Ces éléments peuvent se compléter ou s'exclure. De plus, nombre d'entre eux sont quantifiables. Nous les analyserons donc en détail. Nous n'ambitionnons pas une définition définitive du terme corpus, mais l'étude de l'ensemble des critères qui le régissent et qui le constituent.

Le troisième et dernier chapitre de notre thèse détaillera la constitution proprement dite de notre corpus[3], où nous présenterons les différentes étapes de la constitution. Nous exposerons donc le détail de nos protocoles d'enregistrement qui montreront nos choix quant aux locuteurs sollicités et aux types de données enregistrées. Comme nous le disions, cette thèse ne se veut pas un guide technique pour la constitution des corpus oraux, nous ne discuterons donc que brièvement des outils matériels et logiciels auxquels nous avons eu recours. La

[3] Le corpus est entièrement consultable sur la clé USB jointe à cette thèse. Nous proposons en annexe 1 un guide d'utilisation simplifié permettant l'installation du logiciel, l'écoute des entretiens et la lecture des transcriptions.

présentation du logiciel que nous avons utilisé, CLAN, se voudra une justification du recours à ce logiciel plutôt qu'à un autre, et non pas un guide d'utilisateur. C'est sur la tâche de la transcription que nous insisterons plus particulièrement : le protocole de transcription sera détaillé et les difficultés rencontrées seront illustrées par les exemples que nous avons relevés du corpus. La transcription constituera le noyau de notre réflexion, non parce qu'elle est la procédure la plus chronophage, mais parce que d'elle dépend la qualité du corpus mis à disposition et son potentiel d'exploitation. En outre, nous montrerons qu'une transcription neutre, dénuée de toute interprétation n'existe pas, et que la transcription engage donc la responsabilité du transcripteur.

Ce travail n'est donc pas une recherche linguistique au sens où nous ne défendons pas une thèse du langage. Il constitue une thèse en linguistique appliquée à un problème particulier, celui de la constitution d'un corpus oral d'apprenants en FLE. L'objectif est donc essentiellement méthodologique, et vise à élaborer des démarches rigoureuses pour la création d'un corpus oral d'interlangue et sa libre mise à disposition.

1.Chapitre 1 : Historique des corpus oraux

1.1 Introduction du premier chapitre

Les grammaires classiques ayant longtemps exclu l'oralité et considéré que seule la langue écrite était digne d'être étudiée, analysée ou enseignée, il s'agit de comprendre quels ont été les événements ou les courants de pensée qui amenèrent les linguistes s'intéresser à la langue parlée. En effet, jusqu'à la fin du XIX^{ème} siècle et le début du XX^{ème}, les traditions normatives et prescriptives des grammaires laissaient peu de place aux productions orales spontanées des utilisateurs d'une langue. Jusqu'à Saussure et pour employer sa terminologie, le problème ne fut pas l'absence de distinction entre « langue » et « parole », mais la dénégation de la « parole », et c'est au début du siècle précédent que la grammaire, jusqu'alors limitée à la prescription du bon usage en se référant aux textes écrits, entame une suite de conversions pour devenir une linguistique moderne englobant toutes les modalités du langage, dont la langue parlée. Laks dit à ce propos :

> La linguistique moderne se dégage progressivement de la Grammaire, qu'elle soit normative ou non, en opposant à une rhétorique de l'*exemplum*, une science du *datum*, c'est à dire en définitive en proposant une nouvelle ***approche empirique des faits de langue***. C'est en effet sur la base d'un empirisme affirmé et d'une attention particulière aux faits empiriques de langue saisis comme des usages, c'est-à-dire comme des produits sociaux et culturels que Saussure et Whitney arrachent la linguistique à la grammaire et constituent la science du langage.[4] (Laks, 2008 : 16)

La distinction saussurienne entre « langue » et « parole » et son refus de la primauté traditionnelle de l'écrit sur l'oral, n'a pas habilité la langue parlée en tant qu'objet d'étude sérieux et reconnu d'un jour à l'autre. Très longtemps après Saussure, la langue parlée était encore considérée comme une langue inférieure ou déclassée par rapport à l'écrit, ou comme une « langue populaire », indigne d'être étudiée, tel que le rapportent Blanche-Benveniste & Jeanjean :

> Le français parlé est compris comme du français populaire. C'est une constante, de 1900 à nos jours ; comme si le « non populaire » ne se parlait pas ; ou comme si, parlé, il n'avait aucune caractéristique remarquable. La restriction est de taille ; quantité d'ouvrages qui portent en titre « français parlé » ne s'occupent pas du tout de ce qui se dit en français, oralement, mais seulement de ce qui dit « le peuple ». (Claire Blanche-Benveniste & Jeanjean, 1987, p. 11)

[4] En italique et en gras dans le texte.

Blanche-Benveniste & Jeanjean (1987, p. 12) illustrent leurs propos avec certains exemples d'ouvrages où le terme « populaire », usité pour la qualification du français parlé, est accompagné « d'adjectifs péjoratifs » (français « relâché », français « populaire et argotique », français « familier »). Les auteures constatent que cette tendance a duré jusqu'à des dates bien avancées du XX^{ème} siècle, de surcroît chez des auteurs de renom, comme en témoigne cette citation de Martinet datant de 1969 :

> Ce ne serait pas une boutade de dire que le français populaire n'est pas vraiment le français (…) [il est en effet parlé] par de larges couches prolétariennes et paysannes qui n'ont ni le loisir ni les moyens de cultiver chez eux-mêmes des exigences linguistiques. (cité par Claire Blanche-Benveniste & Jeanjean, 1987, p. 13)

D'un point de vue théorique, Ferdinand de Saussure n'est pas non plus à être considéré comme un pionnier dans l'étude de la langue parlée car l'intérêt des enseignants-phonéticiens comme Viëtor, Passy, l'abbé Rousselot, Gouin ou Schweitzer pour la langue parlée fut antérieure aux travaux de Saussure[5].

Pour rester dans le domaine de l'enseignement des langues, les recherches de Fries[6] et de Gougenheim, Sauvageot & Rivenc[7], qui élaborèrent respectivement *Structure of English* et *Le Français fondamental* dans les années 1950, furent induites par la volonté de proposer un manuel d'enseignement simplifié aux apprenants de l'anglais et du français, et non par les théories linguistiques de Saussure. Dans d'autres domaines, telle l'acquisition du langage chez l'enfant, l'intérêt pour la langue parlée est né avec la discipline elle-même, de par la nature de cette dernière. Il en va de même pour des domaines comme la dialectologie qui remonte aux années 1820[8]. Le rapport entre langue et parole et la distinction opérée par Saussure est donc une théorisation qui, comme le rapporte Rastier, avait déjà été opérée dans d'autres domaines, et qui est à rapprocher de la distinction que fera Chomsky entre « compétence » et « performance » :

> Traditionnellement, le rapport entre une grammaire et les productions linguistiques qu'elle règle est conçu comme un rapport entre la *puissance* et l'*acte* (dans la tradition aristotélicienne), ou encore entre *energeia* et *ergon* (selon Humboldt qui la reprend),

[5] Cf. 1.7.1.

[6] Cf. 1.6.2.

[7] Cf. 1.7.2.3.

[8] Cf. 1.5.1.

> ou enfin entre *compétence* et *performance* (selon Chomsky, qui se recommandait de Humboldt sur ce point). (Rastier, 2005)

Ainsi, le *Cours de linguistique générale* n'influa donc pas directement et immédiatement sur l'étude de la langue orale, si ce n'est sur les travaux de Damourette & Pichon qui s'en revendiquèrent[9]. La preuve pourrait en être également que, près de quarante ans plus tard, Gougenheim *et al.* furent « violemment accusés de fournir une mauvaise image de la langue française » (Blanche-Benveniste, 2010 : 10) lors de la parution du *Français fondamental* et, en 1987, lorsque Blanche-Benveniste & Jeanjean publient *Le français parlé*, il s'agissait du premier état des lieux faisant le bilan des recherches sur l'oral car, même à cette époque, « peu de gens y voyaient un objet légitime d'étude, même chez les linguistes ». L'intérêt pour l'oral dans les différents domaines de la linguistique ne fut pas le résultat d'un mouvement de pensée unique et bien défini. Chaque discipline voit quelques-uns de ses spécialistes se pencher sur la langue parlée mais à des époques parfois très éloignées les unes des autres, preuve qu'il ne s'est point agi d'une progression théorique commune.

Quelques événements viendront pourtant susciter ou accélérer l'intérêt pour l'oral dans l'ensemble des domaines de la linguistique, mais ils sont d'ordre technique. L'invention d'appareils d'enregistrement fut ainsi un moteur, sinon le déclencheur des études de la langue orale. En 1911, le grammairien et historien de la langue française Ferdinand Brunot, titulaire depuis 1900 de la chaire d'histoire de la langue française à la Sorbonne, inaugure « Les Archives de la Parole »[10] ; concernant le recueil et l'archivage des données sonores, c'est une première institutionnelle en France. Inspirés par des projets étrangers comme le « Phonogrammarchiv » de Vienne[11] et se basant sur les travaux de Rousselot, les objectifs des « Archives de la parole » étaient les suivants[12] :

- enregistrer une représentation fidèle des accents en France, telle « la parole nuancée d'accents faubourien ou provincial » ;

[9] Cf. 1.6.1.

[10] http://gallicadossiers.bnf.fr/ArchivesParole/

[11] Le Phonogrammarchiv de Vienne a été créé en 1899 par des membres de l'Académie autrichienne des sciences: c'est le plus ancien fonds d'archives sonores du monde. Le Phonogrammarchiv a constitué des collections qui comptent aujourd'hui plus de 50.000 documents enregistrés et représentent près de 7.000 heures d'enregistrement.

[12] Cf. le site des Archives de la Parole.

- enregistrer les patois et les dialectes, en vue de constituer un atlas linguistique et phonographique de la France ;
- enseigner les langues étrangères, en l'occurrence le français en tant que langue étrangère ;
- traiter des pathologies de la parole.

Ferdinand Brunot s'impliqua lui-même dans la collecte des données, et enregistra patois, dialectes et accents dans différentes régions de France[13]. La disponibilité d'appareils d'enregistrement constituait pour lui une avancée majeure, sur laquelle il insista lors du discours d'inauguration des « Archives de la parole » :

> La création d'appareils qui enregistrent et reproduisent la voix humaine, complète une série d'inventions qui a commencé le jour où l'homme a imaginé de dessiner un premier symbole pour représenter sa pensée. Après l'écriture, après l'imprimerie, il restait encore un progrès essentiel à faire, car ni l'une ni l'autre ne fixe ni ne transmet la parole dans son intégrité absolue. (…) Aucune écriture phonétique, si chargée qu'elle soit de signes diacritiques, ne nous rendrait les accents, les intonations que nous n'avons pas entendus. (Brunot, 1911)

Comme le prévoyait Brunot et comme le prévoira William Stern[14], l'histoire des sciences humaines ayant trait à la parole en général, et celle des corpus oraux en particulier, fut un corollaire des premières « machines parlantes »[15] qui seront suivies des enregistreurs portatifs, puis du numérique. Plusieurs auteurs insistent sur ce point. Michael A.K. Halliday, cité par Blanche-Benveniste (2005, 2010; Claire Blanche-Benveniste & Jeanjean, 1987) voyait l'invention des enregistreurs comme la naissance de la linguistique moderne. Brunot, dans son discours (cité par Galazzi, 1995), ainsi que Queneau (1965), bien plus tard, considérèrent l'invention du phonographe comme ayant provoqué en linguistique une révolution comparable à celle du microscope en sciences. L'ère où l'on ne pouvait citer l'oral que par intuition ou par mémorisation (Blanche-Benveniste (2010 : 9) parle d'un maximum de 20 secondes) tel que le faisaient Damourette & Pichon, était révolue. Les premiers phonographes

[13] En outre, en visionnaire de l'analyse de la langue orale, Ferdinand Brunot avait enregistré quelques disques témoignant de la langue orale spontanée de l'époque : deux conversations entre lui-même et deux autres individus, et deux disques retraçant les énoncés d'un marchand ambulant (cf. Gougenheim, Rivenc, & Sauvageot, 1956 : 62).

[14] Cf. 1.3.1.

[15] Les premiers phonographes étaient ainsi nommés.

nécessitaient toutefois des laboratoires d'expérimentation et de fait, la collecte était difficile et onéreuse. La véritable révolution vint donc avec l'invention des magnétophones portables. De Fornel & Léon disent à ce propos :

> Enfin il faut noter que l'intérêt pour les données enregistrées connaît un essor sans précédent avec l'invention des magnétophones portables dans les années 60. Ceux-ci offrent des possibilités inédites, comme celles d'enregistrer un nombre beaucoup plus grand d'interactions et d'écouter et réécouter indéfiniment les exemples de parole spontanée. (De Fornel & Léon, 2000 : 136)

Les enquêtes ne sont plus limitées aux laboratoires, et ne nécessitent plus le transport d'un matériel lourd et peu maniable, comme ce fut le cas pour le *Français fondamental*, dont l'appareil enregistreur pesait plusieurs kilos[16]. Il est dorénavant possible d'enregistrer les locuteurs sur le terrain, telles les enquêtes sociolinguistiques de Labov[17], ou de mener des enquêtes longitudinales sur l'acquisition du langage chez l'enfant en l'enregistrant chez lui, dans son milieu quotidien. Luzzati résume la période ainsi :

> À partir des années 60, à la suite du français fondamental, les corpus ont commencé à fleurir, parallèlement à la diffusion du magnétophone. Les appareils deviennent plus petits et plus autonomes. Les microphones s'améliorent et deviennent directionnels. Et surtout les supports évoluent. Les disques de carton ou de cire sont remplacés par des bandes magnétiques puis par des cassettes d'une autonomie de plus d'1h. Enregistrer devient peu à peu facile, courant, presque banal. (Luzzati, 2009 : 2)

Cet essor, dans les années 1960, coïncida avec les premiers ordinateurs mis à la disposition des équipes de recherche, et donc avec la possibilité de numériser le texte, puis également le son ; c'est l'époque des premiers « machine-readable corpora ». L'apport de l'informatique aux corpus oraux est indéniable ; pour la première fois, il fut possible d'enrichir et de modifier le flux sonore, dont la pérennité était assurée grâce à la numérisation, *a contrario* de tout support magnétique qui subit inévitablement une dégradation avec le temps. Les ordinateurs facilitèrent également la manipulation des données ; les informaticiens élaborèrent des logiciels qui permettent d'aligner le son avec sa transcription et il est devenu alors possible d'écouter en particulier une partie de l'enregistrement, tout en visualisant la transcription qui en a été faite. Cette représentation du signal sonore sous forme transcrite numérisée augmente substantiellement la rapidité et la fiabilité de deux démarches : 1) la vérification, la

[16] Cf. 1.7.2.3.

[17] Cf. 1.4.2.

modification ou l'annotation des données. 2) la recherche d'occurrences particulières et l'analyse des données.

Un essor important est donc dû aux magnétophones portables dans les années 1970, tel que le rapportent Blanche-Benveniste & Jeanjean, sans que ceci ne signifiât pour autant la création d'un grand nombre de corpus exploitables, en raison de la problématique de la transcription :

> La vogue du magnétophone un provoqué un déferlement de documents de toute sorte : « roman-magnéto » des années 1965-1975, tranches de vie, etc. RADIO-France a édité en 1983 un catalogue des cassettes et transcriptions qu'elle vend au public ; ce sont des émissions de radio, des interviews, des enquêtes, « une sélection établie à l'intention des établissements scolaires ». Les transcriptions, même lorsqu'elles essaient d'être « très fidèles », ne sont pas utilisables pour une étude linguistique. (Blanche-Benveniste & Jeanjean, 1987 : 47)

Aux côtés des ordinateurs grand public, des magnétophones portables et des moyens de stockage, la technologie tenta donc de trouver une solution à l'un des principaux obstacles sur le parcours de la constitution d'un corpus oral, qui est la transcription des données, en mettant au point des techniques de reconnaissance automatique vocale (Automatic Speech Recognition ou ASR[18]). Geoffrey Leech (in Aijmer & Altenberg, 1991 : 11) dit à propos de l'ASR que tant qu'elle ne sera pas au même niveau que la reconnaissance optique de caractères (Optical Character Recognition ou OCR), la collecte de données de langue orale à même échelle que de la langue écrite restera un rêve dans l'avenir : « the collection of spoken discourse on the same scale as written text will remain a dream of the future ». Plus de vingt ans après ce constat, force est de constater que la transcription manuelle est loin d'être devenue ce que les transcripteurs avaient souhaité qu'elle devînt : un mauvais souvenir, et les données orales disponibles sont encore loin d'être à l'échelle des données écrites. La révolution se limite donc à la facilité de *collecte* des données, surtout à l'époque d'Internet, ne serait-ce que pour les données audio-visuelles.

Selon les auteurs, les tournants majeurs en linguistique de corpus sont différents. Ainsi Hardie, McEnery &Wilson (McEnery & Hardie, 2011; McEnery & Wilson, 2001) utilisent les termes « Early corpus linguistics » pour définir les linguistiques de corpus antérieures aux théories de Chomsky, et lient ainsi l'évolution des corpus à la théorisation linguistique. Certains, comme nous l'avons vu, mettent l'accent sur les appareils d'enregistrement.

[18] Cf. 2.7.2.1.

D'autres considèrent la machine informatique comme le tournant majeur de la linguistique de corpus, à l'instar de Francis (1991) qui offre un panorama pré-informatique dans son exposé sur l'épistémologie des linguistiques de corpus.

Si la langue parlée a longtemps été négligée, cela n'est dû qu'en partie à une certaine forme de conservatisme des scientifiques d'avant le XX^{ème} siècle ; la raison est en réalité essentiellement dans le fait qu'ils n'avaient pas la possibilité de capturer le flux sonore pour l'étudier[19]. Les technologies évoluant et se diffusant par à-coups, la constitution des corpus oraux a suivi le même chemin saccadé, plus ou moins irrégulier selon les différentes disciplines. Après quelques rares exceptions jusque dans les années 1950, les corpus commencent à se multiplier dans les années 1970. Il eût peut-être encore été possible d'opérer un recensement vers la fin des années 1980, car les années 1990 et les années 2000 virent la multiplication des corpus avec la démocratisation des ordinateurs. De nos jours, le nombre de corpus disponibles est encore moindre que celui des corpus en projet. Afin de retracer plus en détail leurs parcours, nous avons opté une présentation par discipline. Mais pour souligner l'importance de la technologie dans l'histoire des corpus oraux, nous commencerons par un retour sur le domaine linguistique le plus intimement lié à l'informatique, à savoir le TAL. Pour chacune des disciplines que nous présenterons, notre démarche sera la suivante : nous commencerons par un retour sur les premiers corpus oraux constitués (ou ayant une partie orale) ou ébauchés pour terminer par un état des lieux des corpus oraux actuels en français. Notamment en ce qui concerne ces derniers, nous nous efforcerons de ne traiter que des corpus mis à disposition, où à défaut ceux dont la constitution a permis une avancée sur les plans méthodologique et théorique de la linguistique de corpus.

1.2 Corpus et TAL

Le « Traitement Automatique des Langues » ou TAL, parfois appelé « Ingénierie Linguistique », est à la croisée de la linguistique et de l'informatique sous toutes les formes de cette dernière : des premiers calculateurs électroniques à l'intelligence artificielle en passant par les ordinateurs personnels les plus récents. Le TAL a pour objet l'application de procédures informatiques capables de traiter automatiquement, ou semi-automatiquement les

[19] Nous reviendrons sur la dépendance des corpus oraux aux évolutions technologiques en 1.6.3.

données des langues naturelles. Quel que soit le champ d'application du TAL[20], l'intérêt d'un traitement automatique réside en particulier dans la rapidité du traitement des données et dans la fiabilité de ces traitements. Pour que ces procédures soient possibles, le TAL nécessite :

1) la mise à disposition de données langagières aux chercheurs travaillant sur le projet afin de constituer des outils formels d'une part, et l'exploitation des outils informatiques sur ces données d'autre part, soit la mise à disposition de corpus exploitables ;
2) des outils linguistiques qui traitent des diverses informations relatives aux langues traitées ;
3) des outils formels qui exprimeront ces connaissances dans un langage artificiel, à savoir un formalisme qui conviendra à un traitement automatique des données ;
4) des outils informatiques (machines, logiciels, concordanciers, parseurs etc.) pour la mise en application des possibilités des outils 2) et 3) sur le corpus.

La création et l'exploitation de corpus électroniques a donc permis des mises en application statistiques basées sur la théorie de l'information de Claude Shannon (1949) et des calculs probabilistes (Manning & Schütze, 1999). Nous citerons pour exemple la création de dictionnaires entièrement basés sur corpus, tel le *Collins COBUILD English Language Dictionary* en 1987, ou les calculs fréquentiels récents qui visent à la création de « lexiques de base » destinés à l'apprentissage des langues.

Il ne faut pourtant pas confondre TAL et linguistique de corpus, même si le corpus est l'instrument de travail principale de ces deux disciplines qui se différencient du point de vue de leurs objectifs affichés. En effet, alors que l'objectif du TAL reste l'automatisation et la création d'outils permettant les observations langagières, le TAL demeure un outil pour la linguistique de corpus, dont l'objectif est la langue dans sa globalité. Rastier dit à ce propos :

> Pour la linguistique de corpus, l'informatique n'est qu'un instrument, non un modèle théorique, car la linguistique appartient pleinement aux sciences de la culture. Il serait fallacieux de postuler une épistémologie propre et donc une autonomie scientifique des «Traitements Automatiques du Langage». Il n'est aucunement certain en effet que l'informatique, technologie sémiotique, soit en outre une science, car le traitement de l'information n'est pas un objet scientifique, mais un objectif ; et l'information reste

[20] Nous citerons pour exemples la traduction automatique, la correction orthographique, la fouille de textes, l'annotation de corpus, la synthèse de la parole, la reconnaissance vocale, l'archivage de documents, la reconnaissance automatique de caractères, etc. Le TAL n'a pas « inventé » de nouvelles linguistiques appliquées, mais a rendu rapides et fiables des procédures le plus souvent chronophages et inévitablement approximatives.

> un objet des mathématiques. Un traitement constitue évidemment un objectif technique et non un objet scientifique : confondre les deux, c'est instituer une technoscience qui trouverait sa légitimité dans les outils. De fait, les TAL doivent leur scientificité à la linguistique dont ils constituent un secteur d'application et dont ils doivent pour ainsi dire « hériter les propriétés ». (Rastier, 2005 : 41)

Il convient de revenir sur les origines du TAL, en raison de ce que représente cette discipline au service de la linguistique de corpus. Pour ce faire, c'est la traduction automatique (TA) qu'il faut étudier car elle représente le domaine pionnier du traitement automatique des langues. Gross, évoquant l'histoire de la TA, note en effet qu'« il est vraisemblable qu'elle ne peut pas être séparée de l'histoire de l'intelligence artificielle », donc « d'un secteur de l'informatique » (1972 : 40) et Anis la présente comme « la locomotive de la linguistique formelle et de ce qui va devenir la linguistique computationnelle » (1994 : 112). À ses prémisses, les raisons de l'intérêt quasi exclusif du TAL pour la TA sont attribuées au contexte historique des années 1940 : les débuts de la guerre froide et la rivalité technologique de l'Occident avec le bloc URSS (course à l'armement et conquête de l'espace) poussent les gouvernements à subventionner essentiellement les recherches en TA pour les traductions du russe[21]. Ce n'est qu'avec l'apparition des machines électroniques à usage non militaire, dans les années 1940, qui a donc donné naissance aux premières réelles tentatives et théories d'un traitement automatisé des langues : William Weaver préconise dans un « Memorandum »[22,23] l'utilisation des techniques du déchiffrage cryptographique pour traduire des textes de façon automatique, considérant que celles-ci sont suffisantes à une TA de qualité. Un premier rapport de Yehoshua Bar-Hillel paraît en 1951, et en 1952 se tient au M.I.T. la première conférence sur la TA. En 1954, une expérience est tentée à New-York sur un petit nombre de

[21] Nombre d'articles et d'ouvrages traitent du conflit russo-occidental et de ses conséquences sur la TA (Archaimbault & Léon, 1997; Cori & Léon, 2002; Fuchs, Danlos, Lacheret-Dujour, Luzzati, & Victorri, 1993; Léon, 2001; Loffler-Laurian, 1996). Vers le milieu des années 1930 et donc à l'ère pré-informatique, il est vrai qu'il y eut une tentative, du moins une réflexion, avec l'idée du Franco-Arménien Georges Artsrouni et du Russe Petr Troyanskii d'une machine à traduire automatisée (voir Hutchins, 2005b). Ils proposèrent non seulement la possibilité d'un dictionnaire bilingue automatique, mais également un schéma de codage grammatical qui opérerait l'analyse et la synthèse des textes à traiter. Mais leur travaux ne furent connus que vers la fin des années 1950, les premiers ordinateurs étaient déjà nés.

[22] Ce « Memorandum » est intitulé « Translation ». Il date de 1949.

[23] Il y a eu, en 1947, des tentatives de la part de R.H. Richens et A.D. Booth, mais c'est le « Memorandum » de Weaver qui lance la TA, voir à ce sujet « Ordinateurs et traduction : survol d'un demi-siècle » (Anis, 1994) et « Les débuts de la traduction automatique en France (1959-1968): à contretemps? » (Léon, 1998).

phrases russes, avec un vocabulaire de 250 mots et six règles de grammaire élaborées à la Georgetown University. La médiatisation de l'expérience accélère les recherches au Canada, en Grande-Bretagne, en Italie et en URSS. Aux États-Unis, les recherches sont subventionnées par la « National Science Foundation », mais aussi par la CIA. Les moyens financiers et humains engagés dans la TA durant les années cinquante sont considérables[24]. Dans le domaine de la TA travaillaient à la fois des ingénieurs, des informaticiens, des mathématiciens, mais également des linguistes[25], et c'est là le point qui nous intéresse : nous allons voir dans la section suivante que la TA fut l'une des premières disciplines où il fut exprimé des points de vue scientifiques quant aux corpus. Bien que ces premiers points de vue fussent négatifs, ils représentaient pourtant la naissance de la linguistique de corpus.

1.2.1 TAL, TA et théories linguistiques

Dans son « Memorandum », Weaver évoque certains aspects théoriques de la TA en abordant des pistes de réflexion telles les structures logiques du langage ou les grammaires universelles. De même, Bar-Hillel, Zellig S. Harris et Victor Yngve[26] aborderont la TA d'un point de vue linguistique. Ainsi Bar-Hillel (1953a) présente la nécessité d'une syntaxe opérationnelle qui puisse déterminer la structure syntaxique d'une séquence donnée[27]. Yngve (1955) voit dans la syntaxe un moyen de reconnaissance et d'analyse des structures, alors que les autres groupes de la TA voyaient en elle uniquement une procédure de réarrangement de l'ordre des mots dans le texte cible. Enfin Harris, dans son article intitulé « Transfer Grammar » (1954), élabore un modèle qui permettrait le transfert des structures syntaxiques, morphologiques et phonologiques d'une langue à une autre, soit un modèle linguistique exploitable par et pour la machine.

Néanmoins, la prise en compte des problématiques théoriques linguistiques ne donne pas lieu à des applications et les résultats se sont pour la plupart soldés par des échecs, que certains

[24] Pour une histoire institutionnelle plus détaillée de la TA, voir les travaux de Hutchins (Hutchins, 2003, 2005a, 2005b; Hutchins & Somers, 1992) et de Léon (Archaimbault & Léon, 1997, 1997; Léon, 1998, 2001, 2002).

[25] La plupart des acteurs de la TA étaient des ingénieurs, des mathématiciens, des philosophes et des spécialistes de langues naturelles ; les linguistes étaient minoritaires.

[26] Il ne s'agit pas ici de détailler toutes les références traitant de la théorisation linguistique de la TA, mais Bar-Hillel, Yngve et Harris en sont les figures de proue.

[27] Cf. 1.2.1.1.

auteurs (Anis, 1994; Fuchs *et al.*, 1993) imputent à la méthode adoptée, dite méthode d'approche directe et qui consiste à une simple automatisation de la traduction mot à mot. Fuchs *et al.* (1993) notent que seules les recherches menées en URSS à cette époque avaient une approche de la TA plus théorique grâce aux travaux de linguistes russes qui ont formalisé les notions de « langage intermédiaire universel » et de « système intermédiaire ». Les deux principaux courants linguistiques américains de l'époque, à savoir le structuralisme et le générativisme, furent éloignés de la linguistique de corpus et du TAL pour des raisons que nous allons présenter.

1.2.1.1 Structuralisme et TAL/TA

Le fait que le TAL fût en grande partie orienté vers la TA uniquement éloigna les linguistes structuralistes des évolutions du TAL pour des raisons que Bar-Hillel expose dans un article (1953b), où il démontre l'incompatibilité des ambitions de la TA avec les théories du structuralisme. Les incompatibilités ont été principalement d'ordre sémantique et Hutchins (2005b) les nomme les *semantic barriers*. Dans son article, Bar-Hillel rapporte quatre obstacles théoriques majeurs à une TA totalement indépendante de l'intervention humaine : 1) l'élaboration d'une syntaxe opérationnelle 2) la question de l'intraduisibilité des langues naturelles 3) les idiomes et expressions idiomatiques 4) les catégories syntaxiques universelles. Selon cet auteur, deux de ces points sont insolubles du point de vue de la linguistique structurale :

1) Syntaxe opérationnelle

Bar-Hillel pose comme condition *sine qua non* à la construction d'une machine de traduction l'élaboration d'une syntaxe opérationnelle capable de déconstruire une séquence donnée en ses éléments constitutifs, et de déterminer la fonction syntaxique de chacun de ces éléments. Bar-Hillel propose pour cette démarche la notion de « analytic syntax ». Or, la linguistique structurale fournit une description exhaustive des éléments constitutifs de la langue, mais n'est pas en mesure de déconstruire une séquence donnée en ses éléments constitutifs. Elle opère ce que Bar-Hillel appelle une « synthetic syntax », qui est insuffisante pour une TA opérationnelle.

2) Idiomes et expressions idiomatiques

Bar-Hillel rapporte la définition de « idiom » suivante : « an expression in the usage of a language, that is peculiar to itself either in grammatical construction or in having a meaning which cannot be derived as a whole from the conjoined meanings of its elements (as, the more the merrier, a picture of the king's, to make friends with him) », et note que le terme « meaning » y apparaît deux fois. Pour traduire, la machine doit aborder la question du sens, incompatible avec les théories structuralistes pour lesquelles le langage peut être décrit exhaustivement sans avoir recours à la sémantique. Or la machine appliquant la TA doit, à un moment ou un autre, pouvoir comprendre telle ou telle séquence : « some step has to be taken which directly or indirectly depends upon the machine's ability to understand the text on which it operates. »

Plus tard et en raison des échecs répétés des équipes ayant travaillé sur la TA durant les années 1950, leurs méthodes empiriques souvent dénuées de méthodologie linguistique élaborée et l'impossibilité de mise en application du modèle de Harris, Bar-Hillel publie « A demonstration of the nonfeasibility of fully automatic high quality translation » (Bar-Hillel, 1960), dans lequel l'auteur reconnaît l'impossibilité, à court terme, d'une TA entièrement automatisée de qualité :

> One of the reasons why we do not as yet have any translation centers (…) is the reluctance of many MT workers to recognize that the idea of inventing a method for fully automatic high quality translation (FAHQT) is just a dream which will not come true in the foreseeable future. (Bar-Hillel, 1960)

Dans son article, Bar-Hillel présente les obstacles techniques, insurmontables en 1960 : une machine ayant une capacité de mémorisation suffisante au stockage d'un lexique encyclopédique universel ainsi que la syntaxe opérationnelle qu'il avait évoquée en 1953. De plus cette machine devait pouvoir être suffisamment puissante pour effectuer les calculs imposés. Si de nos jours ces objectifs techniques sont beaucoup moins inenvisageables, de telles performances semblaient impossibles alors :

> Whenever I offered this argument to one of my colleagues working on MT, their first reaction was: "But why not envisage a system which will put this knowledge at the disposal of the translation machine?" Understandable as this reaction is, it is very easy to show its futility. What such a suggestion amounts to, if taken seriously, is the requirement that a translation machine should not only be supplied with a dictionary

> but also with a universal encyclopedia. This is surely utterly chimerical and hardly deserves any further discussion. (Bar-Hillel, 1960)

Cependant Bar-Hillel imagine que, quand bien même une machine de ce type pourrait être créée, deux obstacles restent insurmontables dans le cadre d'une FAHQT: la machine ne pourra jamais déduire des réalités à partir de celles présentées, et la machine ne pourra jamais aborder les problématiques liées à la polysémie dont la résolution suppose la mise en œuvre de principes de référenciation, apanage de l'esprit humain. Pour étayer son propos, Bar-Hillel propose un exemple demeuré célèbre : « the box was in the pen », dans lequel « box » signifie boîte à jouets, et « pen » indique parc. C'est pourquoi il faut concevoir une « high quality translation by a machine-post-editor partnership » qu'Anis (1994 : 114) exprime en ces termes : « traduction de grande qualité grâce à la collaboration entre la machine et le réviseur ». Il est à noter que Weaver, dans son « Memorandum » de 1949, écartait déjà l'hypothèse d'une traduction parfaite qui n'ait pas recours à l'humain. Ainsi, assez tôt dans la relation humain-machine dans le traitement du langage, les théoriciens ont constaté l'irremplaçabilité de l'intuition humaine. Le rôle du linguiste dans le TAL, le degré d'intuition que l'on doit laisser s'exprimer et le degré de neutralité supposé ou possible sont des questions qui sont toujours d'actualité, *a fortiori* dans le domaine de la linguistique de corpus et des multiples méthodologies dans la manière d'appréhender les données.

1.2.1.2 Générativisme et TAL/TA

En 1955, Noam Chomsky rejoint le Laboratory of Electronics au M.I.T. Research, sous la direction d'Yngve, pour travailler un projet de TA, avec une subvention de la National Science Foundation. Il n'a pourtant à aucun moment travaillé directement sur la TA, ni sur aucun projet de TAL. Après la démission de Yngve et la dissolution de l'équipe, Chomsky rejette de se rattacher au traitement automatique des langues et lui et ses partisans feront maintes fois état de leur hostilité au principe de recherches en TA/TAL (Cori & Léon, 2002; Gross, 1972). Nous pouvons constater que la grammaire générative a aussi bien rejeté l'idée d'un traitement informatique du langage que l'idée de l'exploitation de corpus pour l'analyse des langues. L'approche empirique de la langue, inhérente au TAL et à la linguistique de corpus, n'est pas compatible avec les théories rationalistes de Chomsky[28]. D'autre part, Gross (1972 : 42) relève que, de toute manière, « les théories linguistiques modernes (théories

[28] Nous détaillerons le point de vue des génératives vis-à-vis des méthodologies empiriques en 2.2.

transformationnelles de Z. S. Harris et de N. Chomsky) bien qu'extrêmement fructueuses, et riches d'applications potentielles sont trop récentes pour qu'on puisse envisager leur usage pour des réalisations de type industriel, même à un niveau expérimental ».

1.2.2 Déclin temporaire du TAL aux États-Unis, naissance du TAL en France

Suite aux échecs successifs d'une TA viable, au désintérêt progressif des linguistes pour ce sujet et en raison de l'article de Bar-Hillel, l'Automatic Language Processing Advisory Committee (ALPAC) est fondé en 1964 pour des raisons politiques par le gouvernement américain, dans le but d'évaluer le TAL en général et la TA en particulier. L'ALPAC publie son rapport en 1966[29]. Dans ce dernier, l'apport de l'informatique à la linguistique est indéniable et clairement énoncé, et les perspectives envisagées sont qualifiées de révolutionnaires :

> The advent of computational linguistics promises to work a revolution in the study of natural languages (…). The revolution in linguistics has not been solely a result of attempts at machine translation and parsing, but it is unlikely that the revolution would have been extensive or significant without these attempts. (« Language and Machines », 1966, p. 29□30)

De plus, le rapport donne deux raisons possibles motivant le TAL : la constitution de modèles formels, et la possibilité d'accès à de grands corpus. Cependant, sur la base de données statistiques des coûts de revient, le rapport concentre son évaluation sur la TA des documents russes vers l'anglais, et les résultats sont jugés selon le besoin militaire du gouvernement américain. Les résultats n'étant pas à la mesure des investissements colossaux, la TA est discréditée et les subventions furent stoppées, retardant quelque peu l'évolution du TAL pour quelques années. En effet, le rapport mentionne : « There is no immediate or predictable prospect of useful machine translation. », et les recommandations préconisent une assistance technique aux traducteurs humains et l'élaboration de dictionnaires automatiques bilingues.

[29] Ce rapport est consultable en ligne :
http://www.nap.edu/openbook.php?record_id=9547
John Hutchins (2003) en fit un commentaire détaillé et en évaluera l'impact sur le TAL.

Les recherches aux États-Unis et dans le reste du monde en TAL commencent alors à s'intéresser à autre chose que la TA uniquement[30].

En France, l'intérêt pour le TAL commence ainsi à la date où l'article de Bar-Hillel constate l'échec de la TA. Les raisons sont diverses et principalement institutionnelles. La linguistique des années 1940/1950, traditionnelle, est essentiellement dominée par la stylistique et la philologie des langues classiques et des formes anciennes des langues européennes (Chevalier & Encrevé, 1984). La France manque de spécialistes des nouvelles technologies et le calcul électronique n'est pris au sérieux ni par les entreprises, ni par les universitaires. Ainsi, des linguistes comme Émile Benveniste, Georges Gougenheim, Michel Lejeune ou André Martinet ne collaboreront à aucun projet lié au TAL. Il fallut attendre 1959, où la politique gaullienne de favorisation de la recherche ainsi que les demandes nouvelles de traduction du russe et de la documentation technologique lancèrent le TAL, qui doit donc son essor en France à des raisons politiques et à l'argentisation du secteur. Les démarches d'Émile Delavenay, alors directeur du service des documents et des publications de l'UNESCO, aboutissent à la constitution autour de lui d'un groupe de travail sur la TA au courant des travaux américains et russes. Cette année, Delavenay créé l' « Association pour l'étude et le développement de la Traduction Automatique et de la Linguistique Appliquée » (ATALA), devenue l'« Association pour le Traitement Automatique des Langues ». Bernard Vauquois et Jean Kuntzmann fondent le « Centre d'Études sur la Traduction Automatique » (CETA) qui deviendra le « Groupe d'Étude en Traduction Automatique » (GETA) puis le « Groupe d'Étude en Traitement Automatisé des Langues et de la Parole » (GETALP). Les termes « Computational Linguistics » et, dans un second temps, « Natural Language Processing » auront différentes traductions au fil du temps. Les termes « Traitement Automatique des Langues » s'imposèrent en France au début des années 1990[31].

[30] Ceci de manière officielle car, comme le rapportent Cori & Léon (2002 : 25-26), des études très diverses eurent lieu sous couvert de la TA, alors qu'elles n'avaient en commun que l'automatisation des traitements : « dictionnaires électroniques fondés sur l'analyse morphologique, analyseurs syntaxiques, mise au point de langues intermédiaires, ébauches de mémoires de traduction, résolution d'ambiguïtés à l'aide de méthodes statistiques et traitement des unités lexicales composées ».

[31] Pour une étude plus détaillée sur l'histoire du TAL en France, voir les travaux de Léon (Cori & Léon, 2002; Léon, 1998, 2001, 2002).

1.2.3 La relation entre TAL et corpus

Le numérique, l'informatique et les différents outils du TAL furent et sont toujours d'un apport indéniable à la constitution, la consultation et l'exploitation de corpus modernes et, de fait, à la linguistique de corpus. Malgré ces liens, nous démontrerons tout au long de ce premier chapitre que le corpus n'est pas né avec l'informatique, car nous verrons que la notion de corpus a précédé l'informatique quel que soit le domaine, bien que la terminologie elle-même n'apparût pas avec la naissance des premières bases de données. La seconde question qui découle est de savoir si, de nos jours, un corpus contemporain peut se passer de la numérisation et de la machine informatique[32]. Mais quelque décisive que paraisse l'influence du TAL sur l'évolution de la linguistique de corpus, des corpus furent créés et exploités avant l'avènement de la machine informatique, et bien avant la démocratisation des ordinateurs personnels. Afin de démontrer ce propos, nous allons discuter ici d'un projet qui, à l'origine, n'employa pas les outils du TAL : le « Survey of English Usage » (SEU). Il nous faut d'emblée présenter ce corpus pour les raisons suivantes : c'est l'un des premiers corpus modernes, il n'était pas numérique à sa constitution et il a été exploité dans plusieurs branches de la linguistique. Son importance et son impact direct sur nombre de corpus qui suivront seront abordés également.

1.2.3.1 Le projet SEU, un corpus moderne non numérique

En 1959, Randolph Quirk, un linguiste anglais, fonde le SEU, qui est alors le premier centre de recherche sur corpus en Europe. Le SEU est établi au département de langue et littérature anglaises à l'University College London. Les motivations de Quirk trouvent leurs sources dans les travaux de Henry Sweet et de James Murray[33], ceci d'après Léon :

> Quirk se rattache à la lignée de l'empirisme britannique par l'importance qu'il donne à la « lexicographical syntax » d'Henry Sweet, à la recherche de patterns grammaticaux, et à la tradition des dictionnaires de l'anglais fondés sur l'usage de la fin du XIXe siècle comme le *National English Dictionary* de James Murray. (Léon, 2008 : 20)

[32] Il s'agit ici d'un critère définitoire des corpus et nous approfondirons donc la question en 2.5.

[33] Cf. 1.4.1.

Toujours d'après Léon (2008 : 21), ses rencontres avec Charles Fries[34] « à qui il doit l'idée d'utiliser des conversations enregistrées, totalement cohérente avec la tradition britannique favorisant l'étude du langage parlé », et Freeman Twaddell avec lequel il collaborera lors de la constitution du corpus Brown[35], amèneront Quirk à entreprendre le SEU sur un objectif principalement didactique, celui de constituer une grammaire d'enseignement de l'anglais, les années 1950 ayant été une époque où la demande en enseignement de l'anglais et du français fut très forte, en raison du contexte géopolitique d'après-guerre et de décolonisation. Le désir de Quirk était que la conception de cette grammaire fût constituée sur base de données attestées, qu'il nomma « a corpus of natural usage », « a body of full and objective data », « a copious body of actually recorded usage »[36], autrement dit de proposer une grammaire descriptive des usages réels de l'anglais. Plusieurs versions de cette grammaire paraîtront, dont *A grammar of contemporary English* (Quirk, Greenbaum, Leech, & Svartvik, 1972) et *A comprehensive grammar of the english language* (Quirk & Crystal, 1987). Pour constituer son corpus, Quirk entreprit la collecte des données : le projet initial (voir Quirk, 1960) devait réunir 200 échantillons écrits et oraux, de 5000 mots chacun pour un total d'un million de mots. Les données orales étaient alors enregistrées sur des bandes magnétiques, retranscrites sur papier et indexées sur des fiches. Les données sont de deux types : des données authentiques et attestées, mais aussi des données produites artificiellement en laboratoire. Le corpus fut ensuite enrichi d'annotations prosodiques et paralinguistiques dont les schémas sont détaillés par Crystal & Quirk (1964). La consultation du corpus nécessitait le déplacement au SEU, à Londres.

Bien que non numérisé à ses débuts, le SEU est un projet pionnier, annonciateur de la nature des corpus modernes. Il est en effet constitué selon des protocoles visant à en faire un corpus représentatif et équilibré. De plus, le SEU est un projet étant destiné à l'élaboration de grammaires de l'anglais, mais dont les données, collectées sur 30 ans, permettent également des études diachroniques. Nous avons évoqué les rencontres entre Quirk et l'équipe du Brown Corpus, ce dernier étant considéré comme le premier corpus numérique, et Léon (2005 : 40) rapporte qu'en 1963, Quirk était présent à la conférence durant laquelle les principales

[34] Cf. 1.6.2.

[35] Cf. 2.10.1.

[36] Cité par Léon (2008 : 21).

décisions de ce que serait le Brown Corpus, d'ailleurs construit sur le modèle du SEU, furent prises.

D'autres corpus importants s'inspirèrent ou dérivèrent du SEU : Le London-Lund Corpus of Spoken English (LLC) en 1975 ou le Lancaster-Oslo-Bergen Corpus of British English en 1978 (LOB). Ensuite, en 1988, Sidney Greenbaum[37] proposa le projet « International Corpus of English » (ICE), dont l'objectif est la compilation des corpus des diverses variétés d'anglais de par le monde. C'est à cette occasion que les données du SEU furent numérisées. Les échantillons linguistiques du corpus ICE sont tous construit sur le modèle des échantillons du SEU Corpus original : ainsi le ICE-GB, la composante britannique du ICE, contient à l'instar des autres corpus, un million de mots de l'écrit et de l'oral. La similarité des échantillons fait des corpus du ICE des corpus comparables qui permettent des études synchroniques, et le ICE-GB est annoté morphosyntaxiquement ainsi qu'arboré[38]. Comme nous l'avons mentionné, un nombre de chercheurs se sont intéressés à l'étude diachronique du langage, et un projet récent, se chargea d'arborer la partie orale du LLC, soit des données datant des années 1960 aux années 1980, selon les mêmes schémas utilisés pour la partie orale du ICE-GB. L'ensemble forme un corpus diachronique oral de 800 000 mots, nommé « Diachronic Corpus of Present-Day Spoken English » (DCPSE). Le DCPSE est disponible sur CD sur le site du SEU[39].

Contrairement au Brown qui ne comportera que des données écrites, le SEU ne défavorisa pas la langue parlée. Il est vrai qu'à l'heure du Brown, qui se voulait le premier corpus électronique, la numérisation du son n'était pas encore possible. C'est donc dans sa volonté de recouvrir l'ensemble du langage et dans ses échantillonnages à caractère scientifique que le SEU est considéré comme pionnier, qu'il inspira nombre de projets et qu'il est un centre encore en activité aujourd'hui. Le projet est sans doute l'exemple le plus flagrant que la linguistique de corpus s'est certes grandement développée grâce au TAL, mais ce dernier reste un outil de la linguistique de corpus.

[37] Directeur du SEU de 1983 jusqu'à son décès en 1996.

[38] Cf. 2.7.3 où seront discutées ces annotations.

[39] http://www.ucl.ac.uk/english-usage/projects/dcpse/index.htm

1.3 Corpus et acquisition du langage

Nous allons discuter dans cette section d'un autre champ d'application et d'exploitation des corpus qui émergea bien avant toute technologie, celui de l'acquisition du langage chez l'enfant ou ontogénèse linguistique. L'étude de l'ontogénèse est un des domaines précurseurs où les chercheurs s'intéressèrent à la langue orale pour des raisons évidentes : il va de soi que durant les premières années de l'enfant, la langue écrite n'est pas maîtrisée et que le seul moyen d'analyser la langue en cours d'acquisition et ses mécanismes ne peut se faire que par le biais de données orales de productions d'enfants[40]. Le plus ancien document[41] pouvant être apparenté à un corpus est le journal que tint Jean Héroard sur la vie quotidienne de Louis XIII, *Le journal d'un roi*, dès la naissance de ce dernier en 1601 et jusqu'à la mort de l'auteur en 1628. L'ouvrage était essentiellement tenu pour des raisons médicales, cependant Jean Héroard y consigna tous les événements de la vie du dauphin : ses horaires et habitudes détaillés, ainsi que tout épisode public ou privé. Ce « procès-verbal d'expérience »[42] rapporte également les productions langagières de l'enfant, transcrites (pour certains cas une transcription semi-phonétique) et parfois commentées[43]. Ce journal, unique en son genre pour son époque, constitue un trésor pour les historiens, les psychologues et les linguistes, et

[40] Pour davantage de détails concernant l'histoire des corpus oraux d'enfants et des sciences de l'acquisition du langage, voir *L'enfant dans la langue* (Morgenstern, 2009).

[41] À titre anecdotique, MacWhinney (1996) rapporte que Saint Augustin, en 397, s'intéressa à l'acquisition du langage dans *Les Confessions* où, dans un chapitre nommé « L'apprentissage de la parole », il affirma se souvenir du processus par lequel il apprit à parler : « Cela, je m'en souviens, et comment j'ai appris à parler, je l'ai remarqué par la suite. Ce n'étaient pas les grandes personnes qui me l'enseignaient en me fournissant des mots selon l'ordre déterminé d'une science, comme, peu après, pour les lettres. Mais c'est moi-même qui, grâce à l'esprit que tu m'as donné, mon Dieu, l'apprenais, lorsque par des gémissements et des cris variés, ainsi que par des mouvements variés de mes membres, je voulais mettre au jour les sentiments de mon cœur afin que l'on obéît à ma volonté, et que je ne pouvais ni exprimer tout ce que je voulais, ni l'exprimer à tous ceux que je voulais. Lorsqu'ils dénommaient quelque chose et mouvaient leur corps vers un objet en accord avec cette parole, je m'en emparais dans ma mémoire ; je voyais et je retenais que telle chose était appelée du nom qu'ils faisaient résonner lorsqu'ils voulaient la montrer. » (Augustin d'Hippone, 397, *Les Confessions, Livre I*, nous rapportons la traduction de Jean-Claude Fraisse).

[42] L'expression est empruntée à Madeleine Foisil, à qui l'on doit une nouvelle publication du journal de Héroard en 1989 (Héroard & Foisil, 1989).

[43] Des analyses détaillées sur la transcription des paroles du dauphin sont disponibles dans l'étude d'Ernst (Ernst, 1985) et la republication de Madeleine Foisil (Héroard & Foisil, 1989).

l'intérêt d'un tel corpus est tel que, quatre siècles après sa création, il est encore étudié et exploité[44]. Mais ce sont deux siècles plus tard que des savants s'intéressèrent de manière plus précise aux productions orales de leurs enfants, phénomène qui a été suivi par la constitution de larges corpus transversaux et longitudinaux tout au long du XX^{ème} siècle. Nous présenterons ces travaux et terminerons par la présentation de la situation actuelle en France.

1.3.1 Les premiers *baby books* ou *diary's note*

Au XIX^{ème} siècle, les expériences du type *baby books* ou *diary's note*, dans lesquels le chercheur tenait un journal sur le développement de son ou de ses enfants, s'inscrivent dans le courant des sciences naturelles et évolutionnistes. Charles Darwin en est le représentant le plus célèbre, et tint lui-même un journal sur son fils aîné et qui fera l'objet d'un article, « A biographical sketch of an infant » (Darwin, 1877), publié dans *Mind*. Il avait en cela été inspiré par Hyppolite Taine qui, un an plus tôt, avait publié « Note sur l'acquisition du langage chez les enfants et dans l'espèce humaine dans Revue Philosophique de la France et de l'étranger » (Taine, 1876). Après ces pionniers, nombreux prirent le relais, et nous citerons notamment les allemands Wilhelm Preyer et William Stern, pour les détails riches qu'ils nous ont laissés sur la constitution de leurs corpus

Preyer observe son fils quotidiennement et note avec précision toutes ses remarques. Il est le premier à transcrire phoniquement et de manière très détaillée les productions langagières de son fils, dont il fera un compte rendu dans « Die Seele des Kindes » (Preyer, 1884). L'étude couvre les productions langagières de l'enfant dès ses premières semaines et jusqu'à la fin de sa troisième année. Les travaux de Preyer restent néanmoins ceux d'un psychologue, qui s'est principalement intéressé aux processus cognitifs de l'acquisition du langage. Ce n'est que quelques années plus tard qu'une étude focalisée principalement sur l'acquisition du langage a eu lieu : William Stern et son épouse Clara ont tenu des journaux sur leurs trois enfants durant dix-huit ans, et ont publié « Die Kindersprache » (Stern & Stern, 1907), le premier journal entièrement consacré au langage de l'enfant. Nous rapportons ici quelques-unes des

[44] Voir « L'observation du langage d'un enfant royal au XVIIe siècle » (Gougenheim, 1931) et « A diary study on the acquisition of Middle French: A preliminary report of the early language acquisition of Louis XIII » (Ingram & Le Normand, 1996). Blanche-Benveniste (2010 : 55) cite également l'ouvrage, rapportant que « cet enfant sautait souvent le *ne* de négation, qu'il interrogeait avec *est-ce que* et qu'il omettait souvent certains subjonctifs ».

remarques faites par Morgenstern (2009) qui montrent l'importance et la qualité des travaux des Stern. Tous les deux faisaient la distinction entre observation et interprétation et autant que faire se peut, le couple tentait de tenir le journal sans que les enfants ne s'en rendissent compte en faisant déjà la distinction entre rapport d'événement et commentaire. En cela, leur travail met en relief le rapport, encore flou de nous jours, entre transcription et annotation. William Stern envisagea également très tôt que le « phonographe » et la « photographie » changeront sans doute la recherche sur le développement de l'enfant en permettant aux chercheurs de faire des enregistrements sans que l'enfant ne s'en rende compte.

1.3.2 Corpus transversaux et corpus longitudinaux

Nous avons constaté dans le paragraphe qui précède qu'il a existé deux types d'approches dans les corpus d'enfants. La première approche consiste à rechercher des universaux en collectant les données d'un grand nombre d'enfants, puis à les comparer entre elles ; le corpus constitué est alors de type transversal. Un corpus transversal est donc un corpus regroupant les productions de plusieurs enfants à un seul moment de leur développement, sans les suivre individuellement. La seconde approche consiste à étudier l'ontogénèse d'un enfant en particulier, afin de pouvoir suivre les mécanismes d'acquisition et leur évolution dans le temps ; le corpus constitué est alors un corpus de type longitudinal. Les corpus constitués furent alors auprès d'un nombre restreint d'enfants mais s'inscrivant dans la durée en enregistrant régulièrement les enfants participant à l'expérience.

Si les premières études de Taine et des Stern que nous avons présentées étaient européennes, les études transversales en acquisition du langage qui ont suivi ont été principalement américaines, entre 1926-1927 et jusqu'en 1957 (McEnery & Wilson, 2001 : 3). Les Américains (voir tableau ci-dessous) prirent le relais car ils considérèrent ces études européennes comme étant aléatoires, peu scientifiques, peu fiables et comme décrivant des enfants qui ne reflétaient pas un standard. Les travaux américains de la première moitié du XX^{ème} siècle ont donc voulu remédier à ces lacunes en adaptant de nouvelles méthodologies cherchant à octroyer une scientificité à ces corpus en instaurant des critères tels l'échantillonnage, la prise en compte de critères métalinguistiques (situation d'énonciation, sexe, milieu socio-économique, âge et enfants spécifiques comme les jumeaux ou les enfants doués) et l'homogénéité des données. Ces critères, comme nous le verrons tout au long du 2ème chapitre, feront partie des critères de constitution des corpus scientifiques modernes.

L'objectif formel des corpus transversaux est l'établissement de normes dans l'acquisition du langage grâce à de larges études quantitatives et comparatives. Voici un tableau regroupant les principales études transversales menées entre 1926 et 1957[45] :

Tableau 1 : Études transversales sur l'acquisition du langage entre 1926 et 1957

Auteur	**Date**	**Nombre d'enfants**	**Âge**	**Échantillons**	**Thème de recherche**
Smith	1926	124	2 à 5	1 heure de conversation	Longueur d'énoncé/ développement
McCarthy	1930	140	1;5 à 4;6	50 énoncés	Longueur d'énoncé/ développement
Day	1932	160	2 à 5	50 énoncés	Langage des jumeaux
Fisher	1934	72	1;6 à 4;6	3 heures par échantillon	Enfants doués
Davis	1937	173/166	5;6 à 6;6	50 énoncés	Jumeaux/ enfants uniques
Young	1941	74	2;6 à 5;5	6 heures de conversations	Classe sociale
Templin	1957	430	3 à 8	50 énoncés	Longueur d'énoncé/ développement

Comme nous pouvons le voir, la principale caractéristique des corpus transversaux est la collecte de données auprès d'un grand nombre d'enfants, selon un échantillonnage qui dépend d'un protocole scientifique (échantillons égaux en taille, critères de sélection des enfants à enregistrer).

La période des corpus transversaux dura jusqu'en 1957. Les chercheurs ont alors délaissé le souci quantitatif et comparatif pour s'intéresser à l'évolution dans le temps des compétences linguistiques d'un ou de plusieurs sujets. Les raisons principales du délaissement des études transversales au profit d'études longitudinales furent de deux ordres, l'un technique, l'autre théorique. Le premier fut la démocratisation du magnétophone[46] qui permit le suivi d'un enfant plus facilement qu'auparavant. Le second fut la parution de *Syntactic structures* (Chomsky, 1957) qui amena les chercheurs à orienter leur travaux sur la naissance de la syntaxe ; pour cela, ils eurent besoin de corpus longitudinaux leur permettent de suivre l'évolution de la syntaxe d'un seul et même enfant et donc, théoriquement, de comprendre le

[45] Ce tableau a été emprunté à Morgenstern (2009 : 194).

[46] Cependant, en raison de l'importance du non-verbal dans le domaine de l'acquisition du langage, la véritable avancée technologique dans ce domaine ne se fera qu'avec la démocratisation des enregistrements vidéo.

processus universel d'acquisition du langage. Les principaux projets entre 1963 et 1970 sont les suivants :

Tableau 2 : Principales études longitudinales sur l'acquisition du langage entre 1957 et 1973

Auteur	Année	Nombre d'enfants	Âge en début de projet	Durée de l'étude	Intermittence de recueil des données
Braine	1963	3	Entre 19 et 23 mois		Continue + 12 sessions de 4 heures
Miller & Ervin	1964	5	Entre 21 et 24 mois	2 ans	Deux à trois séances de quatre à cinq heures tous les deux mois.
Bloom	1970	3	Entre 21 et 27 mois		Deux heures toutes les deux semaines + une demi-heure toutes les semaines
Brown	1973	3	Premiers mots de l'enfant	Fin 3ème année	Hebdomadaire ou bimensuel

En comparaison avec le tableau précédant, nous constatons le nombre beaucoup plus restreint d'enfants enregistrés, mais le suivi des enfants sur des périodes relativement plus étendues par rapport aux données transversales. Ce type de collecte s'étala jusqu'à la moitié des années 1970 environ ; depuis, la situation est une synthèse des deux approches comme nous allons le voir dans le paragraphe qui suit.

1.3.3 Situation actuelle

Nous avons vu que les études transversales américaines avaient pour principales motivations le souci de rigueur ainsi que la recherche de scientificité des analyses et des résultats, éléments qui faisaient défaut dans les premières approches européennes. Les données représentaient les productions d'un grand nombre d'enfants sans avoir pu opérer de suivis longitudinaux en raison de l'absence des techniques appropriées entre 1927 et 1957. Les études longitudinales qui ont suivi à partir de 1957 ont délaissé la représentativité et l'échantillonnage en faveur du suivi du développement de l'enfant ; là encore, en raison de l'absence de moyens techniques et humains à mener des études longitudinales d'envergure. Les corpus uniquement transversaux ou uniquement longitudinaux ont montré leurs limites. Or les valeurs transversales ou longitudinales d'un corpus ne sont pas exclusives. Depuis les années 1970, les chercheurs désirent avoir à disposition des corpus à la fois transversaux et longitudinaux afin de pouvoir à la fois comprendre les processus d'acquisition en étudiant les corpus longitudinaux, mais aussi de vérifier leurs résultats, de les comparer et de les

compléter en ayant sous la main des données transversales, soit des études longitudinales effectuées sur un grand nombre d'enfants.

C'est dans ce contexte que, les moyens techniques aidant, le projet « Child Language Data Exchange System » dit CHILDES (Brian MacWhinney & Snow, 1985) vit le jour en 1984. Il s'agit de la première base de données orale numérique participative internationale de langue orale et les premières données à y être intégrées sont celles du projet Brown. Elle contient aujourd'hui un grand nombre de productions d'enfants collectées de par le monde, et représente un corpus longitudinal, transversal et multilingue. Notre corpus a été transcrit au moyen de l'un des outils de CHILDES, le logiciel CLAN, que nous présenterons en détail au 3ème chapitre. En outre, CLAN est le logiciel de transcription de l'un des corpus français les plus importants en ce qui concerne le développement du langage entre 1 et 3 ans, à savoir le corpus du projet ANR Colaje[47], qui regroupe des corpus audio-visuels, leurs transcriptions et des tests de langage. Les enfants sont filmés dans leurs familles et l'ensemble des documents est mis à disposition.

L'apport de l'étude de l'ontogénèse à la linguistique de corpus, en prenant en compte le potentiel des données à représenter des universaux, et donc en instaurant une base scientifique aux bases de données, est ainsi considérable au niveau méthodologique. Il faut également retenir que l'idée de suivi de cohorte est née dans ce domaine, et que les corpus longitudinaux sont aujourd'hui fréquents dans l'étude de l'interlangue.

1.4 Corpus et lexicographie

Outre l'acquisition du langage, la lexicographie est un autre domaine où les bases de données ont précédé l'informatique. En effet, la création et la consultation de corpus d'un côté, et la constitution de dictionnaires sont des disciplines qui s'entrecroisent et se retrouvent, car la

[47] http://colaje.risc.cnrs.fr
http://anr-leonard.ens-lsh.fr/

lexicographie est par définition un recensement qui nécessite inévitablement un corpus de travail, que celui-ci soit constitué[48] ou pas ; Geyken dit à ce propos :

> Un des domaines dans lesquels, dès le départ, les corpus électroniques ont joué un rôle important est l'élaboration de dictionnaires. À la différence de la linguistique générale, la lexicographie, dans sa méthodologie, a toujours été une discipline empirique dans le sens où elle s'est toujours appuyée sur des énoncés attestés. (2008 : 77)

En raison du souci d'exhaustivité des lexicographes, un des critères prédominants dans les corpus lexicographiques est la taille de ces corpus. La taille des corpus est généralement déterminée par le nombre de *tokens* (c'est-à-dire une chaîne de caractères entre deux blancs) ainsi que par le nombre de *types* ou *mots-types* (c'est-à-dire des *tokens* différents dans le corpus). Selon la loi de Zipf[49], la majorité des mots-types d'une base de données apparaît très rarement. Ceci est confirmé pour le corpus lexicographique par Sinclair (1991 : 18), Meyer (2002 : 14) ou Geyken (2008 : 81) : dans un corpus donné, un nombre relativement restreint d'occurrences apparaît un grand nombre de fois (*function words*), et un grand nombre d'occurrences apparaît de manière beaucoup plus limitée (*content words*). Pour représenter ceci, nous empruntons à Geyken (2008 : 81) le nombre de types du corpus DWDS-E, et le nombre d'occurrence de ceux-ci :

[48] Par constitution, nous entendons le respect des différents critères qui font d'une accumulation de données un corpus, comme la représentativité ou la documentation du corpus. Ces critères seront étudiés dans le second chapitre.

[49] George Zipf (1902-1950) fut un linguiste américain qui étudia les statistiques appliquées à la linguistique dans plusieurs langues. Il publie son ouvrage majeur en 1949, *Human Behavior and the Principle of Least Effort* (Zipf, 1949) dans lequel, en s'appuyant sur des faits statistiques, il démontre que la longueur d'un mot est très étroitement liée à la fréquence de son emploi : plus grande est cette dernière, plus bref est le mot. Bully (1969 : 24) justifie par exemple que « le langage courant tend à abréger cinématographe en cinéma, radiophonie en radio ou pneumatique en pneu » par la loi de Zipf, arguant que la fréquence de ces termes amène les locuteurs à vouloir les abréger. Zipf constate également que le premier mot dans la hiérarchie de fréquence revient en moyenne tous les dix mots ; le second revient tous les vingt mots, le troisième tous les trente mots et ainsi de suite. Ce qui est nommé de nos jours loi de Zipf est sujet à polémique : les résultats et la fréquence de distribution semblent trop précis pour être exacts. Nous ne discuterons pas de ceci ici, mais utiliserons les termes « loi de Zipf » dans le sens où la fréquence d'occurrence d'un mot dans une liste est liée à son rang dans l'ordre des fréquences. Pour davantage d'informations concernant les statistiques appliquées à la linguistique, il conviendra de consulter les ouvrages de référence de Oakes (1998) et de Baayen (2008). Nous mentionnons également que l'article de Bully (1969) est le premier à évoquer les travaux de Zipf en France.

Tableau 3 : Effectifs du corpus DWDS-E

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

La constitution de corpus dits représentatifs est donc une nécessité pour les lexicographes modernes. D'autre part, considérons une définition du corpus proposée par Ooi :

> A corpus, like a dictionary, is a mere snapshot of the language at a certain point of time, and therefore may need to be continually updated for changing and new patterns of usage. (Ooi, 1998 : 55)

Selon cette définition, les corpus destinés à des fins lexicographiques impliquent la notion de *monitor corpus* ou corpus de référence, soit un corpus qui se veut représentatif de l'ensemble du langage[50], avec les contraintes de taille et d'échantillonnage que cela implique, et que nous discuterons dans la section consacrée à la représentativité[51]. Ceci étant, la constitution des dictionnaires ne repose pas uniquement sur des exemples authentiques et attestés. Geyken (2008 : 82) rapporte qu'un grand nombre d'entrées des grands dictionnaires monolingues ne sont pas présentes dans tous les corpus, y compris des corpus unanimement considérés comme équilibrés et représentatifs, tel le British National Corpus[52].

D'après Perry *et al.* (2008), le document le plus ancien comportant une liste de mots explicités date de 2300 av. J.-C., et représente une collection de tablettes trouvées dans l'actuelle Ebla en Syrie, en sumérien traduits en akkadien. Il ne s'agit pas d'un dictionnaire mais d'un glossaire de type bilingue[53]. Le premier glossaire unilingue connu est le dictionnaire de

[50] Cf. 2.6.2.

[51] Cf. 2.6

[52] Cf. 2.9.2.

[53] Un glossaire est une œuvre que l'on nomme ainsi depuis le XVI^{ème} siècle et qui rassemble uniquement une liste non exhaustive de termes jugés difficiles. En cela, ils se rapprochent de ce que Niklas-Salminen (2005)

chinois Erya qui daterait du III[ème] siècle av. J.-C. (Karlgren, 1932). Au VIII[ème] siècle apr. J.-C., en langue arabe, *Kitāb al-ʿayn*[54] est le premier dictionnaire extensif et dont les occurrences sont classées[55] (Paoli, 2007). Les Grecs et les Romains ne possédèrent pas non plus de dictionnaires extensifs, mais des glossaires, des gloses ou des sommes des mots comme l'*Onomasticon* de Julius Pollux. Au Moyen Âge, plusieurs glossaires sont édités en France, en Allemagne, en Espagne et en Italie, tel le *Dictionarum* de l'érudit italien Ambrogio Calepino[56,57]. Mais c'est en Angleterre que fut ébauchée la méthodologie de constitution d'un dictionnaire à la fois extensif et sur base de corpus, comme nous allons le voir dans la section suivante.

1.4.1 Les corpus lexicographiques pré-informatiques

Au XVII[ème] et XVIII[ème] siècle, Francis (1991) note qu'un des principes des lexicographes était l'usage d'un corpus destiné à illustrer, par des citations, les polysémies et usages des items des dictionnaires. Leurs méthodes et cahiers de notes ne nous sont pas parvenus, et nous ignorons donc comment ces corpus étaient constitués ; nous ne savons pas s'ils étaient oraux, écrits ou documentés. Durant plusieurs siècles, le travail d'un lexicographe dépendait beaucoup de celui de ses prédécesseurs. Néanmoins, Samuel Johnson donne un premier exemple de méthodologie en publiant *Plan of an English Dictionary* en 1747, dont l'un des points fut la collecte d'un imposant corpus pour l'élaboration du *Dictionary of the English Language* en 1755, qui contient 40000 entrées illustrées par 150000 citations. Johnson notait les citations qu'il comptait employer « sur des morceaux de papier » (Francis, 1991), mais la

qualifie de « dictionnaires intensifs », en opposition aux « dictionnaires extensifs » qui visent à l'exhaustivité dans le répertoriage des occurrences.

[54] *Kitāb al-ʿayn* est attribué à Ḫalīl Ibn Aḥmad, mais Paoli (2007) met sa paternité en doute et discute des auteurs possibles.

[55] Dans *Kitāb al-ʿayn*, le classement des mots n'est pas organisé par ordre alphabétique mais par ordre phonétique, des sons les plus postérieurs (glottales, laryngales, pharyngales) aux plus antérieurs (labiales).

[56] L'ouvrage de Calepino était désigné du nom de son auteur : le calepin. Comme le seront plus tard le Larousse ou le Robert.

[57] Pour une épistémologie détaillée de l'histoire de la lexicographie en général, voir *The History of Lexicography* (Hartmann, 1986) ; pour une présentation détaillée de l'histoire des dictionnaires français, voir *Les Dictionnaires du français moderne, 1539-1863: étude sur leur histoire, leurs types et leurs méthodes* (Quemada, 1968).

taille et les classements du corpus nous sont inconnus[58]. Francis rapporte que le plan publié, élaboré en amont, n'échappe pas à ce qui advient aux protocoles modernes : il a évolué au fil du travail en fonction des contraintes et problématiques. La méthodologie de Johnson est considérée par Williams comme la naissance de la linguistique de corpus :

> 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. (Williams, 2006 : 152)

Le *Dictionary of the English Language* restera le dictionnaire de référence en langue anglaise jusqu'à la parution de l'*Oxford English Dictionary* (OED), cent cinquante ans plus tard. James Murray en est l'un des premiers éditeurs, de 1879 à sa mort[59], et il a basé son travail sur environ quatre millions de citations authentiques, collectées au fil du temps par des milliers de volontaires[60]. Le corpus de travail de l'OED est, depuis, en constante évolution, et est donc un corpus ouvert. Aux États-Unis, en 1934, George et Charles Merriam publient le *Webster's New International Dictionary, Second Edition*, qui résulte d'un travail de collecte des citations plus rigoureux, assuré par une équipe de lexicographes professionnels à plein temps. La collecte est « systématique », garantissant l'exhaustivité dans l'exploration des sources et des domaines. Toute la période depuis 1747 et jusqu'au dé but des années 1980 est caractérisée par la collecte et l'analyse manuelles des données. Selon Kilgarriff & Tugwell (2002), cette première période est à distinguer d'une seconde dans l'histoire de la lexicographie qui intervient avec la démocratisation de l'informatique.

1.4.2 Les corpus lexicographiques modernes

À partir des années 1980, les corpus des lexicographes sont stockés sur ordinateur. Mais ces corpus de travail en deviennent tellement grands que les lexicographes ont besoin d'outils pour les consulter. La coopération entre linguistes et informaticiens donne ainsi naissance aux premiers concordanciers. Les lexicographes pouvaient ainsi, pour la première fois, appréhender des exemples sans passer par le filtre d'un jugement arbitraire sur la pertinence de l'exemple, mais en se basant sur des résultats statistiques. Béjoint (2007) discute des

[58] Une des prouesses de Samuel Johnson fut de réaliser son dictionnaire seul, en neuf ans.

[59] James Murray décéda en 1915, avant la première parution de l'OED en 1929.

[60] Le recours à un tel nombre de volontaires est discuté en 2.7.1.

possibilités d'un corpus numérique analysé avec les outils appropriés, que nous résumons comme suit :

- évaluer la pertinence des éléments, ceci permettant d'exclure les formes trop rares et d'inclure celles dont l'indice de fréquence ;
- établir les usages syntagmatiques réels de par les études de collocations ;
- définir les éléments de manière plus précise, en résolvant ainsi la problématique soulevée par Teubert, et que nous avons évoquée plus haut ;
- catégoriser les éléments selon leur contexte, grâce aux concordanciers.

Ces développements ne permettent cependant pas la constitution d'un corpus unanimement reconnu comme représentatif, qui reste « un idéal visé par le lexicographe faute de pouvoir atteindre l'exhaustivité »[61] pour Béjoint, qui souligne également la sous-représentation de langue orale :

> La subjectivité reste donc présente dans la lexicographie moderne, au moins au moment de la constitution du corpus. L'une des « déformations » évidentes en lexicographie traditionnelle est la sous-représentation des données concernant la langue orale, pour diverses raisons, en particulier le fait qu'elles sont trop coûteuses à rassembler. (Béjoint, 2007 : 15)

Les avancées que permet le projet « Collins Birmingham University International Language Database » (COBUILD) sont toutefois substantielles. Le COBUILD est un projet de recherche qui a été dirigé par John Sinclair. Il a été initié en au début des années 1980 à l'université de Birmingham et son objectif était la constitution d'un corpus de référence pour la langue anglaise et d'exploiter ce corpus pour la création d'un dictionnaire pour apprenants uniquement basé sur l'analyse d'un corpus. La première concrétisation des objectifs du projet est la parution du dictionnaire *Collins COBUILD English Language Dictionary* en 1987. La constitution et l'exploitation du corpus sont détaillés dans « Looking Up: an account of the COBUILD Project in Lexical Computing » (Sinclair, 1990), où Sinclair explique que la conception du dictionnaire ne reposait sur aucun travail antérieur, et que ses deux caractéristiques principales étaient les suivantes : 1) le dictionnaire était entièrement basé sur un corpus numérique, et qu'en tant que dictionnaire pour apprenants ; 2) il reflétait la réalité du langage anglais en raison du fait que le corpus intégrait des données orales. En constituant

[61] L'idéal d'un corpus représentatif ne concerne pas uniquement, comme nous le verrons en 2.6, la lexicographie.

un dictionnaire entièrement basé sur corpus, Sinclair déforme le bon mot de Fillmore[62], en présentant son travail comme « une lexicographie de corpus » et non « une lexicographie de fauteuil ».

Le projet COBUILD nous intéresse car ses retombées sont vastes et variées. En effet, son exploitation ne se limita pas à la constitution du dictionnaire, mais un nombre important d'application découle du projet : le corpus fut exploité et analysé et donna lieu à des études linguistiques, des grammaires et des méthodes d'apprentissage[63]. En ce qui concerne la constitution des corpus, Williams (2006 : 155) note que la constitution du corpus du COBUILD a entraîné la réaction des autres éditeurs dont un consortium créa le British National Corpus (BNC)[64]. Le corpus COBUILD a en outre continué à évoluer jusqu'à nos jours, pour devenir l'actuel Bank of English (BoE)[65]. Nous remarquons ainsi que les besoins des lexicographes sont à l'origine de la constitution d'un grand nombre de corpus qui ont permis, au fil des ans, l'élaboration d'un ensemble de normes qui régissent la constitution des corpus modernes, notamment la prise en compte de la représentativité des données, l'intégration de données orales, la création des premiers corpus de référence et de la standardisation des données. La lexicographie basée sur corpus a également développé les outils et les théories des approches empiriques en ayant recours aux listes de fréquence qui donnaient une idée plus exacte sur les réels emplois des items. Si l'approche empirique n'est pas l'apanage des lexicographes, elle le fut en tout cas lorsque ces derniers constituèrent les dictionnaires pour apprenants du COBUILD, qui sont les premiers à entièrement être constitués d'exemples authentiques, alors que la tradition était dans les méthodes d'apprentissage aux exemples forgés. D'autre part, ce sont des phénomènes lexicaux comme

[62] Fillmore illustre la linguistique d'introspection comme une « linguistique de fauteuil », par rapport à la linguistique de corpus qui est une « linguistique de terrain » ; nous en discutons en 2.2.2.

[63] Nous noterons également que les dictionnaires entièrement basés sur corpus comme celui du projet COBUILD ont également profité à la traduction automatique. Par exemple, Pinkham & Smetst (2002), qui ont construit le système de traduction automatique MSR-MT, rapportent que les dictionnaires bilingues élaborés à la main n'amélioraient pas la traduction, et qu'ils pouvaient même nuire à sa qualité. Leur choix fut de n'utiliser ainsi que les dictionnaires basés sur corpus.

[64] Le BNC est un corpus dont la taille, l'annotation et la diffusion en XML ont fait de lui un corpus de référence en matière de normes dans la création de corpus. Il fera l'objet d'une présentation détaillée dans la section 2.9.2, notamment pour l'importance de sa version XML.

[65] Le corpus BoE est un corpus dont nous discutons en 2.10.1 en tant qu'exemple de la notion de « corpus de référence » d'une part, et de « corpus ouvert », soit enrichi avec le temps, d'autre part.

la polysémie ou l'homographie qui sont à l'origine de systèmes d'annotations complexes qui permettent des définitions plus précises. Les corpus lexicographiques ont ainsi permis des innovations méthodologiques en mettant l'accent sur l'authenticité des données et des innovations techniques en constituant des corpus vastes et annotés. Ce type de corpus fait d'ailleurs cruellement défaut en France, encore de nos jours. En effet, il n'y a pas toujours pas en France, plus de 25 ans après le COBUILD, de corpus similaires aux BoE et BNC anglais, c'est-à-dire intégrant des données écrites variées et non pas uniquement littéraires comme Frantext, ainsi que des données orales.

1.5 Corpus et études de la variation

Outre l'acquisition du langage chez les enfants et la lexicographie (qu'elle soit destinée à constituer des dictionnaires pour apprenants ou pour natifs), l'étude de la variation est un autre champ d'étude où la constitution de corpus, ou du moins la consultation d'exemples authentiques se sont imposées en tant que nécessité. En effet, quelles que soient les variations étudiées, le linguiste ne saurait baser ses recherches sur des exemples forgés par introspection, car le matériau de base de toute analyse de la variation est la performance et non la compétence du locuteur. En outre, l'étude des variables phonétiques implique l'analyse de données orales et donc la constitution de corpus oraux. Cette section se consacrera donc à la présentation de la place des corpus au sein de deux disciplines, la dialectologie et la sociolinguistique.

D'un point de vue théorique, ni le structuralisme, ni le générationalisme n'attribuèrent une place réelle à la préoccupation première de ces deux disciplines, à savoir l'étude des variations individuelles. Le structuralisme, de par sa conception de la langue comme un objet homogène et l'exclusion de la spécificité du locuteur, insista sur les relations qu'il peut y avoir entre les systèmes linguistiques d'une même langue, sans aborder leur hétérogénéité. Quant au générativisme, cette hétérogénéité n'en est pas une, puisque le générativisme propose un ensemble de règles qui régissent méthodiquement toute variation ; par ailleurs, la variation reste du domaine de la performance et non de la compétence qui est le domaine de prédilection de la linguistique générative.

Ainsi, jusqu'aux années 1960, l'étude des variations se faisait uniquement par le biais d'une sociolinguistique externe qui regroupait, selon Delais-Roussarie & Durand (Delais-Roussarie

& Durand, 2003 : 12), « les travaux qui prennent pour objet d'étude les rapports généraux qui existent entre langage et société », comme par exemple « la politique ou la planification linguistique ». Cette sociolinguistique ne se préoccupa donc que des relations entre les systèmes linguistiques en présence et les rapports de force qui les régissent. La sociolinguistique externe diffère de la sociolinguistique interne qui s'intéresse à la structure interne des systèmes linguistiques en plaçant au centre de ses intérêts l'analyse des variations individuelles des locuteurs, en se basant sur des facteurs sociaux et démographiques. Seul le facteur géographique est cité et étudié jusqu'aux travaux de Uriel Weinreich et William Labov, d'où la dialectologie. Ce sera donc Labov qui consolidera les bases théoriques de la modélisation de la variation, bases qui avaient été jetées par Uriel Weinreich qui décéda très jeune, mais dont l'influence sur Labov, et de fait sur la sociolinguistique est revendiquée par Labov lui-même :

> He died suddenly, of cancer, at the age of 39. Going through his papers in later years, I found that he had written up projects for research that anticipated most of the things I wanted to do. So to this day, I do not know how many of my ideas I brought to linguistics, and how many I got from Weinreich. (Labov, 2001a)

Maintenant que nous avons quelque peu souligné la nuance entre dialectologie et sociolinguistique, nous allons voir comment chacune de ces deux disciplines a abordé la notion de corpus, et surtout quand et comment l'oralité a été prise en compte dans les données étudiées. En outre, nous terminerons cette partie par une section concernant l'analyse conversationnelle, discipline qui se trouve dans ses démarches méthodologiques très proche des études de la variation.

1.5.1 Corpus et dialectologie

La dialectologie est une science qui nécessite la collecte de corpus conséquents mais qui, d'après Francis (1991 : 23), ne se développa qu'au début du XIX^{ème} siècle, vers 1820, soit en retard par rapport à la lexicographie : ceci dû au fait que les dialectes furent longtemps considérés comme des versions corrompues et perverties du langage standard. Francis lie l'intérêt pour la dialectologie à ses débuts en partie à l'intérêt accordé à la linguistique historique et la linguistique comparative d'une part, et sur le romantisme dominant de l'époque : l'attention se porte sur le « langage même des hommes », et les linguistes s'intéressent aux termes omis par les lexicographes pour les raisons citées plus haut.

En France, il y eut très tôt des enquêtes s'intéressant à la variation dialectale, comme l'enquête de l'Empire (1806 – 1812), mais dont la technique reste le questionnaire écrit[66]. L'approche par le biais de l'écrit des variations, même phonétiques, durera longtemps, comme en témoigne *La prononciation du français contemporain* (Martinet, 1945), dont l'enquête a été menée en adressant des questions écrites aux participants, en leur demandant de préciser comment ils prononçaient tel ou tel mot ; les déductions se faisaient sur la base des déclarations des participants, évidemment subjectives. Néanmoins dans *Introduction à l'étude des patois*, paru en 1887, Rousselot soulignait déjà l'importance de l'interaction par rapport au recueil de données écrites :

> Il y a plusieurs manières de recueillir les mots d'un patois. Toutes n'ont pas la même valeur. Mais souvent on n'a pas le choix. La meilleure, c'est le tête à tête avec des parents ou des amis. Grâce au laisser-aller de la conversation, on peut faire les observations les plus profondes, recueillir les faits les plus curieux, pénétrer dans les secrets de la syntaxe (…) On peut demander des traits de chronique locale, des contes, des dictons, des proverbes, le nom des objets que l'on a sous les yeux. (Rousselot, cité par Meunier-Crespo, 2008 : 9)

Ainsi, l'une des premières et plus célèbres enquêtes directes sur la variation qui recueillait des données orales fut celle du Suisse Jules Gilliéron, dont le collaborateur Edmond Edmont sillonna la France à bicyclette de 1897 à 1901. La collecte des données se déroula de la sorte :

> Le travail consistait à vérifier la forme lexicale et la prononciation d'une liste d'environ 1500 mots à partir d'un questionnaire régulièrement révisé en cours d'enquête. 737 témoins sont cités et Edmond leur posait surtout des questions directes (du type *comment appelle-t-on un* ***chat*** *par ici ?*), et transcrivait immédiatement la réponse. Ses interviews portaient sur un seul locuteur par localité. (Delais-Roussarie & Durand, 2003 : 18)

Les méthodologies d'Edmont ne correspondent évidemment pas encore aux standards actuels, et ne permettent pas d'aller dans l'étude au-delà du vocabulaire et de certains aspects de la prononciation. Par ailleurs, la transcription à la volée ne met pas l'enquêteur à l'abri de l'erreur et la représentativité sociolinguistique ou géographique de leur corpus ne fut pas

[66] Il était demandé aux participants d'écrire la parabole de l'enfant prodigue dans leur dialecte, ainsi que de fournir pour chaque idiome une chanson et un conte. Cette technique a été reprise en 1875 par Georg Wenker en Allemagne. Ce type d'approche s'est donc intéressé aux variations lexicales. Voir à ce sujet Delais-Roussarie & Durand (2003 : 17).

construite scientifiquement mais due au hasard des déplacements professionnels d'Edmont[67]. Néanmoins, les résultats obtenus ont permis l'élaboration de *l'Atlas linguistique de la France* (Gilliéron & Edmont, 1902), « un travail de pionniers qui reste un grand ouvrage de référence » (Delais-Roussarie & Durand, 2003 : 18). Chevalier & Encrevé évoquent également l'importance de l'*Atlas*, mais la mort de Gilliéron en 1926, laisse la place vide et la dialectologie en France s'en trouve retardée :

> La dialectologie enfin, elle aussi relativement autonome, est en plein recul. Gilliéron (mort en 1926) avait donné à la recherche française une autorité internationale avec l'Atlas linguistique, et avait placé ses résultats au cœur du débat théorique avec les néogrammainens. Ses élèves (Roques, Bruneau, Bloch) ne prendront pas la relève dans la nouvelle donne théorique, au contraire. (Chevalier & Encrevé, 1984 : 66)

En Angleterre, la dialectologie naquit à la même époque qu'en France. Ainsi la « English Dialect Society » est fondée en 1873, et le linguiste Joseph Wright publie *The English Dialect Dictionary* (1898) et *The English Dialect Grammar* (1905) ; la collecte des données pour ces deux ouvrages suit les méthodologies de James Murray dans la constitution du OED. La consultation du second ouvrage, *The English Dialect Grammar*, montre que les cinq premiers chapitres sont – malgré le titre – des études de phonétique ; seul le sixième et dernier chapitre s'intéresse à la grammaire des dialectes, approche néanmoins pionnière pour l'époque. Ce sixième chapitre comporte des titres tels que :

- les articles définis et indéfinis ;
- la formation du pluriel ;
- remarques générales sur les adjectifs ;
- les pronoms personnels, possessifs, réflexifs, démonstratifs, interrogatifs, relatifs ;
- les classifications des verbes ;
- les adverbes.

Cependant, le corpus de dialectologie d'importance de cette époque est celui d'Alexander J. Ellis, qu'il collecta pour la publication de *The Existing Phonology of English Dialects* (1889). La constitution de son corpus l'occupa durant vingt ans, avec l'aide de 811 personnes et la

[67] Edmont se déplaçait de foire en foire pour acheter des fromages. Dans l'atlas, il en résulte que les zones de haute densité de points d'enquête sont des zones fromagères, et que l'absence de production laitière s'est traduit par des vides géolinguistiques (Delais-Roussarie & Durand, 2003 : 18).

collecte de données en 1145 lieux géographiques différents en Angleterre et en Écosse (Francis, 1991 : 23). La constitution du corpus d'Ellis, à visée phonologique, dut surmonter plusieurs obstacles aujourd'hui résolus : l'inexistence d'appareils d'enregistrement, la disponibilité d'un alphabet phonétique consensuel et l'outil informatique pour le stockage et l'analyse des données. Aux États-Unis, la *Linguistic Society of America* crée en 1929 le « linguistic atlas », afin de permettre aux linguistes de poursuivre leurs travaux entre 1929 et 1939, pendant la Grande Dépression. Après la Seconde Guerre mondiale, De Fornel & Léon (2000 : 134) rapportent que dans le monde anglo-saxon, « les changements de la linguistique et les transformations de la société font apparaître la dialectologie rurale comme vieillotte et sans intérêt. Elle laisse alors la place aux données enregistrées en milieu urbain et à la sociolinguistique ».

L'évolution de la dialectologie d'une dialectologie rurale vers une dialectologie englobant les parlers urbains est tardive en France par rapport aux États-Unis, comme le montre le constat de Blanche-Benveniste & Jeanjean :

> Ce populaire délinquant, sauvé par la littérature, est opposé à l'image du peuple qui existe en dialectologie. Là, le peuple rural est paré de toutes les vertus conservatrices. Là l'étude de la langue ne se fait pas d'après les témoignages littéraires mais d'après les enquêtes, avec la caution scientifique de sociétés savantes (…). Mais pour ce qui n'est pas dialecte, pour le parler des villes par exemple, rien de tel (…). Jamais Bauche[68] ni les auteurs intéressés par le français parlé urbain n'ont trouvé de cercles linguistiques pour les accueillir. (Blanche-Benveniste & Jeanjean, 1987 : 15)

Ainsi, nous n'avons pas connaissance d'un corpus de dialectologie urbaine qui soit antérieur aux années 2000, et ce n'est que très récemment qu'a été constitué le « Corpus de Français Parlé Parisien des années 2000 » (CFPP2000), qui est un projet de l'équipe SYLED[69] à l'Université Sorbonne nouvelle, Paris 3. Sur le site du CFPP2000[70], le corpus est présenté ainsi :

> Il s'agit de la première étape d'un projet qui concerne les formes du français parlé dans l'agglomération parisienne et qui doit permettre d'aborder sur des bases solides la question du rôle que jouent les pratiques linguistiques de Paris pour la France entière et pour la francophonie en général. Des études portant sur des grandes villes comme

[68] Henri Bauche publia en 1920 *Le langage populaire* (Bauche, 1920).

[69] SYLED : **SY**stèmes **L**inguistiques, **E**nonciation et **D**iscours.

[70] Site du CFPP2000 : http://cfpp2000.univ-paris3.fr/index.html

> Londres ont en effet montré que les variétés langagières pratiquées dans les métropoles sont les moteurs du changement linguistique.

La collecte des données a été effectuée par le biais d'entretiens d'une durée moyenne d'une heure à une heure et demie. D'après les responsables du corpus, les enquêtés sont contactés à partir de réseaux de connaissances, et l'enregistrement de couples (ménages ou amis) a été privilégié afin de minimiser l'impact de la situation d'enregistrement. En mars 2012, le corpus comportait 535 000 mots environ, soit 37h75 pour 29 entretiens (dialogues ou multilogues). D'après Branca-Rosoff *et al.* (2009 : 3), « la collecte des données doit se prolonger dans les années à venir jusqu'à atteindre un million de mots ». Les données sont transcrites, documentées, annotées morphosyntaxiquement et disponibles sur le site du CFPP200.

Par ailleurs, mais toujours dans le domaine de la dialectologie en France, nous citerons le projet Corpus de la Parole[71], qui est un projet de numérisation des corpus oraux dialectologiques, sous la tutelle de la « Délégation générale à la langue française et aux langues de France » (DGLFLF), dans le cadre du Plan de numérisation du Ministère de la Culture en partenariat avec le CNRS. Le projet vise à la sauvegarde, la diffusion et l'enseignement du patrimoine oral. Le Corpus de la Parole est actuellement une plateforme regroupant des données orales de 10 dialectes parlés sur le territoire français, 7 langues non territoriales, 10 dialectes d'Outre-mer et des langues de Polynésie française, Wallis et Futuna, Nouvelle-Calédonie et Mayotte.

1.5.2 Corpus et sociolinguistique

Les liens entre corpus oraux et sociolinguistique sont exclusivement illustrés par les travaux de William Labov. C'est donc principalement ses travaux que nous présenterons ici. Labov est né en 1927, et l'importance de ses travaux se situe à deux niveaux. Le premier niveau est l'apport de Labov à la sociolinguistique, si ce n'est l'invention de celle-ci. Le second niveau est l'apport méthodologique de Labov qui proposa une approche empirique des données, et donc la nécessité d'une linguistique de terrain ; soit une linguistique de corpus qu'il mettra en application dans le cadre de ses propres travaux. Nous allons exposer brièvement ce qui a amené Labov à s'intéresser aux variantes linguistiques et quelle a été sa méthodologie dans le recueil des données.

[71] Site du Corpus de la Parole : http://corpusdelaparole.in2p3.fr/spip.php

En 1961, après avoir travaillé durant 11 ans dans le secteur de la chimie, Labov reprend ses études à l'université de Columbia en tant que doctorant. Sa thèse, « The social stratification of English in New York City Department Stores » (Labov, 1966), fut soutenue en 1963 sous la direction d'Uriel Weinreich. Nous présenterons également son étude « Language in the Inner City : studies in the black English vernacular » (Labov, 1972a).

La linguistique que propose Labov, qui deviendra la sociolinguistique, n'a pas pour objectif la construction, à l'instar de la grammaire générative alors en vogue, d'un système génératif de tous les énoncés de la langue. Son objectif n'est pas non plus, comme nous l'avons vu, l'étude des relations entre monde, société et politique d'un côté, et la langue d'un autre, soit le point de vue structuraliste d'une sociolinguistique externe. Labov préconise une démarche empirique qui analyse des donnés authentiques en vue d'étudier les variations individuelles des locuteurs. Cette démarche ne laisse pas de place à l'introspection, et ouvre la voie aux corpus. À ce sujet, Labov dira en 2009[72] :

> Most of the linguists I met were gathering data by introspection, asking themselves, "Can I say this?" and, "Can I say that?" It occurred to me that I might start a new way of doing linguistics by building the study of language on what people actually said in everyday life.

Comme mentionné, « The Social Stratification of English in New York City Department Stores » est une étude que mena Labov dans le cadre de sa thèse de doctorat, et qui porta sur la prononciation du /r/ à New York. L'hypothèse de départ est posée ainsi : « If any two subgroups of New York City speakers are ranked in a scale of social stratification, then they will be ranked in the same order by their differential use of (r) » (Labov, 1972 : 44). La première étape est donc de rendre compte des classements sociaux, et pour ce faire Labov choisit de mener l'enquête auprès des employés de trois grands magasins new-yorkais, Saks Fifth Avenue, Macy's et S. Klein, dont les clientèles étaient respectivement catégorisées comme luxueuse, de classe moyenne et de classe populaire. À ce propos, Labov poursuit qu'il aurait pu mener son enquête sur trois catégories socioprofessionnelles très distinctes : un groupe d'avocats, un groupe de commis aux dossiers et un groupe de concierges. Cela ne se

[72] In « A Life of Learning: Six People I Have Learned From », communication orale à l'occasion du « Charles Homer Haskins Prize Lecture » (2009), disponible en ligne à l'adresse suivante : http://www.acls.org/publications/audio/labov/default.aspx?id=4462

fit pas car Labov pensait, et voulait démontrer qu'il n'était pas nécessaire de recourir à une différenciation aussi extrême pour prouver son hypothèse :

> Such an extreme example of differentiation would not provide a very exacting test of the hypothesis. It should be possible to show that the hypothesis is so general, and the differential use of (r) pervades New York City so thoroughly, that fine social differences will be reflected in the index as well as gross ones.

Afin d'obtenir des résultats plus subtils, Labov mène donc l'enquête auprès des employés des magasins susnommés, en se basant sur deux hypothèses :

> 1) il y a corrélation entre le niveau social des employés et le niveau social des clientèles des trois magasins ;
> 2) il y a corrélation entre le niveau social d'un poste de travail et le comportement social de la personne qui tient ce poste. Labov reprend une hypothèse de C. Wright Mills qui stipule que : « salegirls in large department stores tend to borrow prestige from their customers, or at least make an effort in that direction » (Labov, 1972 : 45)(les vendeuses des grands magasins tendent à s'approprier du prestige de leur clientèle, ou du moins elles s'efforcent à le faire).

Pour réaliser son enquête, la méthode appliquée par Labov est celle de « l'interview rapide et anonyme » (Delais-Roussarie & Durand, 2003 : 24) ; l'enquêteur (Labov lui-même) se rend dans un des magasins et se fait passer pour un client. Il aborde un employé et lui demande l'emplacement d'un article en particulier, de sorte que l'article soit spécifiquement au quatrième étage « fourth flour », afin que la réponse comportât deux réalisations du /r/. Là, Labov faisait mine de ne pas avoir compris la réponse, pour que l'employé répétât la réponse « spoken in careful style under emphatic stress » (Labov, 1972b : 49) (articulée clairement avec un stress emphatique). Labov notait alors, à l'insu des employés, les informations suivantes : le magasin, l'étage, le sexe de l'employé, son âge (estimé par tranches de cinq ans), le poste occupé, sa race et, éventuellement, son accent. Ces informations constituèrent donc une documentation de corpus constituante de la représentativité recherchée[73]. En répétant la manœuvre autant de fois que possible dans chaque étage, puis en passant à l'étage suivant, Labov obtint des informations de 68 locuteurs à Saks, 125 à Macy's et 71 à Klein. Il estime la durée totale des interviews des 264 locuteurs à 6h30. Les résultats de l'enquête

[73] La relation entre représentativité et documentation d'un corpus sera discutée en 2.6.4.

(Labov, 1966, 1972) confirmèrent l'hypothèse initiale. L'intérêt de la méthodologie de Labov se résume en plusieurs points ; les enquêtés sont dans leur lieu de travail, et la rapidité de l'interview minimalise l'intrusion de l'enquêteur, tout en maximisant l'authenticité de l'échange. Néanmoins, mais cela reste compréhensible pour l'époque, l'absence d'enregistreurs limite qualitativement les données, et ouvre la voie à des biais d'interprétations. Labov fera une autre étude sur la stratification du /r/ en utilisant cette fois des enregistreurs[74].

Plus tard et concerné par les données réelles et la spécificité presque individuelle des variations, Labov mènera une étude sur le langage des jeunes afro-américains de Harlem, nommée *Language in the Inner City : studies in the black English vernacular* (Labov, 1972a). Sur les raisons théoriques qui l'animèrent tout au long de sa carrière, Labov dira 30 ans plus tard :

> Linguists wanted to describe languages, like English or French, but their methods only brought them in contact with a few individuals, mostly highly educated. Whenever someone raised a question about the data, they would answer, "I'm talking about my dialect." The current theories held that every individual had a different system, and they weren't making much progress in describing the English language. (Labov, 2001a)

Les raisons pratiques furent autres : constatant l'échec régulier et systématique des enfants afro-américains (majoritairement défavorisés dans l'Harlem de l'époque) dans leur apprentissage de la lecture, Labov propose à l'Office of Education un projet de recherche visant à étudier leur dialecte, afin d'évaluer son rôle dans l'analphabétisme des jeunes afro-américains. Dans le cadre de notre étude, l'intérêt de ce projet réside dans la méthodologie de recueil des données : Labov considère que la langue d'un jeune locuteur pris isolément dans le cadre d'une enquête ne sera pas celle qu'il pratique au quotidien, il opte alors pour le recueil au sein de groupes réels, où le groupe influence directement le style de discours. D'autre part, Labov stipule que la présence d'enquêteurs blancs pourrait influencer le naturel des échanges et demande donc la collaboration de Clarence Robins et John Lewis, deux chercheurs noirs qui procédèrent au recueil des données. Labov et Paul Cohen, un autre chercheur blanc, se chargèrent de leur interprétation. La collecte se fait sur le terrain, de par des entretiens

[74] Auprès de 122 personnes nées à New York et domiciliées dans le Lower East Side. Voir Labov (1966) ou Delais-Roussarie & Durand (2003 : 27)

individuels, des face à face, des sorties collectives où chaque participant est enregistré sur une piste séparée grâce à un micro-cravate, tandis qu'un micro principal capte l'ensemble des échanges.

L'ensemble de travaux de Labov ont ainsi permis de démontrer que des facteurs métalinguistiques comme l'âge, le sexe, les origines socioprofessionnelles ou géographiques et les situations de communication influaient sur les usages langagiers ; en outre, les méthodologies de recueil des données orales par Labov sont révolutionnaires, au niveau des objectifs mais aussi au niveau des démarches entreprises.

À un autre niveau, les travaux de Labov inspirèrent la constitution du corpus de Sankoff-Cedergren. Constitué de 120 interviews effectuées en 1971 à Montréal, dans les secteurs à prédominance francophone (plus de 64 % de francophones), il s'agit du premier corpus oral francophone informatisé (Sankoff *et al.*, 1977 : 186). Tous les informateurs sont originaires de Montréal ou y ont vécu au moins depuis l'âge de six ans. Les informateurs et informatrices, 60 hommes et 60 femmes, ont été sélectionnés de façon à fournir une image la plus fidèle possible des différentes strates socio-économiques de cette population : il a été déterminé 6 catégories de revenus et l'on a choisi 20 informateurs pour chacune de ces catégories. L'âge et la profession du « chef de famille » sont renseignés. En ce qui concerne le recueil des données, les entrevues ont pris la forme de conversations informelles où l'enquêteur tentait d'amener l'informateur à s'exprimer sur certains thèmes déterminés. Les thèmes couverts sont la vie et les coutumes au Québec dans le passé, la vie moderne à Montréal et les opinions de l'informateur sur la langue. Les transcriptions ont été effectuées en français standard, en prenant en compte les amorces, les pauses, les hésitations. Le corpus représente finalement 100 à 120 heures d'enregistrement sur 120 bobines et leurs copies de sauvegarde, ainsi que 100 000 cartes perforées.

1.5.2.1 Corpus sociolinguistiques actuels en France

Nous précisons bien qu'il s'agit ici de corpus à visée sociolinguistique constitués en France, et non pas de corpus francophones, car les corpus sociolinguistiques en France sont en retard par rapport à ceux constitués en français mais hors de France, comme nous venons de le voir pour le corpus Sankoff-Cedergren, constitué au Canada. Blanche-Benveniste & Jeanjean (1987 : 85) parlent de « la curieuse aliénation du milieu linguistique français » et Cappeau & Gadet évoquent la francophonie hors de France en ces termes :

> Elle s'avère un lieu où la constitution de corpus est une tradition plus ancienne et plus solidement ancrée que dans l'Hexagone, sans doute parce qu'il y est vite apparu qu'on ne pouvait se contenter des intuitions pour travailler sur ces zones. (Cappeau & Gadet, 2007b : 131)

En parlant des corpus hexagonaux et de leur diffusion, les auteurs ajoutent qu'en France, les corpus ont rarement été commandités par les pouvoirs publics, comme cela avait été le cas pour le corpus de Sankoff-Cedergren.

Ainsi, pour évoquer les corpus oraux sociolinguistiques en France, nous présenterons L'Enquête Socio-Linguistique à Orléans (ESLO1), qui date de 1966. L'initiative découle des faits suivants : un nombre de linguistes anglais travaillant sur l'enseignement du FLE, voulurent profiter des nouvelles techniques de l'époque (magnétophones et laboratoires de langue) pour moderniser leurs méthodes. Ils désiraient ne plus s'en tenir au français officiel et littéraire des manuels de l'époque, mais enseigner la langue française telle qu'elle était parlée au quotidien, avec tout ce qu'elle comporte de variations. À ce moment-là, il n'y avait aucun corpus oral réellement constitué en France, car la France prenait rarement l'initiative ; citant ESLO1, Bergounioux *et al.* disent à propos de l'initiative anglaise :

> Il est symptomatique que l'initiative de cette enquête ne provienne pas d'une équipe française — ses promoteurs sont des enseignants anglais ; il l'est plus encore que son exploitation ait été le fait d'universitaires anglais, allemands, néerlandais et belges (…). Ce paradoxe n'est pas sans exemple dans l'histoire de la linguistique et déjà, au XIXe, seule la pression de la concurrence allemande avait permis qu'une romanistique acquière droit de cité dans l'université française pour y assurer un traitement « national » des textes et documents médiévaux. De même, le « Français Fondamental ». (Bergounioux *et al.*, 1992 : 74)

Le projet est donc lancé[75] et parmi son intérêt dans le cadre de la présente étude est multiple et réside dans une partie de ses objectifs : il est pionnier dans sa volonté de constituer un corpus cohérent et documenté d'enregistrements oraux. Les données furent collectées en cinq semaines en 1969. Elles représentaient environ 300 heures, ou 4 500 000 mots, provenant de

[75] En ce qui concerne le financement et le parrainage du projet (Bergounioux, Baraduc, & Dumont, 1992: 76) : « Le travail a bénéficié de divers appuis, scientifiques ou matériels, parmi lesquels le Ministère (britannique) de l'Éducation et de la recherche, l'Ambassade de France à Londres et le B.E.L.C., auxquels il convient d'ajouter le Centre de Sociologie Européenne et, à Orléans, la Mairie, l'I.N.S.E.E., le Centre médico-psycho-pédagogique (C.M.P.P.) et le Centre régional de documentation pédagogique (C.R.D.P.) ».

l'interview d'environ 200 locuteurs. Les données furent documentées (caractérisations sociologiques des témoins, identification de l'enquêteur, date et lieu de passation de l'entretien) mais n'étaient pas homogènes ; un questionnaire de 26 questions concernant les paramètres sociolinguistiques était rempli par les participants, puis un questionnaire de 37 questions sur le rapport de l'interviewé aux pratiques du langage (lecture, écriture, jugements normatifs...). Bergounioux *et al.* (1992 : 86) rapportent que « posé en fin d'interview, ce questionnaire a été souvent réduit, ou est demeuré incomplet. Les réponses semblent avoir, en général, déçu ses initiateurs. » D'autre part, 84 enregistrements divers ont été effectués avec des témoins inconnus, selon la technique du « micro-caché » (dont 48 pour des achats) ; aujourd'hui cela ne serait pas possible pour raisons juridiques.

L'équipe de l'ESLO n'a pu assurer la transcription de l'intégralité des bandes. D'autre part, les bandes ont subi des détériorations, faute d'une conservation efficace. Le projet est repris en 1993 sous le nom de ESLO2, afin de reconstruire le corpus et de le numériser. Le projet ESLO2 fut initié par le « Centre orléanais de recherche en anthropologie et linguistique » (Coral), qui réussit à récupérer en 1993 l'ensemble des documents originaux composés des bandes magnétiques, d'un catalogue dactylographié, de quelques centaines de feuillets de transcription manuscrites (d'une qualité inégale) et des fiches d'identification des locuteurs. L'opération de numérisation s'est avérée une véritable reconstruction du corpus. Les documents sonores ont été regroupés et complétés, puis numérisés à partir des enregistrements L'étape suivante a consisté à transcrire et à baliser l'intégralité du corpus. Les données d'ESLO2 sont aujourd'hui :

- 157 entretiens face à face (± 182h30).
- 79 enregistrements dans des situations sociales ou professionnelles « informelles » (± 27h).
- 51 communications téléphoniques (± 2h10).
- 46 interviews « sur mesure » avec des personnalités (± 47h).
- 29 conférences-débats ou discussions avec plusieurs participants (± 32h).
- 84 enregistrements divers avec des témoins anonymes (± 14h30).
- 41 entretiens au Centre Médico-Psychopédagogique (± 10h).

Bien que les motivations de la constitution d'ESLO1 fussent pour l'amélioration des méthodes de FLE de l'époque, le corpus qui fut constitué est de nature sociolinguistique de par les données collectées, mais surtout en raison de l'exploitation qui peut en être faite de nos

jours, principalement en tant que corpus de référence de l'oral des années 1960-1970 en France.

1.5.3 Corpus et analyse conversationnelle

Suite à ce que nous avons exposé sur l'étude de la variation en général, soit de la dialectologie et de la sociolinguistique en particulier, il nous semble nécessaire de consacrer une section à un domaine qui leur fut proche de par les méthodologies adoptées ainsi que de par l'époque et le lieu où il est né, à savoir celui de l'analyse conversationnelle. Ce domaine nous intéresse pour les méthodologies qu'il employa pour la collecte des données, car il va de soi qu'une analyse conversationnelle repose essentiellement sur la constitution et l'analyse de corpus oraux. Nous discuterons des origines de l'analyse conversationnelle et de ses tenants théoriques, et nous terminerons par un examen de la situation de cette discipline en France.

La linguistique conversationnelle est un champ d'analyse ayant émergé aux États-Unis avec les travaux de Harvey Sacks. Né en 1935, il est juriste et politologue de formation mais sa rencontre avec le sociologue Harold Garfinkel l'amène à s'intéresser à l'anthropologie et à l'ethnométhodologie (De Fornel & Léon, 2000 : 133). En 1963, il est assistant au centre d'études scientifique du suicide à Los Angeles. En travaillant sur les bandes enregistrées et les transcriptions sténographiques des appels téléphoniques, il commence ses premières analyses de conversation ; il s'intéresse par exemple à la manière d'engager la conversation téléphonique des appeleurs et aux méthodes qu'ils utilisent afin de rester anonymes. De Fornel & Léon décrivent son apport méthodologique à l'étude de la parole de la sorte :

> Cet intérêt pour cette réalisation méthodique et reproductible des activités ordinaires marque le début de l'analyse de conversation. Le matériel enregistré, non manipulé et appréhendé dans tous ses détails, devient intéressant comme ressource essentielle pour ce qui est en train d'être accompli dans et par la parole. La parole est appréhendée comme une activité en soi. (De Fornel & Léon, 2000 : 133)

Sacks enregistre d'ailleurs les cours qu'il dispense à l'University of Californy de Los Angeles (UCLA), de 1965 à 1972. Ce corpus est lu et étudié pendant vingt ans, puis publié par Emanuel Schegloff (Harvey Sacks, 1995), un des collaborateurs de Sacks et cofondateur de l'analyse conversationnelle. Sacks disparaît prématurément en 1975, à l'âge de quarante ans, mais son apport à l'analyse de conversation est affilié au développement de disciplines telles que la linguistique interactionnelle, l'ethnographie de communication ou la sociolinguistique,

encore émergentes durant les années soixante et ayant largement recours aux enregistrements audio et/ou vidéo[76]. En ce qui concerne les filiations de l'analyse conversationnelle, De Fornel & Léon (2000 : 134) voient dans cette discipline une continuité de l'intérêt alors en vogue pour l'étude des données enregistrées ; intérêt qui, comme nous l'avons vu, s'était principalement développé dans le cadre de la dialectologie et de la sociolinguistique. Comme le note Mondada (2001 : 142), cette époque était aussi celle des premières grammaires de l'oral et des premiers grands corpus de données orales authentiques (nous rappelons, concernant ces deux points, les travaux de Quirk).

L'apport de la linguistique conversationnelle à la linguistique de corpus se situe dans ses enquêtes de terrain, en raison de l'exigence de travailler sur des données collectées dans leur contexte social d'énonciation. L'inefficacité du recours à l'introspection ou à l'élicitation est liée d'après Mondada (2001 : 144) à deux assomptions : 1) l'interaction sociale est « le lieu prototypique de l'usage des ressources linguistiques » 2) la relation entre le contexte social et les productions linguistiques est bilatérale : les pratiques langagières structurent et sont structurées par la situation d'énonciation. Mondada poursuit de la sorte :

> Pour cela, la linguistique interactionnelle et l'analyse conversationnelle d'inspiration ethnométhodologique pratiquent une démarche de terrain qui fait recours à une intégration du chercheur dans les groupes observés, à une invasivité minimale dans les activités enregistrées, à l'auto-enregistrement de la part des acteurs, à l'utilisation de dispositifs d'enregistrement prévus par les acteurs eux-mêmes pour leurs propres fins pratiques. (Mondada, 2001 : 144)

Ainsi l'analyse conversationnelle a permis la constitution de bases de données qui, par rapport aux données d'introspection ou obtenues grâce à des questionnaires, contiennent des données orales non décontextualisées. La langue est alors considérée comme une pratique sociale, dont les productions en soi sont indissociables des situations d'énonciation. Un autre apport de la linguistique conversationnelle fut que la constitution des bases de données amena les linguistes travaillant sur ces corpus à être confrontés aux problématiques liées à la transcription des données sonores ; ce fut Gail Jefferson (1938-2008), collaboratrice de Sacks

[76] L'article de De Fornel & Léon (2000) détaille l'histoire de l'analyse de conversation et de ses relations et apports aux autre champs linguistiques comme la sociolinguistique ou la grammaire générative. L'article expose également les principaux concepts et méthodes de l'analyse conversationnelle.

et de Schegloff, qui élabora un ensemble de conventions pour la transcription des conversations, connue sous le nom de « Jefferson System »[77].

En ce qui concerne la France, Balthasar & Bert (2005 : 2) affirment que ce sont les structures dont est issu l'actuel laboratoire ICAR[78] qui introduisirent les approches interactionnistes en France vers 1975 et qui surtout favorisèrent un contexte institutionnel qui encouragea la constitution d'un grand nombre de corpus audiovisuels naturels et expérimentaux. Ces corpus dispersés jusqu'à la fin des années 1990 ont été regroupés au sein de la base de données CLAPI[79]. Les corpus ont été numérisés afin de garantir leur pérennité et leur diffusion. Les données sont présentées sur le site comme suit :

> La plateforme Corpus de langues parlées en interaction (CLAPI) du laboratoire ICAR est un environnement d'archivage et d'analyse de corpus d'interactions enregistrées en situation authentique (en famille et entre amis, sur le lieu de travail, dans les institutions les plus diverses, Interactions professionnelles, institutionnelles ou privées, commerciales, didactiques, médicales etc.). 46 corpus, 350 enregistrements (148 h), 600 transcriptions documentés par 75 descripteurs, 47 h balisées pour des traitements d'analyses et de requêtes. Presque toute la base est accessible gratuitement.

Les données orales de CLAPI ont été majoritairement transcrites selon les conventions de transcription ICOR, que nous discuterons dans la section consacrée à la transcription citée plus haut.

1.6 L'oralité dans les corpus généralistes ou de référence

Dans la section présente, nous nous proposons de voir quand et comment la langue orale a été prise en compte non pas dans des corpus spécifiques tels que ceux que nous avons présentés dans les sections précédentes, mais dans des perspectives généralistes destinés à la constitution de corpus de référence censés représenter l'ensemble du langage[80]. Ce type de

[77] Nous reviendrons sur le « Jefferson System » lorsque nous discuterons en détail de la transcription en 2.7.2.

[78] ICAR : Interactions, Corpus, Apprentissages, Représentations. UMR 5191 CNRS.

[79] CLAPI : Corpus de LAngue Parlée en Interaction. L'adresse de la base est la suivante : http://clapi.univ-lyon2.fr/

[80] Par « corpus spécifique », nous considérons ici que la spécificité s'exprime dans les objectifs du corpus. Ainsi les corpus constitués par les lexicographes sont spécifiques dans le sens où ces corpus serviront à la construction de dictionnaires. Quand bien même ces corpus seraient exploités à d'autres fins, nous ne considérons ici que les objectifs primaires et affichés des corpus que nous discutons.

corpus, les corpus de référence, servent soit à l'élaboration de grammaires généralistes, soit à l'étude de phénomènes linguistiques particuliers, mais qui ne sont pas prédéfinis lors de la constitution du corpus lui-même.

En premier lieu, nous voulons revenir plus en détail sur le retard dans le développement de corpus oraux et des études de la langue orale. Il faut préciser que toute notion de retard implique la présence d'un point de repère comparatif qui est celui des corpus écrits. La comparaison avec l'écrit n'est pas uniquement une question de référentiel, l'écrit et l'oral entretenant depuis toujours une relation dialectique. L'oral a cependant été longtemps, comme nous l'avons vu dans l'introduction de ce chapitre, été considéré comme inférieur à l'écrit. Nous avons également discuté de la possibilité de conservation des données, et avons vu que la conservation de l'écrit est connue depuis l'antiquité alors que les modes de conservation du son n'existent que depuis à peu près un siècle. Les deux principaux obstacles à la constitution de grands corpus oraux furent donc, historiquement, la non reconnaissance de l'oral en tant que sujet d'étude au moins aussi légitime que l'étude de l'écrit d'une part, et les possibilités technologiques d'autre part. Si ces deux obstacles peuvent paraître aujourd'hui résolus, ils obstruent pourtant encore l'étude de la langue parlée et la disponibilité de corpus de référence ; nous allons considérer ces deux points pour étayer notre propos.

En ce qui concerne le statut accordé à la langue orale par les théoriciens, force est de constater que la réflexion accordée sur la linguistique des corpus oraux dans les ouvrages de référence récents reste lacunaire par rapport à celle accordée aux données textuelles, comme le note Kennedy (1998 : 20), qui constate que la majorité des études grammaticales et lexicales en anglais est basée sur des corpus textuels, et ceci bien que « la langue parlée soit de loin le moyen de communication le plus répandus ». Kennedy impute les raisons de cette situation aux coûts matériels et temporels à mettre en place pour la constitution de corpus oraux par rapport aux corpus textuels, et les raisons avancés par Habert *et al.* sont similaires, quand les auteurs justifient pourquoi leur ouvrage n'aborde que succinctement la question des corpus oraux :

> Nous avons mis l'accent sur les corpus relevant de l'écrit. Les corpus d'oral transcrits sont encore rares : la transcription proprement dite, les choix qu'elle entraîne, les coûts qu'elle suppose freinent leur développement. (Benoît Habert, Nazarenko, & Salem, 1997 : 13)

Mais il nous semble que les limitations d'ordre matériel ne suffisent pas à justifier la situation, et que les motivations de la majorité des autres linguistes ont été déterminantes. Par exemple, Biber (1993a : 247) reconnaît que la langue parlée représenterait près de 90% de la langue en général, mais qu'il faut privilégier dans la constitution de corpus les données textuelles[81]. La préférence aux données textuelles est également exprimée par Teubert (2010 : 6) qui définit le corpus comme « un ensemble de textes » ou, précise-t-il, « de fragments de textes », sans qu'il ne soit question de langue parlée. Tognini-Bonelli fait de même dans son ouvrage *Corpus Linguistics at Work* (2001) qui traite essentiellement des corpus textuels, et la place accordée à l'oral dans d'autres ouvrages de référence sur la linguistique de corpus est elle aussi marginale[82]. Nous retrouvons cette tendance chez certains auteurs en France, comme Rastier (2005) qui définit un corpus comme « un regroupement structuré de textes intégraux », ou Williams (2005 : 13), qui affirme que « la linguistique de corpus est un domaine qui s'intéresse aux textes, aux textes réels ». Ainsi, la théorisation des problématiques liées aux corpus oraux se retrouve-t-elle très en retard par rapport aux outils et méthodologies proposés pour l'analyse textuelle. Si une minorité des auteurs mirent en avant les obstacles d'ordre matériel, la plupart n'évoquent pas ces raisons et l'orientation vers l'analyse de corpus essentiellement textuels révèle une vision « textuelle » de la langue ; bien que nous ne lisions plus de nos jours des allégations envers la langue orale aussi dépréciatives que celles qui ont pu être écrites jadis, elle n'est toujours pas traitée équitablement à la langue écrite comme le montrent les exemples des manuels que nous venons de présenter.

Revenons-en maintenant au second point ayant ralenti la constitution de corpus oraux, celui des limitations d'ordre technologique. Les avancées ont été substantielles depuis les années 1980, avec la disponibilité de magnétophones puis d'enregistreurs numériques portables, ainsi que de la démocratisation des ordinateurs. Cependant, cet ensemble d'outils ne suffit pas à rendre les données orales aussi accessibles que les données écrites, et les investissements matériels, temporels et humains à investir pour la constitution d'un corpus restent significatifs sinon rédhibitoires, qui plus est lorsqu'il est question de corpus généralistes de référence qui sont nécessairement volumineux. En effet, la collecte des données sur le terrain et leur

[81] Biber défend ici sa vision de « corpus non proportionnel », du point de vue de la représentativité des corpus, nous discuterons de la position de Biber en détail en 2.6.1.

[82] Notamment Biber, Conrad, & Reppen, 1998; McEnery & Hardie, 2011; McEnery & Wilson, 2001; Meyer, 2002.

transcription sont des opérations pouvant s'étaler sur plusieurs années, et la reconnaissance automatique de la parole n'est pas encore une méthodologie totalement au point, malgré ce que prévoyait Sinclair il y a plus de vingt ans :

> A decision I took in 1961 to assemble a corpus of conversation is one of the luckiest I ever made. Even in that time, I was assured that an automatic transcription of speech was 'just around the corner'. It still is. (1991 : 16)

Dans cette section, nous présenterons donc comment s'est peu à peu imposée (ou s'impose encore, le processus n'est pas clos) l'oralité dans les corpus oraux généralistes, par rapport aux deux points que nous venons de détailler. Nous commencerons par présenter les travaux de Damourette et Pichon qui ébauchèrent les prémisses d'une certaine intégration de l'oralité dans l'étude de la langue. Néanmoins, du point de vue de l'oralité des données, ils n'eurent pas de successeurs et n'influèrent pas – toujours du point de vue de la prise en considération de l'oral, nous le précisons – sur les recherches linguistiques. C'est plutôt aux États-Unis, beaucoup plus tard et indépendamment des travaux de Damourette et Pichon, que furent posées les bases d'une méthodologie rigoureuse dans la constitution de corpus oraux généralistes grâce aux travaux de Charles Fries qui constitua un véritable corpus oral, selon des normes rigoureuses. Nous présenterons donc également ses travaux, d'autant que ces derniers inspirèrent Randolph Quirk pour le lancement du projet SEU. Nous terminerons cette section en examinant rapidement la situation des corpus de français parlé contemporains destinés à des études généralistes sur la langue orale. Nous précisons encore une fois que type de corpus diffère des corpus que nous avons vus précédemment, qui sont soit des corpus de spécialisation comme les corpus d'enfants en acquisition du langage, soit des corpus à visée spécifique comme les corpus sociolinguistiques ou régionaux. Les corpus qui nous intéressent ici sont les corpus oraux de référence (ou sous-corpus oraux faisant partie d'un corpus plus vaste), qui reflètent la langue orale en général sans particularités individuelles, sociolinguistiques, régionales ou situationnelles.

1.6.1 Les prémisses de l'étude de l'oral avec Damourette et Pichon

En 1911, Jacques Damourette et son neveu Édouard Pichon, respectivement architecte et médecin[83], entament la rédaction de *Essai de grammaire de la langue française* (Damourette & Pichon, 1930) ; en 1930 paraît le premier tome, et les publications des autres volumes se poursuivront jusqu'en 1950, à titre posthume pour les deux derniers. Leroy & Muni Toke (2007 : 137) rapportent que la publication posthume est possible grâce au travail de Henri Yvon, qui rédige notamment le glossaire contenu dans le volume des *Compléments*, paru en 1952. L'apport de l'*Essai de grammaire de la langue française* à la linguistique française, et comment le situer dans l'histoire des grammaires françaises est abondamment discuté, dès la parution des premiers tomes[84] et nous n'évoquerons pas ici la structure de leur ouvrage et ses fondements théoriques. Ce qui nous intéresse est la prise en compte des données orales et la constitution d'un corpus par Damourette et Pichon.

L'*Essai* est volumineux : sept volumes in-quarto, et y sont traités quelque 31000 exemples (Leroy & Muni Toke, 2007 : 135), empruntés à toutes les époques et à tous les secteurs de la langue. À ce sujet, Benveniste (1939) dira que « l'accumulation de données produit des œuvres de plus en plus massives » (Cité par Leroy & Muni-Toke, (2007, n. 36)) et donne l'*Essai de Grammaire de la Langue Française* comme exemple. Pour constituer cette base de travail, Damourette et Pichon imitèrent leurs prédécesseurs en utilisant les sources classiques qui sont les œuvres littéraires, mais menèrent également des enquêtes sur la langue orale en archivant des entretiens destinés à cette fin, où en notant des remarques entendues dans la rue ou durant les consultations médicales de Pichon. Concernant cet intérêt pour l'oral, Damourette et Pichon se rapprochent de Saussure qui, dans le *Cours de linguistique générale* dénonçait, comme eux, « la suprématie usurpée de l'écrit sur l'oral » ; à ce propos, Damourette & Pichon écrivent :

> Et c'est ici le lieu de s'inscrire formellement en faux contre les assertions des auteurs qui vont affirmant au monde que dans la France d'aujourd'hui il y ait un écart sensible entre le langage écrit et le langage parlé. Tout ce qui s'écrit se parle. (2007 : 135-136)

[83] La mauvaise santé de Damourette l'empêcha de pratiquer son métier d'architecte, et Pichon était atteint de la maladie de Bouillaud qui le handicapa et l'emporta à 49 ans.

[84] Cf. à ce sujet les travaux de Leroy & Muni Toke (Leroy & Muni Toke, 2007; Muni Toke, 2007)

Les exemples oraux constituent une part importante de l'ensemble et sont par ailleurs documentés : âge, sexe, classe socio-professionnelle mais aussi religion et origine régionale. Mais le corpus ne peut être validé selon les standards actuels, d'abord en raison de la transcription que l'on ne peut supposer qu'hasardeuse. En effet, la transcription des exemples recueillis se faisait à la volée, en transcrivant de mémoire ce qu'ils avaient entendu. Cette méthode comporte toutes les imprécisions que note Blanche-Benveniste (Claire Blanche-Benveniste, 2010; Claire Blanche-Benveniste & Jeanjean, 1987) : transcriptions courtes et invérifiables. D'autre part, Damourette et Pichon modifiaient certaines données métalinguistiques afin qu'elles puissent corroborer leur théories sociolinguistiques (Muni-Toke, 2011).

Quels que soient les défauts des données orales recueillies pour l'*Essai,* il reste que Damourette et Pichon ont été des pionniers dans le traitement de la langue orale et dans la constitution de corpus. Leur linguistique de corpus en tant que méthodologie les a amenés à être confrontés aux problématiques liées aux « exemples singuliers » ou *hapax*[85] : l'exemple fournit-il systématiquement une information sur le système linguistique ? Les conclusions tirées de l'analyse d'un exemple sont-elles systématiquement généralisables à l'ensemble de la langue ? En un mot, l'exemple est-il systématiquement recevable ou certains exemples doivent-ils être considérés, selon Benveniste[86], en tant que « bévue[s] sans portée ou déviation[s] individuelle[s] » ? Ainsi, l'importance accordée par Damourette & Pichon aux énoncés oraux annonce le débat des années 1960 et 1970 entre linguistique d'introspection et linguistique observationnelle ou, en termes chomskyens, la valeur à accorder aux compétences linguistiques du locuteur (I-Language) par rapport à ses performances linguistiques (E-Language). Damourette & Pichon se positionnent pour une linguistique d'observation, « ils affirment la supériorité de l'énoncé collecté sur l'énoncé construit et défendent leur exemple singulier comme une donnée linguistique légitime, quand bien même elle serait perçue comme non acceptable par la majorité des locuteurs du français. » (Muni-Toke & Habert, 2011). L'importance des travaux de Damourette & Pichon est toutefois à

[85] Les exemples singuliers ou hapax sont des « énoncés perçus comme non interprétables hors contexte ou non représentatifs d'une règle générale pour la langue décrite » (Muni-Toke & Habert, 2011). Les auteurs citent également Pichon qui évoque ces hapax : « […] les tours singuliers, les phrases négligées, les lapsus seront appelés à concourir, à côté des emplois littéraires. En effet, c'est dans ces aberrances que se révéleront, sans freins rationnels ni normatifs, les tendances profondes du sentiment linguistique. »

[86] Cité par Muni-Toke & Habert (2011).

situer sur le plan théorique en termes de linguistique de corpus. Leur corpus n'est pas un corpus constitué au sens où il n'a pas de documentation et que les données ne suivent pas des protocoles de collecte définis. C'est plus tard, aux États-Unis, que Charles Fries posa les bases rigoureuses de la constitution d'un corpus oral généraliste, comme nous allons le voir.

1.6.2 Les travaux fondateurs de Fries

Charles C. Fries (1887-1967) est un linguiste américain, qui s'est principalement intéressé à l'enseignement de la langue anglaise en tant que langue première et seconde. Il publia notamment *Teaching and learning English as a Foreign Language* (1945), et *The structure of English: An introduction to the construction of English sentences* (1952). Ce ne sont pas ses travaux en tant qu'enseignant qui nous importent ici, mais la valorisation des données orales dans son analyse généraliste de la langue, à l'instar des objectifs de Damourette et Pichon, et plus précisément pour la constitution d'un corpus oral selon une méthodologie scientifique. L'influence de Fries sur l'analyse de l'oral, et notamment son influence sur Randolph Quirk, l'inscrit dans une lignée que n'ont pas connue Damourette & Pichon.

En ne considérant pas la langue orale comme une dépravation de la langue écrite et authentique, Fries se démarque de la tendance des professeurs de langue de son époque ; époque où l'un de ses collègues, F.N. Scott, compare le langage des enfants à l'école « au langage des animaux desquels ils descendent », et le décrit comme constitué « de modulations de sons primitifs, qui remontent probablement à l'enfance de la race »[87]. Fries voyait au contraire dans la langue parlée « le langage réel » qu'il lui fallait étudier et analyser, au lieu des habitudes de son époque qui décrivaient le langage principalement à partir des textes littéraires classiques. Ainsi, dans son ouvrage *Structure of English*, Fries cherche à identifier les caractéristiques de la langue orale en analysant cinquante heures de conversations téléphoniques, effectuées auprès de 300 locuteurs différents. De Fornel & Léon notent que cette étude est celle qui amena Fries à adopter une approche de la phrase spécifique à l'oral, et à introduire la notion d'« utterance » (énoncé) :

> Fries est sans doute un des premiers à rechercher une unité minimale de la conversation. Dans Structure of English, il se demande quelles procédures le linguiste doit utiliser pour définir, dans un corpus de données enregistrées, les portions

[87] Cité par Peter H. Fries (2008 : 98), la traduction est la nôtre.

> d'énoncés qui ne sont pas des parties de construction plus grande et pour déterminer les formes linguistiques susceptibles d'être isolées comme des énoncés indépendants. Il propose le tour de parole comme unité de conversation. (De Fornel & Léon, 2000 : 134-135)

Les travaux de Fries coïncidèrent avec la disponibilité des magnétophones portables et ce qu'ils permettent en termes de situations d'enregistrement sur le terrain et, de fait, une collecte de données orales plus représentative que celles limitées par les conditions de collecte en laboratoire. De Fornel & Léon (2000 : 136) notent que l'influence des approches de Fries est notable chez les sociologues, les ethnographes et les analystes conversationnels.

1.6.3 Les corpus oraux de référence contemporains en France

Comme nous l'avons déjà rapporté, l'étude de la langue orale a été longtemps déconsidérée en France, au moins jusqu'aux travaux de Blanche-Benveniste & Jeanjean, et ce malgré les tentatives de Damourette et Pichon qui n'instaurèrent pas une tradition d'étude de la langue parlée. Cependant, à l'instar de la situation à l'international, le fait que la langue parlée soit reconnue en tant qu'objet d'étude scientifique et légitime en France n'a pas résolu le problème de la disponibilité des données orales. La situation des corpus de référence en France est d'ailleurs encore plus préoccupante que dans d'autres pays ; ainsi, les corpus oraux de langue française disponibles à l'étranger sont non seulement plus conséquents, mais surtout consultables et disponibles. Nous citerons par exemple le Ottawa-Hull French Corpus au Canada[88], constitué de 3,5 millions de mots collectés auprès d'une population de 120 personnes, censée représenter la population francophone d'Ottawa ; ou le corpus Valibel[89] en Belgique, constitué d'environ quatre millions de mots, et consultable par un formulaire en ligne depuis 2006.

Nous insistons sur le fait qu'il ne s'agit pas uniquement de constituer un corpus, mais également de le rendre disponible, et les deux processus sont bloqués en France. Pour illustrer notre propos sur la disponibilité des données, nous citerons les trois plus importants corpus

[88] Site du laboratoire de sociolinguistique d'Ottawa :
http://www.sociolinguistique.uottawa.ca/materiel/canadien-fa.html

[89] Valibel : Centre de recherche sur les Variétés linguistiques du français en Belgique :
http://www.uclouvain.be/valibel
Nous reviendrons sur le corpus du Valibel, notamment sur ses conventions de transcription, en 2.7.2.2.

oraux de référence constitués en France ; le corpus du Groupe Aixois de Recherche en Syntaxe (GARS), le corpus de l'ancienne équipe d'accueil DELIC[90] et le Corpus de Référence du Français Parlé[91] (CRFP). Comme le montre l'inventaire de Cappeau & Seijido (2005), les données du corpus du GARS, dont la partie numérisée compte un million de mots, ne sont consultables que sur place. En ce qui concerne le corpus du DELIC, qui comprend presque 1 500 000 mots, il n'est pas mis à disposition. Enfin, le projet CRFP qui aurait dû, ou pu devenir ce que laissait entendre son intitulé, semble être suspendu et le site du CRFP ne donne pas d'informations, tandis que le dernier article sur le sujet date de nombreuses années (DELIC, 2004).

Le constat est donc le suivant : le linguiste qui voudrait consulter un corpus de français parlé n'a pas les moyens, à l'heure actuelle, de consulter un corpus à moins de se déplacer, et il ne trouvera d'ailleurs sur place qu'un volume de données très inférieur à ceux disponibles dans d'autres pays. Ce constat date, puisque l'on peut lire sous la plume de Blanche-Benveniste & Jeanjean (1987 : 4) que les corpus de français oraux faisaient défaut en France, et que ceux-ci étaient encore plus nombreux au Québec, en Belgique ou en Grande-Bretagne que dans les universités françaises. Si en 1987, les auteures disaient vouloir « commencer à combler ce retard », la situation demeure inchangée 20 ans plus tard, comme le notent Cappeau & Gadet :

> Il n'existe pas un très gros corpus de français parlé et, en particulier, il n'y a pas eu en France de volonté institutionnelle qui aurait conduit à la constitution d'un grand corpus oral. C'est, en contraste, ce qui a été fait pour l'écrit (*Frantext*) qui tend à s'imposer comme référence (210 millions de mots en 2004). (Cappeau & Gadet, 2007b : 130)

La situation est effectivement très différente en ce qui concerne les corpus anglo-saxons, les linguistes anglais et américains disposant de corpus tels le British National Corpus (BNC), le Bank of English (BoE), l'American National Corpus (ANC) ou le Corpus of Contemporary American English (COCA)[92]. Il ne s'agit pas d'un retard de la France par rapport à l'Angleterre et aux États-Unis, mais bien d'un retard français général ; en Europe et dans le monde, ont été constitués et diffusés le Reference Corpus of Contemporary Portuguese, le

[90] DELIC : Description Linguistique Informatisée sur Corpus, ancienne équipe d'accueil (EA 3779) de l'Université de Provence. A aujourd'hui rejoint TALEP : Traitement Automatique du Langage Ecrit et Parlé.

[91] Le site du projet est disponible ici : http://sites.univ-provence.fr/delic/crfp/

[92] Ces quatre corpus sont présentés plus en détail en 2.10.1.

Göteborg Spoken Language Corpus, The Spoken Turkish Corpus, la Bavarian Archive for Speech Signals Corpora et d'autres.

Ce retard est évidemment notoire. La preuve en est les nombreuses initiatives institutionnelles ou d'équipes que présentaient Cappeau & Gadet (2007b : 130-131) en 2007. Ces initiatives firent remarquer aux auteurs que malgré la pauvreté des ressources en France, « les choses sont en train de changer, et très vite ». Or il semble que, là encore, la situation n'ait pas beaucoup changé depuis, comme en témoigne l'appel à contribution pour deux journées d'études, organisées par le CNRS en mars 2013 :

> L'Institut de Linguistique Française (CNRS, 2393 FR) organise pour la seconde fois deux journées de réflexion et de débats sur le thème « Initiative Corpus de référence du français ». Le comité directeur de l'ILF a jugé qu'il entre effectivement dans les missions de la fédération de lancer une telle initiative au niveau national (corpus de textes écrits, corpus oraux, corpus de référence ouvert à la diachronie longue). La France, contrairement à d'autres pays ne s'est pas dotée d'un tel type de corpus. Or, les avancées scientifiques et technologiques, le développement des programmes et des infrastructures dans le domaine de ce qu'il est convenu d'appeler « Corpus » permettent de considérer que le contexte actuel se prête à l'accueil favorable d'une telle initiative.

Les raisons de ce retard français sont diverses et dépendent des corpus dont il est question. Elles peuvent être résumées comme suit :

- les formats et les codages des bandes son, ainsi que des transcriptions ne sont pas standardisées, certaines sont obsolètes ;
- certains corpus constitués durant les années 1980 rencontrent des obstacles juridiques. D'une part, l'exploitation des données n'a pas été en bonne et due forme permise par les locuteurs. D'autre part, l'absence de licence rend difficile l'identification des propriétaires ou des responsables des données ;
- les institutions françaises tardent à prendre l'initiative de financement des corpus oraux, ainsi que les entreprises françaises de l'édition ;
- les chercheurs français ne souhaitent pas systématiquement mettre leurs données à disposition.

Cette indisponibilité des corpus est due soit aux volontés des équipes, soit aux difficultés de diffusion de ces corpus dont les protocoles de constitution sont trop spécifiques, comme le détaille Luzzati :

> Ces corpus analogiques, corpus d'Orléans et corpus du GARS y compris, ont plusieurs caractéristiques. Tout d'abord, ce sont des « données propriétaires », difficilement partageables et qui ne circulent pas. Ils donnent lieu à thèses, à publications, mais l'accès aux données est limité à ceux qui les ont réalisées et à leurs groupes de recherches (Benveniste, Jeanjean 1987). Ce n'est pas que leur coût soit élevé, mais ils ont suscité un gros effort de transcription et ils ont souvent acquis un prix affectif pour leurs auteurs-réalisateurs qui sont les seuls à pouvoir pleinement les transcrire et les exploiter. (Luzzati, 2009 : 3)

Dans d'autres cas, le corpus est constitué de façon optimale pour une recherche précise et des transcriptions ou des annotations trop spécifiques le rendent inexploitables sous d'autres perspectives. Toutes ces raisons font qu'il y a en France un nombre important de « corpus fantômes », pour employer les termes de Baude & Abouda (2006 : 3), soit des corpus inexploitables pour des raisons juridiques, personnelles ou scientifiques.

Toutefois, un autre obstacle nous paraît déterminant dans le fait que la France, malgré son statut, ne possède pas encore ce type de corpus, et qui est un certain conservatisme français qui subsiste encore vis-à-vis de la langue parlée. Boulton évoque cet aspect en ces termes :

> En attendant, le plus grand corpus actuellement disponible au public est Frantext avec ses quelques 200 millions de mots. La composition même de ce corpus (principalement des textes littéraires datant du 16^{e} au 20^{e} siècle) est révélatrice de l'importance accordée au « bon usage » de la langue française. En effet, le manque relatif d'intérêt pour la linguistique de corpus en France peut être attribué à une méfiance envers une approche purement descriptive de l'emploi courant de la langue. Ce genre de barrière culturelle n'est donc pas à sous-estimer. (Boulton, 2007 : 37)

Debaisieux (2009 : 42-43) évoque également que « les représentations sur la langue parlée comme lieu de 'déviance' n'ont pas disparu des esprits », et déplore le fait qu' « il n'existe pratiquement aucun outil de consultation convivial, adapté à un usage d'apprentissage et surtout libre d'utilisation » en France ; selon l'auteur, « tout ou presque est à construire ». Aujourd'hui, les efforts fournis à cet effet le sont essentiellement par le consortium Corpus Oraux et Multimodaux (IRCOM), l'un des 9 consortiums du TGIR des humanités numériques

HUMA-NUM qui œuvre à faciliter le tournant numérique de la recherche en sciences humaines et sociales. Les missions principales de l'IRCOM sont les suivantes[93] :

- organiser et accompagner le développement de corpus oraux et multimodaux en linguistique en aidant les chercheurs à s'approprier les outils nécessaires et à développer des standards communs de référence ;
- développer la valorisation, la visibilité et l'accessibilité des fonds existants ;
- améliorer leur mise à disposition, leur mutualisation et leur interopérabilité afin d'intégrer les réseaux internationaux (notamment ERIC-CLARIN) ;
- intégrer la communauté des producteurs et utilisateurs de corpus oraux et multimodaux dans ces pratiques et réflexions.

Nous constatons que les objectifs de l'IRCOM concernent l'accompagnement dans la constitution de corpus et leur mise à disposition. Si cette démarche permettra indubitablement la mise en valeur de certains corpus, le grand corpus oral du français contemporain ne saurait résulter de l'agrégation d'un ensemble de corpus dont les finalités et les protocoles de constitution sont parfois très différents. Nous ne saurions prédire l'évolution des corpus oraux en France, qui reste tributaire d'une décision politique qui octroiera à une institution, ou à une équipe universitaire, les moyens financiers, temporels et humains, de constituer un corpus oral de français contemporain suffisamment représentatif. Cette décision peut survenir dans les mois qui viennent ou se faire attendre encore plusieurs années.

1.7 Corpus et enseignement des langues

Dans notre volonté de retracer la place qu'ont occupée les corpus en général, et les corpus oraux en particulier dans l'histoire de la didactique des langues, notre première tâche sera de discuter de ce parcours en prenant en compte les différents domaines linguistiques dans lesquels les corpus sont nés et ont évolués dans des conditions, des époques et des lieux différents. Nous traiterons dans cette section des liens entretenus entre la linguistique de corpus et l'enseignement des langues. Ces liens sont complexes à plusieurs niveaux car l'exploitation de la linguistique à des fins didactiques n'est pas une science homogène et la linguistique appliquée à la didactique des langues est polymorphe. Nous rappelons par

[93] Cf. site de l'IRCOM : http://ircom.huma-num.fr/site/accueil.php

exemple que le projet COBUILD était un projet lexicographique mais dont les finalités étaient la didactique des langues ; il en va de même pour le projet ESLO, qui représentait quant à lui une sociolinguistique appliquée à la didactique des langues[94]. À cette première difficulté manifestée dans l'hétérogénéité des domaines linguistiques appliqués à la didactique des langues, s'ajoute celle des différents procédés de recours aux corpus lorsqu'il s'agit de l'enseignement des langues. Nous avons pu constater que les corpus sont en relation avec l'enseignement des langues selon trois grands axes[95] que nous présentons comme suit :

1) la linguistique de corpus appliquée à l'enseignement des langues étrangères ;
2) la linguistique de corpus comme outil méthodologique en classe de langue ;
3) les corpus d'apprenants.

La linguistique de corpus appliquée à l'enseignement des langues étrangères

Dans ce premier cas de figure, des corpus sont constitués et exploités pour que le fruit des analyses menées sur ces corpus puisse être exploité en classe de langue. Ce fut le cas pour les projets que nous venons de rappeler qui sont autant d'exemples d'une linguistique de corpus appliquée à l'enseignement des langues : la constitution d'un corpus pour la création d'un dictionnaire destiné à des apprenants (projet COBUILD), la constitution d'un corpus sociolinguistique pour y puiser les usages réels de la langue parlée à des fins didactiques (projet ESLO) ou la constitution d'un corpus permettant la création d'une grammaire destinée aux apprenants (à l'image des grammaires de Quirk et de Fries). Dans cette section, nous détaillerons également comment les linguistes ont constitué des corpus pour l'élaboration des « listes de fréquence » qui furent des listes censées répertorié les items lexicaux et

[94] Il aurait été possible de discuter de ces deux projets au sein de cette section. Notre choix fut néanmoins de les présenter dans des sections propres car malgré l'impact didactique de ces projets, leurs objectifs affichés furent prioritairement orientés vers la lexicographie et la sociolinguistique. Nous consacrons la présente section aux projets, études et mouvements théoriques s'étant prioritairement destinés à la réflexion et à la théorisation de la didactique des langues via le recours aux corpus. Nous justifierons ce choix de présentation par un dernier exemple, celui des travaux de Fries, qui consacra certes nombre de ses travaux à la didactique des langues, mais dont l'ouvrage *Structure of English* (1952) fut davantage orienté vers une grammaire généraliste de la langue que vers son enseignement. Ces choix de présentation ne réduisent pas les perspectives didactiques des projets que nous n'avons pas inclus dans cette section.

[95] Nous précisons qu'il ne s'agit pas ici de la catégorisation de Fligelstone, très souvent reprise dans la littérature, qu'il a proposée dans l'un de ses articles (1993 : 98).

grammaticaux les plus usités au sein d'une langue, afin que l'enseignement des langues prenne en compte ces listes quand il s'agit de définir les priorités d'enseignement. Dans ce premier cas de figure, les apprenants, qu'ils soient en classe de langue ou en apprentissage en ligne ou autonome, ne sont pas en relation directe avec les corpus constitués. Ils ne se servent que des résultats obtenus grâce aux corpus, et peuvent ignorer l'existence des corpus et les méthodologies qui les accompagnent.

La linguistique de corpus comme outil méthodologique en classe de langue

Dans ce second cas de figure, les corpus sont utilisés au sein même de la classe de langue. Les outils méthodologiques sont présentés aux apprenants afin que ces derniers deviennent des utilisateurs directs des corpus, et jouent le rôle d'analystes des phénomènes linguistiques sous la supervision de l'enseignant qui joue le rôle de coordinateur de recherche. La naissance de cette école méthodologique est étroitement liée à la démocratisation de l'outil informatique, et donc au TAL et à l'Acquisition des Langues Assistée par Ordinateur (ALAO). Nous présenterons les détails de cette méthodologie appelée Data-Driven Learning (DLL) par Johns (1991).

Les corpus d'apprenants

Dans ce troisième cas de figure, les corpus constitués sont des corpus de données d'apprenants, soit des corpus écrits ou oraux d'interlangue. Le but de tels corpus est de fournir des bases de données destinées à être exploitées au niveau de l'acquisition des langues étrangères et secondes. La constitution d'un corpus d'interlangue constitue le cœur de cette thèse ; pour cette raison, ce type de corpus sera analysé en détail dans le troisième chapitre. Pour l'instant, nous détaillerons ici les deux premiers cas de figure : nous discuterons principalement de la constitution de corpus pour l'élaboration des listes de fréquence pour illustrer la linguistique de corpus appliquée à l'enseignement des langues, et consacrerons ensuite une partie au DLL.

Mais en premier lieu, il nous importe de retracer la naissance des liens entre les corpus oraux et l'enseignement des langues afin de savoir quand, comment et pourquoi les enseignants-linguistes ont-ils introduit les notions de données ou bases de données orales dans la réflexion sur la didactique des langues. Il nous est apparu que les corpus oraux et l'enseignement des langues ont été liés grâce à la phonétique. Nous allons exposer le parcours commun des deux

disciplines en retraçant non pas l'histoire de la phonétique appliquée à l'enseignement des langues[96].

1.7.1 Phonétique, enseignement des langues et corpus oraux

La consultation empirique de données orales dans le domaine de l'enseignement des langues est un procédé ayant vu le jour grâce à la phonétique. Afin de démontrer ce propos, trois notions nécessitent des précisions ; la première est celle de l'oralité des données, la seconde celle de l'empirisme inhérent aux méthodologies des phonéticiens et la dernière celle de la relation entre phonétique et enseignement des langues. En premier lieu, la phonétique est liée à l'oralité pour des raisons évidentes[97], et les premiers pas de la phonétique moderne furent nécessairement contemporains des premiers phonographes. En second lieu, la phonétique ne saurait reposer sur l'introspection ou les intuitions du chercheur ; l'empirisme des méthodologies des phonéticiens est inévitable en raison du fait que l'expression d'un fait phonétique n'est jamais identique à un autre, alors que l'inverse est possible en ce qui concerne le fait linguistique s'il est limité à sa syntaxe. Labov résume ceci ainsi :

> Le postulat fondamental de la linguistique établi par Bloomfield au début de son développement reste toujours indiscutablement valable: à savoir que certains énoncés sont identiques. Le fait fondamental de la phonétique relève naturellement de l'affirmation opposée: il n'y a pas deux énoncés véritablement semblables. (Labov, 1975) [98]

Ce postulat rend inutile toute introspection et fait reposer l'étude de la phonétique sur la consultation de données authentiques. Ainsi, nous pouvons déduire de ces deux premiers points que toute phonétique est une linguistique de corpus oraux. Il nous faut maintenant nous attarder davantage sur notre troisième point, dans lequel nous expliciterons les liens entre phonétique et enseignement des langues.

[96] Il n'est nullement question de retracer une histoire de la phonétique, qui pourrait remonter jusqu'à la grammaire du sanskrit de Pāṇini, cinq siècles avant notre ère, en passant par le traité de Aelius Herodianus sur l'accentuation en grec ancien, jusqu'aux travaux de Alexander Melville Bell aux États-Unis.

[97] Bien qu'il y ait eu des études de prononciation qui se sont basées sur des questionnaires écrits, voir note n°101.

[98] Pour toutes les citations concernant l'article de Labov « What is a linguistic fact ? » (1975), nous rapportons la traduction parue dans « Marges Linguistiques : Qu'est-ce qu'un fait linguistique ? » (2001b).

Sur le plan théorique, la phonétique naît dans les milieux pédagogiques, du fait des préoccupations d'enseignants phonéticiens préoccupés par l'enseignement d'une langue orale, en réaction aux méthodologies classiques : ce sont en effet les nouveaux besoins sociaux[99] de la fin du XIX^{ème} siècle qui expliquent le progrès de la méthode orale dans l'enseignement des langues vivantes, et c'est la méthode orale qui porta les phonéticiens à la pointe du combat contre les méthodes traditionnelles. L'apport de la phonétique à l'enseignement des langues est « inestimable » (Galazzi, 1995 : 95) et reconnu dès le début du XX^{ème} siècle, comme en témoigne l'article de Théodore Rosset, « Du rôle de la phonétique dans l'enseignement des langues vivantes » (Rosset, 1909). Rosset emploie dans l'article l'expression de « phonétique pratique », et la décrit comme suit :

> La phonétique pratique est un art empirique : elle se propose d'observer et de classer les sons d'une langue donnée, de les comparer à ceux d'une ou de plusieurs autres langues ; elle utilise à cet effet les renseignements des oreilles, des yeux, du toucher, de l'histoire, des appareils, pour donner à ses observations le plus de précision possible ; elle note les fautes que commettent les divers individus en parlant une même langue étrangère et, à côté des fautes, les procédés pratiques de correction. Elle étudie de même la mélodie propre à chaque langue, sa façon d'accentuer les mots, les groupes de mots et les phrases, de les couper, de les chanter etc. L'ensemble de toutes ces remarques finit par constituer une analyse méthodique de la prononciation. (Rosset, cité par Puren, 1988 : 80)

Les figures les plus marquantes sont celles du phonéticien Wilhelm Viëtor (1950-1918) en Allemagne, et de deux enseignants-phonéticiens de langue en France: Paul Passy (1859-1940) et l'abbé Pierre-Jean Rousselot (1846-1924). En 1882, Wilhelm Viëtor, alors professeur de

[99] Concernant ces besoins, Puren (1988 : 66) rapporte que la société de l'époque préconisait que la langue ne devait plus se confiner à « un instrument de culture littéraire ou de gymnastique intellectuelle », mais devenir « un outil de communication au service [du] développement des échanges économiques, politiques, culturels et touristiques qui s'accélère ». D'autre part, Puren (1988 : 67) lie les réformes de l'enseignement des langues aux réformes profondes appliquées par les dirigeants de la Troisième République à partir de 1870 à l'éducation en général, et attribue à ces réformes des raisons idéologiques (les Républicains étaient persuadés que « l'éducation peut changer la société en changeant l'homme ») et militaires (défaite de la France face à l'Allemagne en 1870 et désir de revanche ayant poussé à la formation de citoyens responsables et autonomes) : l'Allemagne apparaissait comme un modèle à imiter pour mieux la surpasser, entre autres en généralisant l'enseignement des langues étrangères. Nous rajouterons que cette période fut celle de la Belle Époque, durant laquelle les progrès sociaux, économiques et politiques en Europe ont sans doute été décisifs pour un enseignement des langues permettant leur acquisition rapidement et efficacement.

phonétique à l'Université de Marburg, écrit un pamphlet intitulé « Der Sprachunterricht muss umkehren » (L'enseignement des langues doit faire volte-face), dans lequel il dénonce la méthodologie traditionnelle alors en vigueur dans l'enseignement des langues étrangères. Viëtor préconise un ensemble de procédés visant à la pratique de la langue orale en classe et critique la dominante traduction-grammaire de l'époque. Il n'est pas le seul. Puren (1988 : 86) rapporte plusieurs citations de théoriciens de l'époque qui commencent à se rendre compte de l'importance de l'oral jusqu'alors négligé : François Gouin (1880 : 93) écrira que « le véritable organe réceptif du langage, c'est l'oreille et point l'œil » ; et Charles Schweitzer (1904, avant-propos : 8) notera que « le livre n'est qu'un aide-mémoire pour l'élève ; il ne saurait prétendre à remplacer la parole vivante du maître qui est l'âme de l'enseignement ». Précisons que durant ces premières années d'habilitation de la langue orale, la seule justification est la nécessité d'assurer la maîtrise de la prononciation. Les notions de syntaxe ou de lexique spécifiques de la langue orale n'apparaissent pas encore dans les théories de l'enseignement des langues étrangères.

En 1886, Paul Passy fonde avec un groupe de professeurs de linguistique une association destinée à promouvoir l'usage d'une notation phonétique dans les écoles pour aider les enfants à acquérir plus facilement la prononciation des langues étrangères. Le groupe s'appelait initialement « Dhi Fonètik Tîcerz' Asóciécon ». En 1889, le nom de l'association devient « L'Association phonétique des professeurs de langues vivantes », et en 1897 fut adopté le nom actuel, « L'Association phonétique internationale » (API), dont Viëtor deviendra le président. L'association parraine et supervise la création de l'Alphabet Phonétique International. Avant la création de l'association, la phonétique est, selon l'expression de Passy, « la marotte de quelques toqués »[100]. L'intérêt de Passy pour la langue orale n'est pas linguistique mais humanitaire, d'après Abry *et al.* (s. d. : 58), qui rapportent que Passy voyait dans la communication et les échanges en langues vivantes un moyen de préserver la paix, et avouait un malaise pour les appareils et les nouvelles technologies.

Ce n'est pas le cas de Rousselot, dont la thèse en 1891, « Les modifications phonétiques du langage dans une famille de Cellefrouin », est inaugurale en investigation linguistique. C'est

[100] L'expression est rapportée par Enrica Galazzi (1995 : 98), de l'ouvrage de Paul Passy « Souvenirs d'un socialiste chrétien », paru aux éditions Je Sers en 1932.

l'une des premières études de la langue parlée faite à l'aide de vérifications instrumentales. L'appareillage associé lors de la soutenance est décrit par Galazzi ainsi :

> Les membres du jury, quelque peu décontenancés par l'apparat instrumental tout à fait inhabituel, avaient associé au jury un professeur de physique (Pellat) de la Faculté des Sciences de Paris (chose inouïe jusque-là). Cette thèse fit sensation dans le monde entier. Pour la première fois y étaient employés : palais artificiel, cylindre enregistreur, tambours, explorateurs de la langue, des lèvres, de la respiration, du nez, du larynx, inscripteur de la parole, spiromètre. (Galazzi, 1995 : 105)

L'accueil aux travaux de Passy et de Rousselot dans le milieu enseignant fut « loin d'être chaleureux », et « l'approche expérimentale, avec sa panoplie d'appareils mystérieux, inspirait la méfiance des linguistes » ; c'est pourquoi Galazzi (1995 : 111-112) rapporte l'essoufflement de la phonétique après la mort des deux enseignants-phonéticiens : la science n'est pas encore enseignée au niveau universitaire, elle manque de spécialistes qualifiés et « la lenteur des instituions » n'améliore pas l'état général malgré quelques disciples et partisans. Il n'y a pas eu depuis, à notre connaissance, de véritable enquête phonologique en France[101] ayant amené à la constitution d'un corpus oral selon des critères scientifiques, jusqu'au projet Phonologie du Français Contemporain[102] (PFC). L'élan de la phonétique appliquée à la

[101] Nous avons néanmoins constaté que la littérature citait les travaux d'André Martinet, en particulier *La prononciation du français contemporain* (1945) dont le sous-titre, « Témoignages recueillis en 1941 dans un camp d'officiers prisonniers » pourrait, selon l'auteur lui-même dans sa préface, dispenser de présentation. Il ne s'agit pas d'une étude prescriptive, mais d'un « ouvrage qui s'intéresse à la façon, répréhensible ou non, dont les différents Français prononcent leur langue maternelle ». Bien que, semble-t-il, il eût été impossible à Martinet d'avoir recours à des appareils d'enregistrement en 1941, il est dommage que nous ne possédions pas de corpus conséquent à cette enquête, mais uniquement les réponses aux 750 questionnaires écrits distribués par l'auteur.

[102] Le projet international Phonologie du Français Contemporain (PFC), est en collaboration entre l'Université d'Ottawa, l'ERSS (Université de Toulouse-Le Mirail), (Université de Paris X) l'Université d'Oslo et l'Université de Tromsø. Il est suivi par un projet similaire pour l'anglais, Phonologie de l'Anglais Contemporain (PAC), codirigé par Philip Carr (Université de Montpellier III) et Jacques Durand. Le projet général « part de la constatation qu'il est nécessaire de poursuivre le travail de description entrepris depuis au moins un siècle par tous les spécialistes de la communication parlée pour » 1) fournir une meilleure image du français parlé dans son unité et sa diversité 2) mettre à l'épreuve les modèles phonologiques et phonétiques sur le plan synchronique et diachronique 3) favoriser les échanges entre les connaissances phonologiques et les outils du traitement automatique de la parole 4) permettre la conservation d'une partie importante du patrimoine linguistique des espaces francophones du monde, et ce en contrepoint aux corpus déjà constitués 5) encourager un renouvellement des données et des analyses pour l'enseignement du français. Les principaux objectifs du projet

didactique des langues est ainsi né et mort en Europe, et s'il initia le concept – jamais formalisée à l'époque – du corpus dans le domaine, l'idée ne fut ni suivie ni exploitée. Comme pour Damourette et Pichon, il fallut attendre que des scientifiques aux États-Unis et en Angleterre reprennent le relais, pour des raisons contextuelles à leurs pays et avec davantage de succès.

1.7.2 Corpus, enseignement des langues et listes de fréquence

Après l'ébauche d'une linguistique de corpus appliquée à l'enseignement des langues en Europe, ce sont les pédagogues anglo-saxons qui repensèrent l'enseignement des langues pour des motivations anglaises et américaines qui furent similaires à celles des Allemands et des Français[103].

Après la Première Guerre mondiale, la France sort victorieuse mais meurtrie par les pertes humaines et matérielles, et ses priorités sont le désobusage et la reconstruction des départements dévastés. Puren (1988 : 147) écrit que « la France de l'après-guerre n'était plus cette nation inquiète, ouverte sur l'étranger à la recherche du renouveau et soucieuse de préparer ses enfants à l'action. On assiste au contraire après 1918 dans toute la pédagogie scolaire officielle à un net repli sur les valeurs « traditionnelles » de formation intellectuelle et culturelle, soit un enseignement insistant, en ce qui concerne les langues étrangères, sur le latin et le grec[104]. En ce qui concerne l'Allemagne vaincue, son état est catastrophique en

sont les études phonologiques et phonétiques : « accentuation, intonation, phénomènes phonétiques fins comme les assimilations de voisement ou la durée de segments phonétiques données ». Le PFC comporte également un volet intitulé « Interphonologie du français contemporain » (IPFC), dédié à l'étude des systèmes phonético-phonologiques des locuteurs non-natifs du français. Pour davantage d'informations concernant le PFC, consulter le site du projet à l'adresse suivante :

http://www.projet-pfc.net/

[103] Nous avons conscience du caractère réducteur que nous apposons à l'Europe en la limitant, dans notre étude, à la France et à l'Allemagne. Nous assumons ce choix pour deux raisons : en premier lieu, les courants innovateurs des enseignants-phonéticiens que nous avons présentés furent principalement allemands et français, et c'est avec ces courants que nous voulons faire le parallélisme anglo-saxon ; en second lieu, il n'aurait pas été très pertinent de pousser nos recherches aux autres pays européens, en raison du caractère mineur des études qui y furent menées d'une part, et cela eût été trop éloigné du cœur de notre problématique principale d'autre part.

[104] Cette situation perdura environ trois décennies. Puren (1988 : 147) rapporte à ce propos le témoignage de Legouis, qui avait commencé sa carrière dans l'enseignement des langues dans les années 1900, et qui constate en 1956 : « Dès la fin de 1919, je constatai avec douleur, au lycée Ampère, l'état dans lequel la Première Guerre

raison du fait qu'elle sort de la guerre non seulement ruinée, mais également contrainte de payer les réparations fixées par le traité de Versailles, qui déboucha sur une crise économique sans précédent pour l'Allemagne durant les années 1920[105]. Dans ce contexte, la jeune république de Weimar a suffisamment de peine à lutter contre l'hyperinflation et la crise monétaire qui engendrèrent émeutes et famines. Il y eut certes l'Âge d'Or de la république de Weimar qui a été rendu possible à partir du plan Dawes en 1923, mais qui fut stoppé par la crise de 1929. Dans ce contexte, les priorités allemandes ne furent pas l'éducation, et encore moins l'enseignement des langues étrangères.

En Angleterre et aux États-Unis, la situation était inverse. Aux États-Unis d'abord, avec la fin de la Première Guerre mondiale débuta l'expansionnisme culturel, économique et militaire américain qui perdure sous de multiples facettes jusqu'à nos jours ; en ce qui concerne l'Angleterre, l'Empire britannique atteignait son extension maximale après le traité de Versailles en 1919 et, dans le nouvel ordre mondial ayant émergé après la Première Guerre, l'Angleterre se rallia du côté des Américains en 1922 avec le Traité de Washington. Cette nouvelle donne internationale, accompagnée de la déchéance de la langue française en tant que langue diplomatique, fit de l'anglais la langue principale des affaires et de la politique dans le monde. Si l'on considère également le fait que le peuple américain est un peuple d'immigrés qui arrivaient rarement sur le sol des États-Unis en parlant déjà l'anglais, nous comprenons que la réflexion sur la didactique des langues fut principalement anglo-saxonne durant trois décennies.

Cette présente section sera divisée en trois sous-parties. En premier lieu, nous allons exposer les critères qui ont défini ces mouvements didactiques anglo-américains, en détaillant leurs origines et motivations philosophiques mais surtout pédagogiques. Dans la seconde sous-partie, nous passerons en revue les principaux projets élaborés en linguistique de corpus appliquée à l'enseignement de l'anglais. Enfin, nous terminerons par la présentation du

mondiale avait mis l'enseignement des langues vivantes : des professeurs vieillis, surmenés, obligés pour vivre de faire trop d'heures supplémentaires (trop facilement obtenues d'ailleurs), des classes encombrées, des enfants mal élevés en l'absence du père. Ce tableau, me dira-t-on, valait pour toutes les disciplines. Mais plus la machine est précise et délicate, plus elle souffre d'avoir à travailler dans de mauvaises conditions. Or, au lieu de faire un patient effort pour la remettre en marche, on préféra s'en prendre à la machine elle-même. C'est bien français, et somme toute la France a depuis trente-cinq ans l'enseignement des langues vivantes qu'elle a mérité alors ».

[105] L'Allemagne continua de payer ses dettes de guerre jusqu'en 2010.

Français fondamental ; il s'agit d'un projet français, pour l'enseignement du français, et dans notre propos qui est que la linguistique de corpus appliquée à l'enseignement des langues fut principalement appliquée à l'enseignement de l'anglais durant cette période, le *Français fondamental* constitue justement une exception qui, comme nous le verrons, fut inspirée des travaux anglophones, et qui ne trouva pas suite en France. L'importance que nous accorderons à ce projet est justifiée par ce caractère exceptionnel d'une part, et par l'originalité du corpus oral constitué d'autre part.

1.7.2.1 Objectifs des courants pédagogiques anglo-saxons et leurs origines philosophiques

Les didacticiens anglo-américains, à partir des années 1920, ont pour volonté d'enseigner l'anglais rapidement et efficacement ; la vitesse de l'apprentissage doit primer sur l'approfondissement des connaissances au travers de textes littéraires. La langue est avant tout un instrument de communication. Il faut enseigner l'anglais *vite* et *bien*. Pour enseigner *vite* et *bien*, les enseignants ont l'intuition de donner la priorité dans l'enseignement à ce qui pourrait le plus probablement être utilisé en situation de communication réelle. Ils entendent s'éloigner de ce que Sinclair qualifiera longtemps plus tard des « manufactured, doctored, lop-sided, unnatural, peculiar, and even bizarre examples »[106], inventés ou élicités pour les méthodes d'enseignement. Ainsi, l'intuition de ces didacticiens suggérait que les apprenants développeraient beaucoup plus rapidement leurs compétences linguistiques si on leur enseignait comment la langue était réellement utilisée. Et par usage réel, il faut entendre usage le plus commun : les apprenants devaient assimiler rapidement le lexique le plus répandu, les sens les plus courants et les constructions les plus fréquentes. Aussi fallait-il donc éviter les formes et les constructions archaïques, rares, désuètes ou appartenant à des registres trop soutenus ou trop spécifiques. Kennedy confirme *a posteriori* l'intuition qu'eurent ses prédécesseurs ainsi :

> Corpus linguistics has held potential relevance for the teaching of languages because responsible language teaching involves selecting what it is worth giving attention to. Since pedagogy attempts to reduce the time that would be necessary to learn a language through exposure alone, potential usefulness and likelihood of occurrence have been seen as relevant for deciding what to teach and learn. (Kennedy, 1992 : 335)

[106] La citation provient d'un article qui n'a pas été publié. Elle est rapportée par De Beaugrande (2000).

Pour enseigner la langue telle qu'elle est réellement utilisée, les exemples doivent remplir donc deux conditions : ils doivent être *courants* et *authentiques*. Considérons quelque peu ces deux conditions avant de poursuivre.

En ce qui concerne le premier critère, il consiste à établir, pour une langue donnée, une liste de vocabulaire simplifiée, qui concentrerait en un nombre restreint le lexique le plus usitée de la langue, afin de permettre à l'apprenant d'acquérir essentiellement les termes qui auront la plus grande probabilité d'utilisation en situation de communication réelle. Cette démarche est basée sur ce qui était alors un postulat : dans une langue donnée, un nombre limité d'occurrences représente la plus grande partie de cette langue, et un grand nombre d'occurrence n'apparaît que très rarement, et ne représente qu'une partie minime de la langue. Ce postulat sera formalisé par la loi de Zipf, repris par Kennedy (1992 : 335), qui affirme que « la linguistique de corpus a maintes fois démontré qu'une proportion importante des formes et des éléments d'un langage n'apparaît que très rarement dans l'usage réel », ou sur le site de la Bank of English[107], où il est écrit que « à peu près de l'anglais oral et écrit est constitué d'environ 3 500 mots ».

Gougenheim *et al.* (1956 : 23-30) offrent un panorama sur les origines philosophiques des vocabulaires simplifiés : volonté de mettre de l'ordre dans la langue pour certains philosophes, le vocabulaire simplifié a été abordé en tant qu'outil pédagogique, comme en témoigne une méthode d'apprentissage créée par l'abbé de l'Épée (1712-1789), qui s'était consacré à l'éducation des enfants sourds-muets, et dont l'essence était la simplification des termes et des paliers d'apprentissage. L'abbé de l'Épée exprime l'esprit de sa méthode sous forme de métaphore :

> Il faut émier le pain qu'on donne à de petits oiseaux, de peur qu'il ne les étrangle au lieu de les nourrir. (de L'Épée, 1776 : 87, cité par Gougenheim *et al.*, 1956 : 24)

Les origines du postulat de la simplification ne sont pas à rechercher uniquement dans le domaine de la didactique des langues. Gougenheim *et al.* citent également le philosophe anglais Jeremy Bentham qui, en tant que juriste et législateur, préconisa l'emploi « d'une langue juridique plus claire, plus vraie, plus conforme à l'analogie », ceci en restreignant le vocabulaire « aux termes compris par tout le monde », et en ne se servant que « d'un seul et

[107] Site de la BoE :
http://www.mycobuild.com/about-collins-corpus.aspx

même mot pour exprimer une seule et même idée ». Ce furent les préconisations de Bentham qui influencèrent le linguiste anglais Charles Kay Ogden (1889-1957) pour l'élaboration du *Basic English*, dont nous parlerons dans la section suivante.

Revenons-en maintenant au second critère que nous évoquions, celui de l'authenticité des données. Nous ne nous attarderons par sur la pertinence de la notion d'authenticité des exemples en classe de langue, car la réflexion a été maintes fois menée, et les chercheurs sont nombreux à reconnaître que la confrontation des apprenants avec des exemples authentiques est bénéfique (voir par exemple Römer (2008 : 116), qui offre une revue littéraire détaillée de ces études). Nous citerons Sinclair (1991 : 5), pour lequel il serait absurde de représenter une langue via des exemples inventés quand il est possible de le faire de par des exemples authentiques ; et Kennedy (1992 : 366), qui affirme que les exemple forgés peuvent donner une version normée, trop soignée du système linguistique visé. Le débat semble tranché, pourtant le recours aux exemples authentiques n'est pas sans controverses, non pas au niveau de l'intérêt des exemples authentiques en classe de langue, mais au niveau du jugement d'authenticité que l'on peut porter sur ces exemples, la problématique pouvant être formulée ainsi : pouvons-nous considérer que des exemples présentés comme authentiques le sont encore lorsqu'ils sont présentés en classe de langue, soit décontextualisés aux niveaux énonciatif et sociolinguistique ? Cette problématique est principalement discutée dans les travaux de Widdowson, qui affirme dans l'un de ses articles (1998 : 712) que le contexte d'un énoncé est un facteur aussi important que les données linguistiques elles-mêmes, et que si ce contexte venait à disparaître, la réalité de l'énoncé se verrait compromise. De ce point de vue, l'authenticité des textes pourrait être condamnée en classe de langue, si l'on considère cette dernière en tant que faux-semblant de situation réelle. Nous n'approfondirons pas la question car là encore, la problématique a maintes fois été discutée[108], et la plupart des chercheurs s'accordent à accepter « l'authenticité de l'interprétation du texte par l'apprenant et l'authenticité de l'activité de l'apprenant », c'est-à-dire « la réalité du contexte de la salle de classe comme l'espace où se joue la vie réelle des enseignants et des apprenants » (Chambers, 2009 : 18).

[108] Pour davantage d'informations concernant l'authenticité des données, nous renvoyons aux travaux de Widdowson, ainsi qu'à l'article de Gilmore (2007) qui fait une synthèse sur la question.

Nous n'avons abordé ces deux critères, qui amenèrent les didacticiens à enseigner une langue *courante* et *authentique*, que pour montrer l'impératif qui en découla naturellement : savoir ce qui est courant et authentique ne peut pas, par définition, reposer sur l'introspection des chercheurs, il leur fallut consulter des corpus et y mener des analyses statistiques[109] afin d'élaborer des listes de vocabulaire simplifié qu'ils proposeront aux apprenants. C'est ce contexte qui amena West (1930 : 511) à affirmer qu'après trois années d'apprentissage, les apprenants ne possédaient pas un lexique basique d'un millier de mots, arguant qu'ils passaient trop de temps sur des activités inutiles, qu'ils apprenaient un nombre important de mots qu'ils n'utilisaient que rarement et que les mots qu'ils « étaient supposés connaître » n'étaient pas maîtrisés. À la lumière de ces constats, West détaille les objectifs de l'enseignement d'une langue :

> The primary thing in learning a language is the acquisition of a vocabulary, and practice in using it (which is the same thing as acquiring). The problem is what

[109] L'origine des statistiques appliquées à la linguistique est à situer en Europe, indépendamment de la didactique des langues. La naissance des études quantitatives appliquées à la linguistique s'effectue grâce aux travaux de Friedrich Wilhelm Kaeding (1843-1928), qui était un journaliste allemand qui publia à la fin du XIX^{ème} siècle le « Häufigkeitswörterbuch der deutschen Sprache » (Dictionnaire de fréquence de la langue allemande) (Kaeding, 1898). Pour construire son dictionnaire, il dirigea la constitution d'un corpus de 11 millions d'occurrences. L'objectif était d'établir la fréquence des mots, des syllabes et des sons en allemand, en vue d'améliorer les connaissances en sténographie. L'enquête a été réalisée en analysant plus de 300 œuvres littéraires, et également de transcriptions de débats parlementaires, de textes administratifs et commerciaux, de journaux et de revue, d'écrits de théologie, de médecine, d'histoire et de lettres personnelles. L'immensité du travail nécessita une centaine de points de comptage dans l'empire allemand et la collaboration officielle de 665 personnes selon Aichele (2005), mais Kennedy (1998 : 16) et McEnery & Wilson (2001 : 12) rapportent que le travail aurait en réalité nécessité l'aide de cinq milliers d'assistants sur une période de plusieurs années. Le corpus de Kaeding, impressionnant même selon les standards actuels, annonce les domaines de la linguistique quantitative et des statistiques en linguistiques. Son successeur Helmut Meier prendra la relève et établira à partir du corpus des statistiques sur la durée de prise de la parole, sur la fréquence avec laquelle les lettres et les sons sont utilisés en allemand et sur la fréquence des phénomènes grammaticaux (Aichele, 2005). En 1968, et donc avant la démocratisation de l'informatique et des calculs électroniques statistiques sur corpus, Yvan Lebrun écrivait (1968 : 772) : « Que le dénombrement soit, dans certains cas, le seul moyen de découvrir la réalité linguistique n'implique évidemment pas que n'importe quel comptage opéré sur un texte soit une contribution à la science du langage ». Lebrun releva des différences entre les calculs de fréquence effectués par Kaeding et ceux de Meier. Il n'est pas surprenant que les calculs de Kaeding n'aient pas été conformes aux calculs antérieurs et aux standards actuels. Il faudra attendre plusieurs décennies, par exemple avec les travaux de Zipf, pour que des ouvrages de référence sur les statistiques du langage voient le jour.

> vocabulary ; and none of these 'modern textbooks in common use in English schools' have attempted to solve the problem. (West, 1930: 514)

C'est ainsi dans ce contexte utilitariste que les didacticiens anglophones élaborèrent les premières listes de fréquence ; nous nous proposons d'examiner les corpus sur lesquels ils se basèrent dans ce qui suit.

1.7.2.2 Les travaux anglo-saxons sur les listes de fréquence

Le premier travail sur corpus en vue d'élaborer une liste de vocabulaire simplifié fut celui d'Edward Thorndike[110] aux États-Unis, qui publia *The teacher's word book* (1921). L'ouvrage est un manuel initialement destiné à améliorer les programmes d'alphabétisation des locuteurs natifs de l'anglais étasunien, en établissant une liste de fréquence basée sur un corpus de 4 500 000 mots provenant de 41 sources différentes. Le *Teacher's Word Book* est une liste alphabétique des 10 000 mots les plus usités d'un calcul de fréquence sur 625 000 mots de la littérature pour enfants ; environ 3 000 000 de mots de la Bible et de l'anglais classique : environ 300 000 mots des manuels scolaires primaires ; environ 50 000 mots de livres portant sur la cuisine, la couture, l'agriculture, le commerce et autres ; environ 90 000 mots de la presse quotidienne et environ 500 000 mots de correspondance. 41 sources différentes ont été utilisées. Le calcul fréquentiel sur corpus effectué par Thorndike, évidemment manuel, devint un principe dans le domaine de l'enseignement des langues première et seconde. L'argument selon lequel un apprenant de la langue devait essentiellement être confronté aux mots les plus usités en premier lieu, soit le « vocabulary control », était lancé.

En parallèle aux travaux de Thorndike, l'Anglais Harold E. Palmer[111], qui enseignait alors au Japon et à la direction du « Institute for Research into English Teaching » à Tokyo, approfondit la notion des listes de vocabulaire simplifié en ne la limitant plus à une simple liste de mots ; il y inclut la notion de la collocation. La collocation, telle qu'elle est définie par

[110] Thorndike (1874-1949) était un psychologue américain, qui travailla essentiellement sur l'intelligence animale et la pédagogie de l'enseignement et de l'éducation.

[111] Harold Edward Palmer (1877-1949) était un linguiste né à Londres et qui enseigna la langue anglaise au Japon durant 14 ans. Ses apports à la didactique des langues sont considérables, d'une part en tant qu'adepte pionnier de la méthode orale, d'autre part pour ses travaux en linguistique de corpus appliquée à l'enseignement des langues.

Palmer (1933) dans son rapport[112], qualifie les cooccurrences fréquentes de deux ou plusieurs mots (*un regard profond*, *rire aux éclats*), sans que la collocation ne soit figée comme pour le cas des locutions ou des formes lexicalisées (*en fin de compte*, *chemin de fer*)[113]. Palmer considère donc que la valeur sémantique d'un mot n'est pas une valeur absolue et isolée, mais dépendante du contexte dans lequel le mot apparaît. Ses travaux donnent indirectement lieu au *Idiomatic and Syntactic English Dictionary*[114] dans lequel est effectué un repérage rigoureux des collocations, qui permirent de dresser un inventaire des « syntactic patterns », soit les constructions de la langue, « verbales d'abord puis nominales et adjectivales dans des éditions ultérieures et chez les concurrents, avec un codage permettant leur repérage, leur classement et leur utilisation » (Béjoint, 2007 : 19).

Ces travaux bénéficièrent d'une politique volontaire et d'investissements financiers et institutionnels ; sous la houlette de la Carnegie Corporation of New York[115] sont tenus deux grandes conférences à New York en 1934 et à Londres en 1935 auxquelles participent des linguistes spécialisés dans l'enseignement du langage dont Michael West qui travaillait alors en Inde, Harold Palmer, Edward Thorndike et Lawrence Faucett en Chine, qui publieront *Interim report on vocabulary selection* (Faucett, Palmer, West, & Thorndike, 1936). Le rapport préconise que dans l'objectif de faire de l'anglais une langue internationale et d'accélérer son apprentissage, le principe de sélection du vocabulaire à enseigner devait être appliqué : les mots qui ont la probabilité d'occurrence la plus élevée dans les corpus devaient être enseignés prioritairement.

[112] Le rapport n'est pas disponible en France. Nous résumons ici ce qu'en disent Williams (2006 : 153) et Béjoint (2007 : 19). Les exemples sont les nôtres.

[113] La notion de collocation annonce le courant linguistique du contextualisme qui sera développé par Firth. Nous reviendrons sur les questions de la collocation et du contextualisme en 2.2.3, pour les liens qu'ils entretiennent encore avec la linguistique de corpus.

[114] Palmer ne fait pas partie des auteurs du dictionnaire en question (Hornby, Gatenby, & Wakefield, 1942) ; c'est néanmoins à la demande de Palmer que Hornby, qui enseignait au Japon depuis 1923, élabora ce dictionnaire en se basant sur le rapport et les réflexions de Palmer. Hornby poursuivra sa carrière en publiant à Oxford ce qui deviendra le célèbre *Oxford Advanced Learner's Dictionary*.

[115] La Carnegie Corporation of New York est une fondation qui a pour vocation (entre autres) l'alphabétisation des adultes, les recherches en science de l'éducation, la facilitation de l'accès des minorités et des femmes à l'éducation ou la promotion des recherches en pédagogie, dans l'intérêt général.

Par la suite, Thorndike accrut son travail avec la collaboration d'Irving Lorge en publiant *The teacher's word book of 30,000 words* (Thorndike & Lorge, 1944), en se basant cette fois-ci sur un corpus plus large de 18 millions de mots. Le travail des auteurs fut d'indiquer, pour chaque mot, le nombre d'occurrence pour un million de mots. Les mots furent également classés selon le nombre d'occurrences : ceux qui paraissaient entre cinquante et cent fois par million de mots furent regroupés, et ceux qui paraissaient au moins cent fois par million de mots constituèrent un autre groupe. Enfin, et pour clore sur les principaux travaux anglo-américains basés sur le calcul fréquentiel des corpus, Michael West élabore une autre liste, la *General Service List of English Words* (West, 1953), qui comporte les 1490 mots les plus usités en anglais où chaque mot est décrit sémantiquement, à partir de l'analyse des corpus de Thorndike et Lorge.

Après la liste de West, soit après 1953, l'élan du calcul fréquentiel est brisé, semble-t-il en raison du caractère fastidieux du repérage manuel, car ce type de calcul ne reprendra que vers la fin des années 1980 avec la démocratisation des outils informatiques. L'exception d'importance fut celle du *Français fondamental*, que nous allons présenter en détail pour trois raisons qui l'opposent aux travaux anglo-américains : l'étude fut l'une des rares études européennes de ce type[116] ; elle a été effectuée dans les années 1950, plus de trente ans après la première étude sur corpus de Thorndike, soit à une époque où la France connut des motivations politiques similaires à celles des États-Unis dans les années 1920 ; et enfin elle reposa sur un corpus exclusivement oral.

[116] Il semble que la seule autre exception soit la publication de *Semantic frequency list for English, French, German, and Spanish: a correlation of the first six thousand words in four single-language frequency lists* (Eaton, 1940), qui recense les 6000 mots les plus fréquents en anglais, français, allemand et espagnol, et compare l'indice de fréquence pour chacune des entrées. Les détails de la constitution du corpus de travail ne sont pas disponibles, mais l'objectif était d'établir des listes de vocabulaire élémentaire. Ce travail, bien que basé sur la linguistique comparative, s'inscrit donc dans la lignée des travaux basés sur corpus et les corpus multilingues permettant le même type d'analyses ne seront récréés que cinquante ans plus tard, dans les années 1990 ; voir à ce sujet McEnery & Wilson, (2001 : 4)

1.7.2.3 Le *Français fondamental*

L'élaboration du français fondamental est la première étude de statistique lexicale menée sur un corpus de conversations spontanées enregistrées en français[117]. Blanche-Benveniste & Jeanjean, rapportent les paroles d'André Sauvageot, lors d'une interview qu'il leur a accordée en 1983 :

> Un jour de 1941, j'écoutais l'émission du Home Service de la BBC que j'avais l'habitude de prendre fréquemment ; j'entends W. Churchill parler du *Basic English*[118] et dire que c'était une arme fondamentale de domination pour la Grande-Bretagne et pour le English Speaking World. (Blanche-Benveniste & Jeanjean, 1987 : 78)

[117] Une polémique naquit du projet, accusé par certains à droite de détériorer le niveau de la langue et de la culture française en développant l'enseignement d'un « français petit-nègre pour étrangers fainéants et incapables », et par d'autres à gauche « d'être un nouvel instrument idéologique du colonialisme » (Puren, 1988 : 208). Cette polémique fut l'une des raisons pour lesquelles le *Français fondamental* fut ainsi nommé après que le premier opuscule intitulé *Le Français Élémentaire* parut en 1954. Les auteurs se défendent en affirmant que « dans l'esprit des promoteurs du français fondamental, il n'a jamais été question de fabriquer un français de seconde zone, mais au contraire de donner aux étrangers le moyen d'explorer de façon progressive, méthodique et de plus en plus différenciée l'ensemble du vocabulaire français » (Gougenheim *et al.*, 1956 : 13). Blanche-Benveniste & Jeanjean (1987 : 79) rapportent davantage de détails concernant cette polémique. Elle nous intéresse ici uniquement en tant qu'illustration des difficultés que connurent les linguistes à affirmer la langue orale en tant qu'objet d'étude légitime.

[118] Le *Basic English* (1934) est une liste de 850 mots d'anglais simplifié élaborée par le linguiste anglais Charles Kay Ogden (1889-1957), qui s'était inspiré des préconisations de simplifications du philosophe Bentham que nous citions plus haut. Il ne s'agit pas d'un calcul lexical sur corpus car la liste résulte des travaux introspectifs d'Ogden et de ses collaborateurs, le *Basic English* se rapproche davantage d'une langue artificielle censée pouvoir exprimer n'importe quel message en se basant uniquement sur les 850 mots de la liste. Suite au succès de la liste et au soutien de Churchill, Lecherbonnier (2005 : 38) rapporte qu'après la Seconde Guerre mondiale, « la commission nationale française de l'Unesco a mis à l'étude un projet identique pour la langue française en vue de l'enseigner au plus grand nombre en Afrique et en confia l'instruction à Léopold Sédar Senghor et à Paul Rivet, directeur du musée de l'Homme ; ce projet fut rapidement abandonné tant il souleva une marée de protestations portées, dans une coalition contre nature, par toutes sortes d'alliés issus d'horizons idéologiques divergents ». Gougenheim *et al.* (1956 : 27) prennent une position favorable à la liste, dont ils commentent le succès comme suit : « C'est le début d'une marche triomphale : au moment où éclate la Seconde Guerre mondiale le Basic est enseigné dans plus de vingt pays répartis sur tous les continents : aussi bien en Chine et au Japon qu'en Australie, dans l'Inde qu'aux États-Unis. Pendant la guerre mondiale 1939-1945 il a été mis en œuvre pour l'enseignement de l'anglais aux deux millions d'hommes de l'armée des Indes. »

Quand *L'élaboration du français fondamental* paraît en 1956 (Gougenheim *et al.*, 1956), son origine première remonte, selon ses auteurs, à une initiative de l'Unesco datant de 1947 ; un comité de linguistes, au sein duquel la France était alors représentée par Aurélien Sauvageot, « envisageait la diffusion de langues de civilisation comme l'un des moyens les plus efficaces de répandre largement l' « éducation de base » ». Certes, à l'époque, la langue française avait perdu de son rayonnement passé et n'est plus la langue universelle du XVIII[ème] siècle[119] ; d'autre part, la plupart des pays qui constituaient les protectorats et les colonies françaises avaient, en 1956, obtenu leur indépendance. Pourtant, la parution de l'ouvrage répond à un élargissement de la nécessité de diffusion de la langue française, pour les raisons suivantes :

- un grand nombre des anciennes colonies et protectorats français ont adopté la langue française comme langue nationale. Quand ce n'est pas le cas, le français reste une langue étrangère privilégiée. L'enseignement du français dans ces pays est une nécessité répondant à un désir des autorités respectives d'accroître la connaissance du français dès la scolarité, ainsi que chez les adultes par voie extra-scolaire ;
- les facultés et écoles françaises voient un grand nombre d'étudiants étrangers venir y faire leurs études ;
- en raison de l'essor économique et industriel d'après-guerre, un grand nombre de techniciens étrangers viennent accomplir des stages professionnels dans les entreprises françaises.

Ces raisons sont celles qui ont poussé ses auteurs à la publication du *Français fondamental*, en s'inspirant du *Basic English*. Leur travail « a constitué, avec de nombreux tâtonnements, une doctrine, nous dirons presque une science, des langues de base », et les auteurs décrivent ce concept de « langue de base » ainsi :

> La conception qui est à l'origine du vocabulaire de base, et, d'une façon plus générale, des langues de base, repose sur la notion de *limitation* du vocabulaire et de la grammaire. Les langues modernes ont un vocabulaire immense, enrichi de tous les apports de la technique et infiniment diversifié par une civilisation complexe (…). Pour assurer la diffusion rapide d'une langue, un décapage s'impose, qui ne laissera que les éléments essentiels. (Gougenheim *et al.*, 1956 : 11)

[119] Gougenheim *et al.* (1956) nuancent d'ailleurs la notion de « langue universelle » en indiquant que d'une part, le français n'était « universel » qu'en Europe, et que d'autre part, ceci ne concernait qu'une aristocratie éduquée mais peu nombreuse.

La question de la limitation des apports aux débutants n'est pas nouvelle. Les manuels traditionnels d'enseignement des langues n'enseignaient pas tout le vocabulaire et toute la grammaire, mais les limitations étaient le résultat de procédures empiriques, que ce soit au niveau de l'élaboration des manuels ou au niveau des jugements personnels des enseignants. Le plus souvent, les manuels offraient aux apprenants un panel extrêmement large de vocabulaire, et leur laissaient le soin d'opérer les choix nécessaires ; il en résultait que la plupart des apprenants avaient « un vocabulaire indigent » (Gougenheim *et al.*, 1956 : 12), d'une ampleur moindre que les vocabulaires de base.

Nous passerons outre ici les différentes étapes institutionnelles et politiques de l'élaboration du *Français fondamental*[120]. Il nous importe de savoir que la liste définitive de 1959 comporte – selon les termes des auteurs – 1475 mots, dont 1222 mots lexicaux et 253 mots grammaticaux. Cette liste fut établie d'après un calcul de fréquence sur un corpus oral recueilli auprès de 275 témoins provenant de 17 régions francophones. La volonté des auteurs à habiliter la langue orale dans l'enseignement des langues étant explicite :

> Or, indépendamment de l'intérêt scientifique que présente l'étude de la langue parlée, on constate qu'actuellement, et depuis un temps plus ou moins long selon les pays, l'enseignement des langues vivantes vise à mettre les élèves en état de comprendre la parole parlée et de parler eux-mêmes (et non pas seulement de lire des textes rédigés dans une langue étrangère et d'écrire dans cette langue). (Gougenheim *et al.*, 1956 : 61)

Afin de constituer leur corpus oral, le témoignage des auteurs sur les conditions d'enregistrement de la langue orale aux premières heures des magnétophones est précieux. Outre l'utilisation de quelques archives sonores[121], les auteurs exploitèrent l'idée d'Aurélien Sauvageot, qui proposa « un dépouillement de fréquence sur des textes de langue parlée, enregistrés au magnétophone ». Ayant conscience de l'impossibilité de travailler sur des faits de langue saisis au vol, les auteurs enregistrèrent donc 301 personnes[122] et transcrivirent ce

[120] Voir à ce sujet Gougenheim *et al.* (1956 : 12-13).

[121] Ces archives proviennent du Musée de la Parole, des Archives de la Radio, du Centre d'Études Radiophoniques et du Musée des Arts et Traditions populaires.

[122] Les auteurs regrettent de n'avoir pu enregistrer – en raison « des difficultés d'ordre matériel comme d'ordre social » – des personnes à leur insu. Ils précisent néanmoins avoir obtenu d'une commerçante parisienne qu'elle consente à laisser dans sa boutique un de leurs appareils durant une matinée. Les propos des clients sont décrits

qu'ils nommèrent « les textes parlés » de 275 témoins, après avoir éliminé 26 enregistrements dont les répliques furent jugées trop courtes. Les données sont notablement documentées (sexe, âge, origine géographique, sociale et professionnelle, niveau culturel et thèmes des témoignages) et aboutirent à un corpus de 312 135 mots. Les descriptions de la constitution du corpus, que nous nommons aujourd'hui protocoles de constitution, sont assez lacunaires par rapport aux standards actuels, bien que satisfaisantes[123]. En ce qui concerne la qualité du son, les auteurs ont conscience des limites de l'appareillage de l'époque (qualité et fidélité du son, durée d'enregistrement possible), mais minimisent – à juste titre – son importance vu leurs objectif linguistiques :

> Évidemment on ne pouvait demander à un appareil de ce genre la reproduction exacte des timbres. Mais cet inconvénient, qui aurait été sérieux pour une enquête de phonétique, était sans importance pour nous. L'interruption dans les conversations que nous imposait le changement des disques toutes les six minutes, au terme de leur durée maxima d'enregistrement, ne présentait pas non plus d'inconvénient. *(Gougenheim et al.*, 1956 : 67)

Néanmoins, les précisions manquent quant à un point essentiel qui est celui de la transcription. Les seules précisions sont les suivantes :

> Les enregistrements étaient transcrits par la personne même qui les avaient effectués dans le délai le plus court (…). Seuls les textes enregistrés par les soins du Centre d'Études Radiophoniques ont été transcrits par des sténotypistes. Les sténotypistes ne nous ont pas caché qu'étant habituées par la pratique des débats et des conférences à rétablir les phrases dans leur intégralité et à corriger les irrégularités et les lapsus, elles éprouvaient quelque difficulté à effectuer une transcription littérale, véritable photographie de la parole. (Gougenheim *et al.*, 1956 : 67)

Nous savons donc que les entretiens étaient retranscrits par l'intervieweur lui-même, sans correction des irrégularités et des lapsus, mais sans plus. Il n'est aujourd'hui pas possible de

comme « des échantillons criants de vérité du parler spontané du peuple de Paris » (Gougenheim *et al.*, 1956 : 64).

[123] La volonté de précision est évidente, notamment dans l'évocation de certains détails : poids du matériel d'enregistrement (6 kg et jugé d'un poids « peu élevé »), emplacement des micros, durée d'enregistrement maximum d'un disque (6 minutes), fragilité des supports ou volonté d'éviter certains sujets (religion, politique, vie privée).

comparer les transcriptions aux bandes originales, car celles-ci n'ont pas été conservées[124] pour des raisons de coût :

> Nous ne nous souciions pas non plus de la conservation des disques. Nous profitions largement de la possibilité qu'offrent les disques en papier magnétisé d'être effacés et de servir ainsi à plusieurs enregistrements successifs. Il aurait été beaucoup trop coûteux de conserver tous les enregistrements comme de bons esprits nous le suggéraient. (Gougenheim *et al.*, 1956 : 67)

Jusqu'au *Français fondamental*, l'étude de l'oral se faisait en France au moyen d'artifices : les exemples oraux parfois forgés de Damourette & Pichon, l'étude du français populaire par le biais de lettres par Henri Frei[125] ou l'étude de la prononciation du français par Martinet à l'aide de questionnaire écrits. Le *Français fondamental* inaugure ainsi l'étude de la langue parlée en France sur la base d'un véritable corpus. Toutefois Gougenheim et ses collaborateurs ouvrirent la voie mais ne furent pas suivis. Les écrits de Blanche-Benveniste & Jeanjean au milieu des années 1980 attestent que, 30 ans après, l'oral n'était pas encore linguistiquement légitime. Ceci ne minimise pas le caractère pionnier et audacieux des auteurs dont le seul manquement fut, à notre sens, de ne pas avoir prévu la conservation d'un tel corpus.

1.7.3 Le Data-Driven Learning

Nous avons vu que la volonté de confronter les apprenants d'une langue étrangère à la langue réelle a amené les didacticiens à élaborer des listes de fréquence, des dictionnaires et des grammaires pour apprenants, soit à appliquer la linguistique de corpus à l'enseignement des langues sous des formes différentes ; il s'agissait, avec l'étude des corpus d'interlangue, de l'une des trois possibilités de l'exploitation des corpus en classe de langue. Dans cette section, nous allons quelque peu détailler une troisième possibilité, qui est celle de l'exploitation des corpus directement en classe de langue.

L'idée d'ouvrir la possibilité aux apprenants de consulter des corpus en classe découle – en partie – des mêmes motivations qui entraînèrent les applications de la linguistique de corpus à

[124] Seules sont conservées les transcriptions originales sur microfilms. Celles-ci peuvent être consultées sur place en prenant contact avec Paul Rivenc à Toulouse.

[125] Henri Frei a publié *La grammaire des fautes* en 1929, dans laquelle il analyse le français « populaire » à partir d'un corpus de lettres écrites par les combattants de la Première Guerre mondiale.

l'enseignement des langues : l'apprenant doit être confronté à des données authentiques et non pas à une langue scolaire qui « ne semble pas exister à l'extérieur des classes de langue » (Mindt, 1996 : 232), ou à « des dialogues de méthode […] élaborés dans une langue qui n'est ni vraiment de l'écrit, ni vraiment de l'oral, et qui comportent des simplifications, voire des invraisemblances discursives » (Carton, 1995 : 73). Cette volonté de confronter les apprenants directement aux données des corpus est corrélée à la démocratisation de l'outil informatique, et cette méthodologie fut principalement théorisée par Tim Johns qui lui donne le nom de Data-Driven Learning (DLL). Pour Johns (1991 : 2), l'apprenant doit, de par la confrontation aux données de manière directe, approcher la langue en tant que chercheur pour en comprendre les fonctionnements ; il n'est plus un récepteur de savoirs transmis par l'enseignant, mais un chercheur qui créé ses propres connaissances de par l'observation de la langue sans informations prescriptives données *a priori*. Landure & Boulton résument la démarche ainsi :

> Plutôt que d'apprendre des « règles », ils explorent ainsi la langue à travers des corpus afin de détecter des « patterns » - des régularités ou des tendances générales, qu'elles soient linguistiques, pragmatiques, culturelles ou autres. Les corpus mettent ainsi en évidence l'usage de la langue dans des contextes authentiques (surtout les lignes de concordance) et révèlent des informations concernant les collocations, la fréquence et la distribution des items linguistiques, etc. (Landure & Boulton, 2010 : 13)

Ainsi le rôle des apprenants est de découvrir la langue, et Johns échelonne cette découverte en trois étapes qui sont 1) l'observation des données d'un corpus via un concordancier ; 2) la classification des données ; 3) la généralisation des principes de fonctionnements à partir des observations et des classifications effectuées.

Dans un tel contexte, le rôle de l'enseignant se voit profondément modifié : sa tâche est de guider les apprenants dans leurs requêtes sur le corpus, dans leurs analyses des données et de répondre à leurs éventuelles questions. Johns (1991 : 3) écrit que l'enseignant doit « apprendre à devenir un directeur et un coordinateur dans l'initiation des apprenants à la recherche ». Comme exemples de DLL en classe de langue[126], nous citerons Mauranen

[126] La concordance de paires de mots proches au niveau sémantique n'est évidemment pas la seule activité possible en DLL. De nombreuses autres possibilités, théoriques ou ayant été mises en application, ont été détaillées. Pour davantage d'informations concernant les activités en DLL, il conviendra de consulter Römer (2008 : 119).

(2004 : 89), qui argue que les corpus ont la capacité « à montrer ce qui est typique, ou commun dans le langage ». Mauranen poursuit en donnant l'exemple du verbe *to think* en anglais : le recours à un concordancier en classe de langue démontre que l'usage le plus fréquent n'est pas celui référant au processus mental, mais celui de « avoir une opinion » (Mauranen, 2004 : 90). Ce type de démarche quantitative permet à l'apprenant d'avoir accès à des éléments du langage non tributaires de l'intuition de l'enseignant.

En ce qui concerne l'efficacité de la méthodologie du DLL, Chambers (2005 : 111) et Römer (2008 : 120) citent un nombre important d'études qui prouvent la productivité du DLL en classe de langue, à différents niveaux : ainsi Cobb (1997 : 313) démontre-t-il que le recours aux corpus en classe de langue améliore l'acquisition du vocabulaire, et Bernardini (2004 : 21) affirme que la présentation des données aux apprenants en tant qu'*échantillons* et non plus en tant qu'*exemples*, leur permet de mieux cerner les régularités et les récurrences linguistiques. Sur un autre plan, Aston (2001 : 41-42) souligne le bénéfice tiré du nouveau rôle de l'enseignant en tant que coordinateur : d'une part il ne fait plus figure d'autorité en classe de langue, et son rôle en tant que « facilitateur » permet aux apprenants d'avoir accès à des informations plus fiables ; d'autre part et en ce qui concerne les apprenants cette fois, les méthodologies du DLL les force à s'impliquer bien davantage que dans une configuration classique et ils sont – de fait – poussés à l'autonomie dans leur apprentissage, une autonomie qu'Aston qualifie de « greatest attraction » des corpus en classe de langue. Bernardini (2004 : 27) discute également de l'apprentissage autonome et évoque son caractère centré sur les besoins, les volontés et les capacités individuelles des apprenants, ce qui renforce leur motivation. Enfin, l'implication de l'apprenant développe chez lui une initiation à la recherche qui lui permet d'acquérir un savoir-faire méthodologique qui développe ses capacités d'apprentissage en plus de son acquisition de la langue. L'apprenant « apprend à apprendre » et devient un « meilleur apprenant » selon les termes récurrents de Johns, pour lequel un « meilleur apprenant » ne signifie pas un apprenant ayant des compétences linguistiques élevées, mais un apprenant qui effectue mieux son apprentissage que celui jouant le rôle classique de récepteur passif de savoir.

Néanmoins, il est de nombreuses raisons pour lesquelles, malgré les avantages du DLL, celui-ci ne s'est pas systématisé dans les classes de langue : le premier obstacle est logistique, le second concerne la formation nécessaire aux enseignants et le dernier porte sur les apprenants eux-mêmes. En premier lieu, puisque le DLL nécessite la possibilité d'avoir recours à des

corpus électroniques en classe, il est donc nécessaire de pouvoir équiper les classes en ordinateurs connectés à Internet et d'avoir accès à des corpus jugés exploitables pour les tâches de DLL au niveau représentatif, corpus dont on aura payé les licences d'exploitation. La mise à disposition de ces outils entraîne un coût notable qui, de surcroît, n'est point limité dans le temps en raison de l'entretien des ordinateurs, leur remplacement, de la connexion Internet et des renouvellements des licences des corpus utilisés, à supposer que ces corpus soient disponibles pour la langue enseignée. Une fois ce dispositif mis en place, il est nécessaire de former les enseignants, à la fois au niveau technique pour l'utilisation des outils informatiques et logiciels, mais surtout au niveau théorique afin de leur permettre une exploitation idoine des corpus et la remise en question de leur rôle traditionnel. Cette formation ne doit pas être prise à la légère car, comme le note Sinclair (2004 : 2), « a corpus is not a simple object, and it is just as easy to derive nonsensical conclusions from the evidence as insightful ones », et les compétences en linguistique de corpus des enseignants est donc essentielle. Sur les deux niveaux technique et théorique, la formation des enseignants peut poser problème, tel que le formule Boulton :

> Certes, le DDL nécessite une remise en question des rôles traditionnels : les enseignants peuvent craindre une perte de contrôle, de pouvoir, de statut d'expert ; les apprenants peuvent refuser la responsabilité qu'ils auront pour leur propre apprentissage. Tous peuvent rejeter la nature « floue » du langage inhérent au travail sur corpus et préférer la certitude et la sécurité (même illusoire) qui accompagnent une approche traditionnelle à base de règles. Certains peuvent prétendre qu'ils manquent de ressources, surtout en termes d'ordinateurs et de logiciels, ou les moyens financiers pour les obtenir. Même lorsqu'elles existent, les ressources TIC font souvent peur et ne font que rebuter les utilisateurs ; un minimum de formation serait indispensable pour une utilisation productive du DDL, mais toute activité pour « apprendre à apprendre » peut être perçue comme une perte de temps qui ne sert qu'à retarder l'apprentissage de la langue elle-même. (Boulton, 2007 : 38)

Enfin, la formation des apprenants à la méthodologie du DLL semble l'obstacle le plus difficilement surmontable, car le revers d'un apprentissage autonome est la nécessité que les apprenants puissent assurer cette autonomie. En effet, un apprenant autonome doit être à même de définir ses besoins, de mener des réflexions linguistiques et de s'auto-évaluer, soit d'avoir un bagage de capacités quelque peu supérieures afin d'acquérir ces compétences. En effet, un travail de recherche inductif sur la langue n'est pas une démarche abordable pour tout le monde, et quand bien même elle le serait, le temps nécessaire à la formation des

apprenants est considérable. Johns lui-même est conscient – très tôt – du phénomène puisqu'il écrit :

> Talking about the DDL approach with other language teachers I am sometimes reproached that while this way of language-teaching by stimulating student questions and by doing linguistic research in the classroom on a cooperative basis may be all very well for students as intelligent, sophisticated, and well-motivated as ours at Birmingham University, it would not work with students as unintelligent, unsophisticated, and poorly-motivated as theirs. (Johns, 1991 : 12)

Pour l'ensemble de ces raisons, le DLL reste une activité très peu répandue dans les classes de langue[127] et, en France, la consultation des corpus dans les classes de langue ou dans l'enseignement en général reste anecdotique, principalement confinée aux milieux universitaires qui sont familiers avec les outils de la linguistique de corpus, tel que le rapporte Chambers :

> Dans l'enseignement secondaire, la quasi-totalité des professeurs de langue ne consultent pas les corpus, et dans l'enseignement supérieur il semble que les professeurs qui présentent des textes extraits de corpus et des concordances à leurs apprenants sont eux-mêmes des chercheurs dans ce domaine. Le jour où la consultation de corpus par l'ensemble des professeurs de langue sera aussi fréquente que la consultation du dictionnaire ou de Google semble être toujours bien loin. (Chambers, 2009 : 27)

Nous pouvons conclure sur le DLL que malgré ses nombreux avantages, il ne peut se répandre sans, avant tout, d'autres études prouvant que ses bénéfices justifient les investissements qui lui sont nécessaires. Ensuite, seule une politique institutionnelle peut mettre en place de telles pratiques sans quoi « le DDL risque de connaître le même sort que ces autres approches, de disparaître et ne représenter qu'une curiosité dans l'histoire de la didactique des langues » (Boulton, 2007 : 44), surtout si la disponibilité d'un corpus de référence représentatif de la langue réelle et donc possédant une partie orale conséquente est elle-même une difficulté comme c'est le cas en France.

[127] Boulton (2007), ainsi que Landure & Boulton (2010) attestent de ceci de par des revues quantitatives et évaluatives du DLL.

1.8 Conclusion du premier chapitre

Les auteurs qui se sont intéressés à la question s'accordent à dire qu'il est difficile de dater la naissance de la linguistique de corpus avec précision (Leech, 1991a : 1 ; Aarts, 2002 : 1). Il ne s'agit pas tant de vouloir adopter une posture d'historien que de comprendre les raisons de ces difficultés. Elles ont à voir avant tout avec la nature de la linguistique de corpus, qui est une notion vaste regroupant des domaines variés où les parcours ont évolué indépendamment les uns des autres. En ce qui concerne la linguistique des corpus oraux en particulier, ce premier chapitre nous a permis de vérifier les constats antérieurs qui ont abordé les corpus d'un point de vue global, si ce n'est exclusif de la langue parlée : l'intérêt pour la langue orale et la constitution de corpus oraux sont des démarches nées au gré d'intérêts ponctuels à des époques et des lieux différents. Mais quelques traits communs sont toutefois à souligner.

Dans un premier temps, nous avons pu constater que dans la plupart des domaines, l'intérêt pour la langue orale est né en Europe avec les travaux de Darwin, Taine, Preyer et Stern pour l'acquisition du langage ; avec les enquêtes de Gilliéron pour la dialectologie ; avec le recours à des exemples oraux par Damourette & Pichon pour la description générale de la langue ou avec l'intérêt d'enseignants comme Viëtor, Rousselot et Passy pour la phonétique appliquée à la didactique des langues. Mais de nombreux facteurs historiques ont fait en sorte que les travaux de ces pionniers n'ancrèrent pas l'étude la langue parlée dans la tradition européenne, et ce fut aux États-Unis et en Angleterre que Quirk, Fries ou Labov réinventèrent l'approche empirique de l'oral et instaurèrent des bases théoriques et méthodologiques que l'Europe, et la France plus particulièrement, ne suivirent que tardivement. Nous avons vu que l'intérêt pour le TAL en France débuta à l'heure où les Américains en constataient les limites ; que les principales bases de données françaises de corpus en acquisition du langage chez les enfants sont directement liées à la base de données CHILDES ; que les corpus sociolinguistiques français font non seulement défaut, mais que des corpus francophones ont été commandités par des pouvoirs publics canadiens (corpus Sankoff-Cedergren) ou sur l'initiative de linguistes anglais (corpus ESLO1) ; que le *Français fondamental* avait été inspiré par le Basic English et qu'il avait rencontré de vives oppositions en France et qu'enfin, et ceci est sans doute l'aspect le plus problématique de la chose, que la France ne possédait pas encore d'un corpus de référence du français parlé contemporain.

Pour citer à nouveau Blanche-Benveniste & Jeanjean (1987 : 85), « la curieuse aliénation du milieu linguistique français » ne peut plus être excusée par l'obstacle technologique. Il nous est clairement apparu que la naissance le magnétophone portable, puis les enregistreurs numériques, la naissance de l'informatique et la démocratisation des ordinateurs sont autant d'avancées technologiques corrélées à l'apparition d'une linguistique de corpus moderne et numérique, dont les corpus Brown ou COBUILD sont les premiers exemples. De nos jours où la France est à l'avant-garde des moyens technologiques et humains, comment justifier alors de son retard en la matière autrement que par un traditionalisme vis-à-vis de la langue parlée ? Certes, la constitution d'un corpus oral est coûteuse, mais l'aspect matériel ne saurait être décisif pour justifier le retard français là où de nombreux autres pays européens, comme nous l'avons vu, se sont dotés d'un ou de plusieurs corpus oraux de référence.

Nous avons également constaté un autre phénomène : les deux démarches que nous évoquions, l'intérêt pour la langue orale et la constitution d'un corpus oral, ne sont pas liées. C'est-à-dire que dans les domaines que nous avons présentés, l'approche empirique de la langue parlée ne fut pas systématiquement accompagnée de la constitution d'un corpus. Certes, ce constat est anachronique puisque nous parlons du « corpus » selon des standards actuels. Mais force est de constater que d'une manière ou d'une autre, ce phénomène persiste en France puisque nombre des corpus rassemblés et analysés pour telle ou telle étude ne sont pas mis à disposition, et qu'ils en deviennent ainsi un « corpus fantôme » selon Baude & Abouda (2006 : 3), ou un corpus « mort-né » selon Habert *et al.* (1997 : 156). À ce titre, les nombreux corpus constitués en France et laissés à l'abandon ne diffèrent pas, quant à leur potentiel d'exploitation, du corpus du *Français fondamental*, qui n'a pas été conservé.

Pour en revenir à la polyvalence des corpus, ce premier chapitre a démontré que l'usage des corpus est multiple et qu'ils ne sont pas rattachés à une branche particulière de la linguistique : les corpus sont exploitables pour la description de la langue, pour sa théorisation ou en linguistique appliquée. Pour ce qui concerne l'anglais, les exploitations de la linguistique de corpus ont permis des avancées « révolutionnaires », ne serait-ce que dans les grammaires généralistes et la constitution de dictionnaires, et « même les personnes qui n'ont jamais entendu parler de corpus utilisent le produit d'exploitations de corpus » (Hunston, 2002 : 96). Avec la multiplication de ressources et d'outils numériques basés sur corpus ces dernières années (outils de traduction en ligne, corpus littéraires de plusieurs milliards de mots), la linguistique de corpus apparaît comme une science jeune, en pleine croissance et aux

perspectives encore incalculables de nos jours. De ce fait, se pose la question de la linguistique de corpus – et du corpus – en soi. Certains y voient une méthodologie au service de la linguistique, à l'image du TAL, d'autres une discipline linguistique à part entière, alors que des théoriciens réfutent tout simplement l'intérêt du corpus en linguistique. Ce sont ces aspects que nous nous proposons de discuter dans le second chapitre.

2.Chapitre 2: Linguistique(s) de corpus, critères définitoires d’un corpus et types de corpus

2.1 Introduction du second chapitre

Historiquement et de par sa nature, l'un des traits principaux de la linguistique de corpus est l'approche empirique de la langue. Son objectif n'est pas de décrire « what people know about language, or what they perceive language to be » mais d'étudier « how language is used » (Tsui, 2004 : 39). D'autre part, comme nous l'avons vu tout au long du premier chapitre, les centres d'intérêt de la linguistique de corpus sont divers, et peuvent concerner l'acquisition des langues premières et secondes, l'élaboration de dictionnaires et de grammaires ou l'étude des variations sociales et régionales en se basant sur des faits linguistiques authentiques. Autrement dit, nous voyons la linguistique de corpus en tant que linguistique appliquée à tel ou tel domaine de la linguistique, et basée sur des corpus linguistiques dans sa démarche. Nous citerons à ce propos ce passage de la biographie d'Otto Jespersen, publiée en 1938, dans lequel il discute de l'approche empirique du langage :

> I am above all an observer; I quite simply cannot help making linguistics observations. In conversations at home and abroad, in railway compartments, when passing people in streets and on roads, I am constantly noticing oddities of pronunciation, forms and sentence constructions (…). For these notes I have found it practical to use small slips of paper… It is impossible for me to put even a remotely accurate number of the quantity of slips I have had or still have. [128]

Ce que décrit Jespersen pourrait être une illustration de la linguistique de corpus, à savoir déduire des généralités à partir de données linguistiques authentiques. Jespersen ne fut évidemment pas le premier linguiste à répertorier des faits linguistiques pour les étudier par la suite ; tel que le formule Laks (2008 : 7), l'observation de faits linguistiques a de tout temps existé :

> De façon générale, on peut dire que depuis l'origine, grammaire, philologie et linguistique sont fondées sur ***l'observation de faits de langue.*** Ces faits sont organisés comme des compendiums, ensembles plus ou moins vastes de données et d'exemples formant ce que Auroux appelle ***un observatoire***.[129] (Laks, 2008 : 7)

Ces « observatoires » peuvent être représentés par les grands corpus lexicographiques anglais du XX^{ème} siècle, principalement celui de Murray, ou par les corpus destinés à l'élaboration de

[128] Otto Jespersen (1860-1943) était un linguiste danois, spécialiste de grammaire anglaise. Le passage que nous rapportons est cité par Svartvik (1992 : 7).

[129] En gras et en italique dans le texte.

grammaires et de listes de fréquence pour apprenants à partir des années 1920 aux États-Unis. Néanmoins, c'est bien avec les premiers corpus électroniques que la linguistique de corpus connut une ère nouvelle grâce à l'informatique, bien plus féconde que tout ce qui précéda. Ainsi, comme l'exprime Mair (in Svartvik, 1992 : 99), la linguistique de corpus commença avant qu'elle ne soit nommée ainsi, mais ce n'est qu'à partir des années 1950 qu'elle fit l'objet de théorisations et de critiques, avec un pic à partir des années 1980, et c'est en 1984, d'après Léon (2008 : 12), que le terme « corpus linguistics » apparaît en tant que titre d'un ouvrage collectif, *Corpus linguistics: new studies in the analysis and exploitation of computer corpora* (Aarts & Meijs, 1984). Comme nous l'avons vu pour nombre de projets et comme nous le verrons pour d'autres, les corpus SEU, Brown, LLC, LOB etc. étaient alors déjà lancés et les termes « corpus linguistics » vinrent désigner l'ensemble des travaux britanniques, scandinaves et néerlandais sur des corpus électroniques de l'anglais. Depuis, les chercheurs en « corpus linguistics » se sont constitués en communauté dont nous montrerons l'hétérogénéité, mais dont la vitalité est manifeste dans les nombreux colloques et ouvrages collectifs qui se tiennent régulièrement ; de même, depuis 1995, une revue internationale est consacrée à la linguistique de corpus, *International Journal of Corpus Linguistics*, ainsi qu'une collection depuis 1998, *Studies in Corpus Linguistics*.

Cette communauté est hétérogène car la linguistique de corpus est elle-même une notion hétérogène qui varie énormément selon le domaine auquel elle s'intéresse, mais également dans la constitution et la consultation des corpus, soit dans les méthodologies employées. Cette diversité dans les conceptions de la linguistique de corpus est formelle dans le titre de l'ouvrage de référence de Habert *et al.* (1997), intitulé *Les linguistiques de corpus*, où le pluriel représente l'hétérogénéité des champs recouverts par le terme. Par ailleurs, Williams s'interroge dans un article, intitulé « La linguistique et le corpus : une affaire prépositionnelle » (2006), sur la traduction des termes « corpus linguistics » en français. L'ambiguïté du terme en anglais qui permet la simple juxtaposition de « corpus » et de « linguistics », et l'invariabilité de ce dernier n'est pas permise en français. Cela implique des choix, linguistique*s* « de », « des » ou « sur corpus ». Évidemment, la problématique de l'article ne relève pas de la traductologie, mais est un prétexte pour exposer l'hétérogénéité de la linguistique de corpus, et ce qu'elle recouvre de diversité. Il n'y a pas une linguistique de corpus, mais des linguistiques de corpus.

Détaillons quelque peu les raisons de la multiplicité des approches. En premier lieu, la linguistique de corpus elle-même, prise au sens strict et minimal de consultation de données authentiques, peut être remise en cause, et nous verrons qu'un certain nombre de linguistes, principalement cognitivistes et notamment Chomsky, considèrent que la réflexion sur la langue ne nécessite aucunement le recours à des données authentiques, et encore moins à la constitution de corpus sophistiqués, soit une linguistique purement « introspective ». Un autre courant, totalement à l'opposé du premier, juge, quant à lui, que la consultation des corpus est non seulement obligatoire, mais également qu'elle se suffit à elle-même. C'est-à-dire que les corpus, une fois constitués, révéleraient des vérités linguistiques. Ce courant positiviste qui « croyait [au XIX^{ème} et au début du XX^{ème} siècle] qu'il était une simple question de temps et de puissance des machines avant que la nature ne rende tous ses secrets » (Scheer, 2004 : 183) est présenté ainsi par Mayaffre :

> Au fond, il n'y avait plus de système, mais seulement des réalisations multiples, variées, imprévisibles qu'il fallait compiler dans des macro-corpus. Finalement, il n'y avait plus de règles ni de structures mais seulement des entorses aux règles, des exceptions particulières que l'on trouvait dans des corpus oraux ou écrits de plus en plus spécifiques. Seul le relevé –si possible exhaustif– d'énoncés authentiques permettait d'avoir un rendu de l'activité langagière. (Mayaffre, 2005, sect. 1)

En raison de ces problématiques, nous débuterons donc ce chapitre par une réflexion détaillée sur ces deux notions qui sont la linguistique introspective et la linguistique observationnelle. Pour ce faire, nous examinerons principalement le type de données auxquelles un linguiste peut avoir recours pour mener sa réflexion, et nous discuterons des limites et des avantages de chacune des démarches, et c'est pour mieux cerner et prendre conscience des limites de la linguistique de corpus que nous aurons recours aux critiques de Chomsky. En second lieu, nous discuterons de la seconde raison qui octroie à la linguistique de corpus sa diversité. Il s'agit de la problématique suivante : la linguistique de corpus, en tant que linguistique appliquée, est-elle une simple méthodologie, ou bien constitue-t-elle une branche indépendante des sciences linguistiques ? La réponse à cette question ne relève pas uniquement de l'ordre terminologique, car les divergences de vue ne sont pas formelles, ou constituées selon des préférences personnelles ou fortuites liées aux parcours de tout un chacun. Selon la vision qu'ont les linguistes des corpus de la linguistique de corpus, la constitution, l'annotation, la transcription et l'exploitation des corpus s'en voient profondément modifiées.

Quant à cette problématique, un premier courant est celui du corpus-based, et dont le principal représentant est Geoffrey Leech. Ce courant considère la linguistique de corpus en tant que stricte méthodologie, tel que le formule Leech :

> But is corpus linguistics really comparable with these other hyphenated branches of linguistics ? No, because « corpus linguistics » refers not to a domain of study, but rather to a methodological basis for pursuing linguistics research. (Leech, 1991a : 105)

De fait, en tant que méthodologie, la linguistique de corpus n'est qu'un outil de vérification indépendant des théories linguistiques, et le corpus devient un ensemble de faits linguistiques dont la fonction est d'infirmer, ou de valider les intuitions ou les théories linguistiques du chercheur, élaborées *a priori*. Quant au second courant, il s'agit du courant corpus-driven, qui fut principalement représenté par John Sinclair, et qui suggère qu'aucune théorie linguistique ne doit précéder la constitution et la consultation d'un corpus, car seul le recours aux données d'un corpus est susceptible de faire apparaître les théories linguistiques. Nous poursuivrons donc dans ce chapitre par une présentation un peu plus détaillée des origines de ces deux courants, de leurs conceptions respectives et des influences qu'ils ont opérées sur les méthodologies de constitution et de consultation des corpus.

Une fois que nous aurons discuté de la linguistique de corpus sur un plan théorique, nous nous intéresserons en détail au corpus en lui-même. Il s'agira d'en définir les critères définitoires et d'analyser en détail chacun de ces critères. Nous n'ambitionnons pas de proposer une définition définitive du corpus, mais de dresser une liste de critères destinés à le différencier d'une accumulation de données non directement exploitables. Pour souligner l'extrême diversité des corpus, nous conclurons d'ailleurs par un examen des différents types de corpus, en proposant pour chacun d'entre eux un exemple représentatif. Tout au long de l'étude du corpus en tant qu'objet linguistique, nous tenterons de valoriser deux facteurs : d'une part la spécificité des corpus oraux, d'autre part les corpus français, quand cela sera possible.

2.2 Les méthodes d'étude des faits linguistiques

Les données linguistiques sur lesquelles un chercheur peut se pencher résultent, grossièrement, de deux procédés : soit le linguiste a recours à l'introspection[130], soit il a

[130] Nous incluons dans le procédé d'introspection celui d'élicitation. L'élicitation est une méthodologie introspective également, mais au lieu de recourir à l'intuition personnelle du chercheur, il est fait appel à

recours à la consultation de faits linguistiques authentiques. La linguistique d'introspection (ou la méthode rationnelle) consiste à formuler une hypothèse, voire un modèle linguistique, puis à forger un ou des exemples à partir desquels le chercheur fera varier les paramètres pertinents pour y vérifier ses hypothèses. Les exemples forgés sont soumis au critère d'acceptabilité mais ne nécessitent que l'intuition du chercheur en tant que locuteur natif de la langue analysée. *A contrario*, la linguistique de corpus (ou méthode observationnelle) consiste à observer les formes qui figurent dans un corpus préexistant à l'étude et l'analyse. Après avoir lu, étudié ou analysé (grâce à aux outils d'analyse des corpus ou sans y avoir recours) le corpus, le chercheur en conclut des informations qui peuvent être vérifiées sur d'autres corpus. Si les vérifications sont satisfaisantes, le chercheur peut alors formuler des généralisations.

Chacune de ces deux approches présente un ensemble d'inconvénients. Nous nous proposons de détailler dans les sections qui suivent les limites de chacune des deux méthodologies, en commençant par les problèmes engendrés par la méthodologie introspective. Nous nous intéresserons ensuite, et plus en détail, aux limites de la méthodologie observationnelle, autrement dit de la linguistique de corpus.

2.2.1 Limites de la méthodologie introspective

Quand un linguiste aborde l'étude d'une langue par un ou des exemples forgés, le premier problème qui se pose est celui de l'acceptabilité de ces exemples. Or différentes expériences démontrent qu'il existe des variations dans les jugements d'acceptabilité. À propos de ces fluctuations dans les jugements, nous nous en référons à un article de Labov, « What is a linguistic fact ? » (1975), traduit en 2001 pour la revue *Marges linguistiques* (Labov, 2001b), dans lequel l'auteur rapporte le cas d'études où ont été confrontés les jugements de linguistes

l'intuition d'un groupe de personnes. Il serait tentant de supposer que le nombre de jugements sur l'acceptabilité des exemples étant plus important, et qu'il en résulterait des exemples plus fiables. Néanmoins, selon Meyer & Nelson (2006 : 99) le nombre de facteurs qui interfèrent avec les jugements de l'élicitation sont trop importants pour tirer pareille conclusion : langue parlée ou écrite, la présence ou non de l'expérimentateur durant le test, présence ou non des autres sujets, le fait d'avoir recours à des locuteurs linguistes ou non linguistes et enfin la forme du test (porter un simple jugement d'acceptabilité, reformuler, remplir des blancs).

aux jugements de locuteurs non spécialistes sur des exemples forgés ; il s'est révélé un taux de désaccord de 40%, et Labov en conclut :

> (…) aucune étude à ce jour n'obtient de résultats radicalement différents de ceux évoqués ci-dessus. À l'heure actuelle, aucun résultat ne permet d'entretenir l'espoir que les jugements introspectifs des linguistes soient fiables, reproductibles ou généralisables dans leur application au langage de la communauté. (Labov, 2001b : 33)

De plus, Labov rapporte le cas d'expériences menées par lui-même (2001b : 46-47), qui montrent de grands écarts entre les affirmations des locuteurs et leurs comportements linguistiques. De nombreux sujets tests utilisaient en interaction des structures syntaxiques qu'ils avaient jugé irrecevables. Ces expériences révèlent donc que l'introspection n'est pas systématiquement fiable, qu'il existe des variations entre individus, des variations individuelles dans le temps ainsi que des variations entre ce que l'individu lui-même juge comme acceptable et les énoncés qu'il produit.

Par ailleurs, dans un article intitulé « Les corpus fondent-ils une nouvelle linguistique ? », Cori & David (2008) rapportent les principales études s'étant penchées sur la constitution de de bases de données de manière introspective, dont celles de Leech (1991a), d'Abney (1996) et de Manning (2003). Nous résumons les constats de Cori & David (2008 : 114-115) comme suit : 1) les données introspectives sont trop peu volumineuses pour être représentatives (Abney, 1996 : 4 ; Manning, 2003 : 295) ; 2) leur caractère « inventé » les déconnecte de l'usage réel, les rendant invérifiables (Manning, 2003 : 296) et inaptes à vérifier les théories (Leech, 1992 : 111) ; 3) les chercheurs se basant sur l'introspection tendent à « nettoyer les données » et à les « ajuster » (Abney, 1996 : 11), et la conséquence en est que ces linguistes « travaillent sur des données artificielles » (Cori & David, 2008 : 114). Le problème de l'introspection est donc son impossibilité à refléter autre chose que l'idiolecte – selon le terme de Corbin (1980) – du linguiste, et cet idiolecte ne pourra rendre compte de tous les faits de langue. Ceci est particulièrement problématique en ce qui concerne certains domaines de la linguistique :

> L'introspection est impuissante à séparer dans un savoir linguistique les traits dialectaux et les traits non dialectaux : on ne fait pas de la sociolinguistique en chambre. Ces remarques valent pour les niveaux de langue. (Corbin, 1980)

Ainsi l'étude de la langue par introspection est tout simplement impossible lorsqu'il s'agit de l'étude des variations, des dialectes, des registres et styles de la langue, des études

diachroniques ou de l'interlangue. Dans ces domaines, l'intérêt du linguiste porte sur la performance, et non sur la compétence. Les terminologies compétence / performance sont celles de Chomsky qui utilisa également les termes de Internalized-Language / Externalized Language. Les générativistes stipulent que la grammaire est une compétence autonome de l'esprit humain, et que notre utilisation du langage est une performance qui ne reflète pas notre compétence ; en effet, les performances sont polluées par des erreurs, de la fatigue, de l'inattention qui ne permettent pas l'étude du langage réel, à savoir la compétence. Chomsky résume la chose en considérant que la performance est « au mieux un épiphénomène » lié à « d'obscurs et complexes éléments sociopolitiques, historiques, culturels et normatifs-théologiques », dont l'intérêt se limite aux études sociologiques, et aucunement à la compréhension de la nature du langage ou des processus cognitifs du locuteur (Chomsky, 2000 : 49).

Comme le note Leech (1991a : 107-108), ces conceptions indiquent un choix d'étude : s'agit-il d'opérer une description linguistique (performance) ou de s'intéresser aux universaux langagiers (compétence) ? Ainsi la description d'un des domaines d'un langage donné en se basant sur les performances de ses locuteurs, est considéré comme « lower-order-theory », tandis que la description des universaux du langage, comme dans la grammaire universelle de Chomsky, est une « higher-order theory », appliquée au langage en général, en tant que faculté humaine. Cependant, ces qualificatifs ne constituent pas une hiérarchie d'importance. Elles ne s'éliminent pas non plus. Elles sont deux approches qui différent de par la taille des domaines abordés, qui se complètent en permettant à la première la vérification des obtenues de la seconde. Halliday formule la même idée de la sorte:

> The "system" and the "instance" are not two distinct phenomena. There is only one phenomenon here, the phenomenon of language : what we have are two different observers, looking at this phenomenon from different depths in time. (Halliday, 1992 : 66)

Néanmoins Leech (1991a : 108) reste critique et indique que d'une part, les données de performance sont observables, alors que l'étude de la compétence repose sur des spéculations, et que d'autre part, la distinction entre compétence / performance n'est pas aussi nette.

2.2.2 Limites de la linguistique de corpus

Pour de nombreux auteurs, les théoriciens de la grammaire génératives sont les principaux détracteurs de la linguistique de corpus. Plus particulièrement, Chomsky est présenté comme un « anti-précurseur théorique » (Léon, 2005 : 47), puisqu'il écrivait dans *Syntactic structures* :

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description of language based on the corpus would be no more than a mere list. (Chomsky, 1957 : 159)

Cette vision est partagée par d'autres génératives : dans un de leurs articles, Biber & Finegan (1991) rapportent[131] une conversation ayant eu lieu au début des années 1960 entre Nelson Francis et Robert Lees[132]. Ce dernier demanda à Francis quelles étaient ses occupations actuelles, et Francis lui répondit qu'il travaillait sur la constitution d'un corpus afin de repérer « les véritables faits de la grammaire anglaise », ce à quoi Lees répondit « avec étonnement » :

> That is a complete waste of your time and the government's money. You are a native speaker of English; in ten minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text.

La position de la grammaire générative vis-à-vis des corpus étant clairement exprimée, intéressons-nous aux raisons théoriques de ce rejet, car l'analyse de ces critiques nous permettra de mieux situer les limites de la linguistique de corpus[133].

[131] La conversation est rapportée à l'origine par Francis, dans un de ses articles datant de 1982, indisponible en France.

[132] Nous rappelons que Francis était l'un des principaux responsables du Brown Corpus. Robert Lees était un linguiste génératiste américain ayant travaillé sur la TA à ses débuts.

[133] Nous nous proposons de discuter du positionnement théorique des génératives vis-à-vis des corpus pour la raison que leurs critiques constituent une sorte de garde-fou au positionnement théorique des méthodologies empiriques. Néanmoins, la personne même de Noam Chomsky est présentée, parfois de manière quelque peu romancée, en tant qu'adversaire acharné envers les corpus ; les supposés méfaits de Chomsky auraient, de plus, ralenti l'avancée des corpus un quart de siècle durant. Les exemples sont presque systématiques dans les ouvrages ou articles unanimement reconnus de référence. McEnery & Wilson (2001 : 4) écrivent ainsi : « The fact remains that we can pinpoint a discontinuity in the development of corpus linguistics fairly accurately in the

Chomsky a défini une procédure d'évaluation des descriptions grammaticales et des théories linguistiques en trois niveaux. Ces niveaux évaluent le degré d'adéquation de la grammaire ou de la théorie et sont : l'adéquation observationnelle, l'adéquation descriptive et l'adéquation explicative. Ces niveaux sont expliqués par Chomsky (1965: 100-101) comme suit : une grammaire atteint le niveau d'adéquation observationnelle si elle est en mesure de décrire les données à disposition et rien de plus. Si une description grammaticale atteint ce premier niveau d'adéquation, cette description grammaticale est alors en mesure d'affirmer la conformité ou la non-conformité d'un énoncé. Pour Chomsky, ce niveau est le plus faible car il se limite à évaluer la recevabilité grammaticale de l'énoncé. Le second niveau est celui de l'adéquation descriptive. Si une description grammaticale ou une théorie linguistique atteint ce niveau, elle sera en mesure non seulement de juger de la recevabilité de l'énoncé, mais également de spécifier les propriétés grammaticales de l'énoncé, sachant que la description portera uniquement sur les propriétés grammaticales ayant permis de juger l'énoncé recevable. Le dernier niveau pouvant être atteint est celui de l'adéquation explicative. À ce stade, la théorie linguistique est en mesure de fournir un nombre limité de règles

late 1950s. After this period the corpus as a source of data underwent a period of almost total unpopularity and neglect. Indeed it is no exaggeration to suggest that as methodology it was widely perceived as being intellectually discredited for a time. This event can be placed so accurately because its source lies almost exclusively with one man and his criticisms of the corpus as a source of information. That man was Noam Chomsky » ; Leech (1991a : 110) : « The impact of Chomskyan linguistics was to place the methods associated with CCL [Computer Corpus Linguistics] in a backwater, where they were neglected for a quarter of a century » ; Kennedy (1998 : 19) : « The Brown Corpus was significant not only because it was the first computer corpus compiled for linguistic research, but also because it was compiled in the face of massive indifference if not outright hostility from those who espoused the conventional wisdom of the new and increasingly dominant paradigm in US linguistics led by Noam Chomsky ». Or il apparaît que nombre de corpus d'importance ont été créés dans les années 1960 et 1970, comme le Brown, le LLC, le LOB, la poursuite du projet SEU et le lancement du projet COBUILD dans les années 1980. Il n'y a, à notre connaissance, que des auteurs français qui ont tempéré l'impact négatif de Chomsky sur la linguistique de corpus, comme l'article de Léon (2005) consacré à la question, ou les propos de Habert *et al.* (1997 : 8) : « Le rejet de principe, formulé par N. Chomsky dès 1957, du recours aux corpus au profit de l'appel à l'intuition du locuteur natif a relégué dans les limbes les travaux de linguistique quantitative et les études empiriques de données attestées. C'est, du moins, l'impression qui domine quand on se retourne sur les quarante dernières années de l'histoire de la linguistique. Cette image est partiellement fausse ». Nous n'irons pas jusqu'à affirmer, comme le fait Léon (2005), qu'il y a eu une démarche réfléchie de victimisation de la part des linguistes de corpus anglo-saxons, mais il est néanmoins étonnant qu'après un constat factuel, Chomsky soit autant vilipendé dans la littérature.

grammaticales pouvant produire tout type d'énoncé. Cette grammaire est l'objectif principal de la théorie générativiste et peut être assimilée à la grammaire universelle proposée par Chomsky. Or, pour Chomsky (1965 : 52), une approche empirique en linguistique ne peut, et ce dans le meilleur des cas, grâce à un corpus regroupant l'ensemble des énoncés possibles d'un langage, accéder qu'au premier niveau d'adéquation observationnelle. Autrement dit, la linguistique de corpus est insuffisante même dans une configuration de travail idéale (soit avoir à disposition un corpus représentatif).

La position des générativistes, dans leur recherche d'une grammaire universelle, peut être soumise à plusieurs critiques. La linguistique de Chomsky présente la syntaxe du langage en tant que système génératif, permettant la production d'un nombre infinis de phrases à partir d'un nombre défini de règles. Cette démarche comporte les risques d'une théorisation en amont, tel que le note Laporte :

> Le danger de cette tendance est de faire perdre de vue qu'une règle générale, mais qui n'est pas en conformité avec la réalité, est une généralisation hâtive, sans valeur scientifique. (Laporte, 2008 : 14)

La critique de Laporte s'inscrit en faux contre les tenants d'une linguistique introspective pure, notamment en raison de l'impossibilité d'un « nombre défini de règles » à englober tout ce que la langue comporte « d'exceptions et d'irrégularités ». Nous tenons à faire remarquer que ces notions ne le sont que si les données sont appréhendées par rapport à des règles prédéfinies. Sans *a priori*, une exception ou une irrégularité n'a pas lieu d'être. Or la grammaire générativiste stipule justement que ces exceptions et irrégularités ne constituent pas un sujet d'étude. Un corpus se retrouvera « souillé » par les performances des locuteurs, selon le terme de Mayaffre :

> En corpus, la grammaire universelle se trouve souillée par la culture, la société, l'humeur ou les pathologies du locuteur, les choix, la sélection de l'analyste, etc. Les corpus de données attestées non seulement ne permettent pas de révéler le système mais le brouillent inévitablement ou le parasitent par divers bruits, le rendant ainsi inaudible au théoricien. (Mayaffre, 2005 : 2)

Or ces « souillures » sont précisément ce qui motive un linguiste de corpus à constituer ou à consulter un corpus. Meyer (2002 : 3) souligne que la complexité des structures est ce à quoi le linguiste de corpus s'intéresse. Contrairement aux grammairiens générativistes, les linguistes de corpus voient dans la complexité et la variation des phénomènes inhérents au

langage, et accordent dans leurs réflexions sur le langage la priorité à l'adéquation descriptive et non à l'adéquation explicative. Le linguiste de corpus est par conséquent sceptique quant aux réflexions linguistiques abstraites et décontextualisées tenues par les grammairiens génératives, principalement en raison du fait que leurs réflexions sont trop éloignées des usages réels du langage. Outre l'incertitude des conclusions génératives, leurs conclusions sont sujettes aux erreurs de jugements personnelles. Nous rapportons ici, pour l'anecdote, une erreur de Chomsky lui-même sur un usage en anglais ; il s'agit d'une conversation entre Chomsky et Hatcher :

- Chomsky: The verb *perform* cannot be used with mass word objects: one can *perform a task* but one cannot *perform labour*.
- Hatcher: How do you know, if you don't use a corpus and have not studied the verb *perform*?
- Chomsky: How do I know? Because I am a native speaker of the English language. (Conversation rapportée par McEnery & Wilson, 2001 : 11)

Les auteurs notent que Chomsky était dans l'erreur, car la structure « perform magic » est possible, sur base de consultation du BNC.

Leech (1991b) argue que dans les années 1950, quand Chomsky présente sa vision de l'approche de la langue par le biais d'un corpus, les outils offrant une démarche empirique de plusieurs centaines de millions de mots étaient encore inconcevables. Cette possibilité, encore accrue de nos jours, est considérée par Leech comme un moyen unique de vérifier la rigueur d'un système linguistique. Il est à noter que Leech suppose que même si Chomsky avait pu prévoir les progrès futurs de la linguistique computationnelle, il n'aurait pas changé d'avis. En ceci, Leech avait raison en 1991, Chomsky déclarant en 2004 que « la linguistique de corpus ne [signifiait] rien » :

> Corpus linguistics doesn't mean anything (…). Maybe the sciences should just collect lots and lots of data and try to develop the results from them. Well if someone wants to try that, fine. (…). My judgment, if you like, is that we learn more about language by following the standard method of the sciences. The standard method of the sciences is not to accumulate huge masses of unanalyzed data and to try to draw some generalization from them (…). [Corpus linguistics] is an odd way of studying the sciences. Very radically different from the method of sciences. Which doesn't prove that it's wrong of course, but it certainly cannot justify itself on the grounds that it's being empirical. It's not what the concept of empirical has ever meant in the natural sciences. (Chomsky, 2004: 97)

À la lumière des critiques des génératives, considérons les trois niveaux auxquels nous situons les limites de la linguistique de corpus. En premier lieu, il est vrai qu'un corpus peut contenir de nombreuses erreurs de performance qui peuvent biaiser le jugement du linguiste et fausser les conclusions qu'il peut en tirer. Se pose alors une question fondamentale qui est la différenciation que devra effectuer le linguiste entre *la faute* liée à la performance, et *l'erreur* liée à la compétence selon Doca (1981). Mais quand bien même cette différenciation serait réalisée, que faire des erreurs, comment les aborder autrement que par le biais d'une théorie linguistique prévalant sur l'analyse du corpus et comment faire en sorte qu'elles ne faussent pas les conclusions d'une analyse ? À la critique de Chomsky (1965 : 8) qui déclare que ces erreurs « dégénéraient la qualité du langage », Labov (1969 : 201) répond que « the great majority of utterances in all contexts are complete sentences », et McEnery & Wilson (2001 : 16) affirment que la plupart des corpus ne comportent qu'une part négligeable d'erreurs. En outre, comme nous l'avons rapporté plus haut, l'intuition introspective n'est pas elle-même exempte de produire des énoncés erronés : le linguiste qui jugerait d'énoncés forgés par introspection se verrait ainsi inscrit dans un cercle vicieux.

Le deuxième niveau auquel se situent les limites de la linguistique de corpus est celui du biais des données en raison de leur nature prédéfinie. La constitution d'un corpus, considérons-le ici comme un ensemble d'exemples, est un processus et non une simple collecte. Ce processus nécessite une réflexion méthodologique en amont, indubitablement indicatrice d'une vision de la langue. Ainsi, Milner (1989 : 129) considère qu'un exemple forgé ne diffère pas d'un exemple tiré d'un corpus, ce dernier étant tributaire de la théorie méthodologique du corpus : « Tout exemple de langue, en tant qu'il permet un raisonnement linguistique suppose déjà un raisonnement linguistique ». Les données d'un corpus, pour les raisons avancées, ne seront jamais neutres, mais décidées en amont par les délimitations du corpus : on n'y trouvera pas autre chose que ce qu'on y a mis. Le linguiste qui constitue un corpus n'inclut pas objectivement toutes les données qui sont à sa disposition, mais opère des choix liés à sa volonté de représenter ce qu'il recherche : ces jugements reposent autant sur l'intuition que les jugements introspectifs. Outre le préformatage théorique des données, Cori & David soulignent que celles-ci sont le plus souvent tributaires des conditions matérielles de sélection d'un corpus :

> Les auteurs qui considèrent que la linguistique fondée sur les corpus est beaucoup plus objective que la linguistique qu'ils qualifient d'introspective oublient qu'il y a de l'arbitraire lors de la constitution des corpus. Tout ce qui est attesté ne peut entrer dans

> un corpus, si grand soit-il. Dans certains cas, on fait entrer dans le corpus les données qui sont disponibles, parce qu'elles ont été stockées sur des supports électroniques, le plus souvent pour de tout autres raisons que celles qui motivent la recherche linguistique. (Cori & David, 2008 : 125)

Concrètement, un corpus, aussi grand soit-il, est un objet fini et ne comporte qu'une partie infime du langage ; à moins de constituer un corpus très spécialisé, il est impossible pour la science actuelle de regrouper en un seul corpus toute forme ou structure d'un langage. Cependant, ces critiques légitimes tiennent pour l'introspection et l'élicitation, et ne concerne donc pas la méthodologie des corpus, mais concrétise un souci largement partagé, celui de l'exhaustivité, qu'ont voulu résoudre les génératives en proposant la grammaire universelle.

La réponse des linguistes de corpus est celle de la notion de représentativité des corpus, qui est devenue un objectif essentiel dans la constitution de tout corpus, et qui vise à lui conférer une neutralité dans la représentation du langage. En raison de l'importance de cette notion, qui est d'ailleurs l'un des critères que nous proposons comme définitoires des corpus, nous consacrerons à la représentativité des corpus une analyse détaillée[134].

Enfin, le dernier type d'argument porté envers la linguistique de corpus concerne son incapacité à la formalisation de principes linguistiques. La linguistique d'introspection formalise des règles, dont la somme peut former une grammaire. La linguistique de corpus prise au pied de la lettre se limite à une accumulation d'exemples ou de statistiques, et les généralisations ne peuvent être qu'hasardeuses tant que nous ne sommes pas certains de la représentativité absolue et évidemment utopique du corpus (Biber, 1993a, 1993b). La « linguistique de corpus pure » (Laporte, 2008), c'est-à-dire se limitant à l'extraction d'exemples et au comptage statistique, est incapable de formaliser. La linguistique de corpus ne doit pourtant pas se limiter à « une vocation patrimoniale », selon le terme de Mayaffre (2005 : 2), qui s'éloigne d'un objectif essentiel de la linguistique : théoriser la langue. Ceci ne peut évidemment se faire sans avoir recours, à un moment ou un autre, à un jugement intuitif et par définition subjectif. La linguistique est une science de l'humain et ne peut aboutir, à l'instar des sciences dures, à des vérités immuables.

L'intérêt ou la réponse de la linguistique de corpus à cette problématique du recours à l'intuition à un moment donné est la possibilité de vérification des hypothèses dans les corpus

[134] Cf. 2.6.

constitués, mais aussi le refus d'une prise de position qui interdirait le recours à l'intuition linguistique. Dans un article demeuré célèbre, Fillmore (1992) caricature les deux approches dans leurs représentations extrêmes. L'approche introspective est représentée par le « armchair linguist » qui, confortablement assis dans son fauteuil, rêve à la nature réelle du langage et ne sort de sa somnolence qu'à la découverte d'une vérité nouvelle. L'autre, le linguiste de corpus, possède un corpus monumental, « a corpus of approximately one zillion running words », et s'échine à déterminer « la fréquence relative des onze parties du discours des premiers mots des phrases vs le second mot des phrases ». Ces deux linguistes ne se parlent pas, le premier considérant que les affirmations du second sont inintéressantes, le second arguant que ce qu'avance le premier est invérifiable. La caricature de Fillmore est une plaidoirie en défaveur de la dichotomie entre introspection et linguistique de corpus, ceci éloquemment formulé de la sorte :

> I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body. (Fillmore, 1992 : 35)

Cette position est aussi celle de Leech (1991b) : dans une rétrospective sur la linguistique de corpus, il plaide pour une interaction homme-machine qui se manifestait à l'époque par la nécessité d'annoter les corpus. Le point intéressant est que, selon ses considérations, ni « la linguistique de corpus pure », évoquée plus haut, ni la linguistique d'introspection, qui rejeta totalement les données d'un corpus, ne sont en mesure de fournir des analyses satisfaisantes du langage. Nous retrouvons ce genre de mise en garde dans la plupart des domaines exploitant la linguistique de corpus. C'est le cas, par exemple, dans le domaine de l'acquisition du langage, où Morgenstern rapporte les indéniables apports des nouvelles technologies qui ont permis la constitution de corpus oraux à la fois longitudinaux et transversaux, mais souligne le danger qu'il y a à la simple exploitation de ces corpus :

> Les nouvelles technologies ont eu un apport immense sur les connaissances en acquisition. Mais le scientifique peut alors perdre la capacité à avoir des intuitions ancrées dans le quotidien puisqu'il est face à des enfants virtuels avec lesquels il n'aura pas noué de liens. (Morgenstern, 2009 : 70)

De la même manière, en lexicographie, Béjoint déplore quelque peu l'actuel état des choses, qui a modifié le travail des lexicographes, et qui ne laisse plus de place « à l'intuition et aux bonnes facultés d'analyse » :

> Indirectement, l'informatique aura donc contribué à la disparition des grands noms de la lexicographie, en réduisant le champ d'application de leur génie : tout rédacteur de dictionnaire sait désormais qu'une proportion importante de ce que le dictionnaire contiendra est fixé, par avance, par l'application mécanique des outils mis à sa disposition, sans qu'il y puisse rien changer. (Béjoint, 2007 : 20)

De son côté, Cappeau, dans une rétrospective sur les corpus oraux de ces trente dernières années en France, évoque le travail titanesque des années 1980 qui imposait au linguiste la lecture de centaines de pages de corpus ; il en résultait que le linguiste « connaissait son corpus »[135], pour l'avoir lu, ce qui n'est plus systématiquement le cas, comme mentionné par l'auteur avec Gadet :

> Avec une conséquence inattendue : le corpus, dès qu'il dépasse une certaine taille, est rarement connu en profondeur. Ce qui est utilisé se limite souvent à des ensembles de lignes qui sont le produit des requêtes formulées. Le plaidoyer en faveur de données réelles et contextualisées débouche de fait sur une vision fragmentée, partielle, de suites de lignes discontinues. On retrouve alors une coupure que l'on avait voulu croire abolie par les grands corpus. (Cappeau & Gadet, 2007a : 102)

Enfin, Laporte (2008) voit dans les traditions apparemment opposées d'une linguistique de corpus et d'une linguistique d'introspection des méthodes complémentaires qui ne s'excluent pas, mais se complètent. Une posture strictement observationnelle ne suffira pas à la formalisation de règles mais ces dernières nécessitent bien une observation intensive et méthodologiquement rigoureuse[136].

Quelles que soient les limites de la linguistique de corpus, elles semblent surmontables et l'usage tend à le prouver. McEnery & Wilson (1997 : 5) témoignent de la vivacité du domaine et Meyer (2002 : 1) rapporte que les barrières entre linguistes de corpus et linguistes d'introspection sont de moins en moins marquées : les premiers ne sont pas uniquement des

[135] Cette expression a été employée par Paul Cappeau lors de sa présentation « Corpus, ô mon beau corpus : regard (critique) sur l'apport des corpus » lors du colloque « (Dés)organisation de l'oral », 24 et 25 mars 2011, LIDILE, Université Rennes 2.

[136] Laporte présente la méthodologie du lexique-grammaire, méthodologie et théorie linguistique à la fois, comme un exemple concret et productif.

descriptivistes obnubilés par les analyses quantitatives, et de nombreux grammairiens génératives ont montré un intérêt pour les données linguistiques en tant qu'outil d'analyse et de vérification, tel que le notent également Legallois & François :

> Les théories les plus récalcitrantes aux explications empiriques, telle l'approche générative, ne sont plus autant réticentes qu'autrefois à raisonner sur des données authentiques, même si l'usage, entendu comme performance, n'est pas considéré comme un objet pertinent. (Legallois & François, 2011 : 7)

Le temps n'a fait que confirmer le statut du corpus dans la linguistique et le TAL : la traduction automatique, la reconnaissance optique de caractères et vocale, la correction orthographique, la classification automatique de documents, pour ne citer qu'eux, sont autant de champs d'applications où les corpus de données linguistiques représentent un outil de travail essentiel. En outre, ces champs d'applications dont le grand public ignorait la connaissance jusqu'à la fin des années 1990, se sont largement démocratisés et sont devenus un outil de travail quotidien.

2.2.3 La linguistique de corpus : théorie linguistique ou simple méthodologie ?

Nous avons examiné dans les sections précédentes les limites et les atouts de la linguistique de corpus et les divergences qui ont pu opposer les linguistes quant au statut du corpus et de l'approche empirique du langage. Toutefois, les partisans d'une approche empiriste ne constituent pas une communauté homogène sur le plan théorique et deux courants majeurs sont à présenter. Le premier courant, appelé corpus-based et principalement représenté par Geoffrey Leech, considère que la linguistique de corpus « refers not to a domain of study, but rather to a methodological basis for pursuing linguistic research» (Leech, 1991a : 105) ; pour le second courant, appelé corpus-driven et constitué autour de John Sinclair, la linguistique de corpus constitue au contraire un domaine à part entière et non une simple méthodologie de travail. Ce second courant est aussi appelé « école de Firth » ou « école firthienne »[137].

[137] Ce courant est ainsi nommé en référence à John Rupert Firth (1890 – 1960), linguiste anglais dont l'influence sur le courant corpus-driven est déterminante dans le sens où lui et Harold E. Palmer développèrent la notion de « collocation », qui est une notion essentielle dans le courant firthien comme nous allons le voir. Notre propos est ici très réducteur, mais les origines de la London School et du contextualisme de Firth ne concernent pas

Par exemple, au sein du courant corpus-based, Kennedy (1998 : 7) écrit qu'il serait erroné de supposer que la linguistique de corpus est une théorie linguistique, au même titre que par exemple la linguistique transformationnelle, ou qu'elle serait une branche nouvelle de la linguistique. Granger (2002 : 4) affirme quant à elle que la linguistique de corpus n'est « ni une nouvelle branche de la linguistique, ni une nouvelle théorie du langage, mais (…) une méthodologie particulièrement efficace », à l'instar de Meyer (2002 : xi), qui y voit « une façon de faire de la linguistique ». Il en va de même pour McEnery & Wilson (2001 : 2) qui considèrent que la linguistique de corpus n'est pas une branche de la linguistique telle que le serait la syntaxe, la sémantique, la sociolinguistique ou autre. Ces disciplines se focalisent sur la description et l'explication des faits de langue alors que la linguistique de corpus est une méthode empirique, et non pas un aspect de la langue requérant une explication ou une description. En ce qui concerne le courant corpus-driven, l'ouvrage de Sinclair (1991) est fondateur, et celui de Tognini-Bonelli (2001) souligne les spécificités du courant par rapport aux autres approches en linguistique de corpus, où l'auteure affirme que « la linguistique de corpus va beaucoup plus loin qu'un rôle exclusivement méthodologique » (Tognini-Bonelli, 2001 : 205).

Si les auteurs de ces deux courants définissent la linguistique de corpus différemment, il faut s'interroger sur ce qui différencie véritablement la conception corpus-based de la conception corpus-driven. En premier lieu, les deux courants diffèrent dans leurs objectifs, ou dans les chemins que les centres de recherches ont privilégiés. Les principales préoccupations de l'école firthienne restent l'enseignement des langues et la lexicographie[138], tandis que celles du courant corpus-based s'orientent davantage vers la création d'outils d'analyse de corpus et le TAL : Williams (2006 : 156) précise toutefois qu'il ne s'agit pas « d'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 ». En effet, McEnery & Wilson (2001 : 2), qui se positionnent pour la linguistique de corpus en tant que méthodologie, concèdent qu'il y a « une théorie syntaxique basée sur corpus en

notre sujet. Les articles de Williams (2006) et de Léon (2008) détaillent davantage l'aspect épistémologique du courant corpus-driven.

[138] Nous rappelons que Sinclair est à l'origine du projet COBUILD qui aboutit à la constitution d'un dictionnaire pour apprenants sur base de l'analyse d'un corpus. Le lien entre enseignement des langues et corpus concerne donc ici l'exploitation des corpus pour la création d'outils didactiques.

opposition à une théorie syntaxique non basée sur corpus, une sémantique basée sur corpus en opposition à une sémantique non basée sur corpus etc. », et que les barrières ne sont donc pas aussi tranchées entre corpus-based et corpus-driven.

Sur le plan théorique, l'option corpus-driven postule qu'aucune position théorique *a priori* ne doit présider aux observations sur corpus. Williams décrit dans l'introduction de sa thèse la démarche comme suit :

> L'approche adoptée ici est celle de la *corpus-driven research*, une recherche dirigée par le corpus. Autrement dit, au lieu de partir de théories toutes faites et de les tester sur le corpus, nous explorons le corpus pour identifier des régularités que nous pourrions exploiter par la suite. (…) nous laissons parler les mots à travers le corpus. (Williams, 1999)

Dans la tendance corpus-based, le corpus est au contraire vu comme un réservoir d'exemples destiné à tester ou vérifier des positions théoriques préexistantes. Mayaffre (2005 : 5) définit ainsi le corpus du courant de Leech comme « un *observatoire* d'une théorie *a priori* » et le corpus firthien comme « un *observé dynamique* qui permet de décrire puis d'élaborer des modèles *a posteriori* ». De façon plus détaillée, Tognini-Bonelli (2001 : 65) discute de cette opposition en disant que la tradition non firthienne utilisait les corpus afin de : « test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study ».

Ces deux positionnements théoriques ont leur influence directe sur la constitution des corpus. Selon Tognini-Bonelli (2001 : 65), la théorisation des corpus du courant corpus-based passe par l'isolation, la standardisation et l'instanciation des corpus, démarches reflétées dans l'annotation des corpus : en résumé, toute annotation déneutralise les données et les teinte d'une théorie linguistique. L'école firthienne, elle, propose donc de considérer les données brutes et intègres, et de ne s'en tenir qu'aux évidences dégagées de ces données vierges. C'est ainsi que Sinclair (2005) s'oppose à tout type d'annotation pour une raison simple : une annotation est l'ajout d'informations aux données brutes et cet ajout ne peut être qu'interprétatif, et donc synonyme d'une théorie linguistique. Le corpus vierge et non annoté constitue ainsi le premier trait distinctif du corpus firthien.

En second lieu, puisque le corpus rôle du corpus firthien est de permettre l'identification de régularités, le mot d'ordre est donc la constitution de corpus géants qui, eux seuls, sont

représentatifs cumulativement, à l'instar du corpus de la BoE dans lequel, nous le rappelons, était impliqué John Sinclair. En outre, la structure du corpus est elle-même concernée, car pour Sinclair, les notions de représentativité et d'équilibre ne sont pas des objectifs atteignables par échantillonnage textuel ou démographique, mais des objectifs probables corrélés à la taille du corpus[139]. Le corpus ouvert et le plus grand possible est le second trait du corpus firthien.

Le corpus firthien, de taille importante et non annoté, est nécessaire afin de pouvoir travailler sur la notion de collocation, « pierre angulaire de la linguistique de corpus » selon Teubert (2010). La collocation « est essentiellement, dans la pratique linguistique, une manifestation de solidarités lexicales, que la lexicologie – et non la grammaire – se doit de mettre en évidence »[140]. Par exemple, Teubert (2010) analyse 450 articles parus après le 11 septembre 2001 et contenant le syntagme « civilised world », et rapporte la liste des collocats[141] sous forme de résumé :

- The *civilised world* is peaceful, legitimate, dignified and solemn, his fundamental responsabilities, is full of innocent people asking for freedom, democracy, reason, and tolerance.
- Its (sworn) enemies are : terrorism, (irrational) fanaticism, (global) terror, terrorists, fanatics, suicide murderers, agressive dictators, Muslims militants, the Muslim world, religious extremists, religious hatreds, bigotry, this cancer.
- The *civilised word* is under (ruthless) attack, assault, terrorist attacks, but when stirred to anger and to action it will wage a (titanic) struggle for freedom and security, it twill bring its ennemies to account, punish them and hound these people down.

[139] Nous discuterons en détail de la représentativité en 2.6, en l'occurrence en présentant la méthode de Sinclair et en l'opposant à la méthode d'échantillonnage *a priori*.

[140] La collocation est un « concept difficile à formaliser et [qui] ne peut être étudié que par rapport à des exemples prototypiques » (Williams, 2001 : 9). Notre propos n'est pas d'approfondir cette question en particulier, mais de l'aborder afin de comprendre les enjeux méthodologiques de la linguistique de corpus firthienne. Les notions de collocation et de colligations sont approfondies par Legallois (2012).

[141] Teubert définit le collocat comme un item co-occurrent dont la fréquence est significative dans le contexte immédiat du *nœud* (en l'occurrence *civilised world*), c'est-à-dire dans un intervalle de cinq items à gauche et cinq items à droite du nœud.

Arguant que pas un seul dictionnaire ne contient l'item « civilised world », Teubert présente les analyses statistiques sur de grands corpus comme étant les seules capables de révéler des cooccurrences signifiantes des « items lexicaux »[142]. Le sens d'un item lexical ne lui est ainsi pas intrinsèque, mais s'interprète par l'analyse de ses collocations.

Les applications de la notion de collocation – et donc de la linguistique de corpus firthienne – sont importantes dans le domaine de l'enseignement des langues. En 1942, paraît au Japon *Idiomatic and Syntactic English Dictionary* (Honby *et al.*, 1942), publié en Angleterre en 1948 sous le nom de *Oxford Advanced Learner's Dictionary of Current English* puis *Oxford Advanced Learner's Dictionary*. La notion de collocation, appliquée à la lexicographie pour apprenants de langue étrangère, a permis la conception de ce qui deviendra la norme des dictionnaires unilingues pour apprenants, basés sur des fiches offrant un large panel de locutions ou constructions de tout type grammatical. Les études rapportant les bénéfices de la collocation dans le développement des compétences en langues secondes et étrangères sont nombreuses, notamment, selon McEnery & Xiao (2011), dans la compréhension des usages réels des items lexicaux ou grammaticaux, dans le fait que la compétence des apprenants est proportionnelle à leur réservoir d'unités préfabriquées, ou que l'apprentissage de ce type d'unité accélère la progression linguistique de l'apprenant en compréhension et en expression[143].

Les courants corpus-based et corpus-driven ne s'excluent pas, mais ont des approches différentes du corpus, essentiellement dans la méthodologie de constitution et d'analyse. Nous terminerons en soulignant un point ayant justement trait à la méthodologie et qui concerne également l'enseignement des langues. Nous venons de voir que le courant firthien était concerné par la didactique des langues, mais uniquement dans la création d'outils destinés aux apprenants ou dans le relevé de collocations. En ce qui concerne les corpus constitués de productions orales ou écrites d'apprenants, Sinclair (2004) considère que ce type de corpus est un corpus spécialisé, et que l'interlangue est un langage trop « extrême » pour être inclus dans un corpus de référence. Il va d'ailleurs de soi que le linguiste ne peut aborder des productions

[142] L'item lexical peut être, selon Teubert (2010), « un ou plusieurs mots, parfois même, comme dans le cas de certains idiomes, une phrase entière ».

[143] McEnery & Xiao (2011) détaillent l'ensemble des études qui démontrent l'intérêt des collocations dans l'enseignement des langues.

d'apprenants que par un biais normatif, et donc basé sur une théorie linguistique, soit une approche corpus-based.

2.3 Le corpus en tant qu'objet linguistique et ses constituants

Tel que le rapporte Rastier (2005), et tel que le prouve l'étymologie du mot « corpus » dans « Trésor de la langue française informatisé » (TLFI)[144]:

> **Étymol. et Hist. 1.** Fin XII^e^ s. *corpus Deu* « hostie » (*Mort Garin,* 132 ds T.-L.); 1206 *corpus Domini* (GUIOT, *Bible,* 1223 ds GDF.) -1584, Benedicti ds *Fr. mod.,* t. 5, 1937, p. 73; *cf.* encore *corpus* « *id.* » (1642, Oudin ds *DG* -1771, *Trév.*); **2.** 1863 *corpus juris, corpus* « collection du droit romain » (LITTRÉ); **3.** 1890 « collection d'inscriptions de l'Antiquité » *(DG)*; **4.** 1961 « ensemble d'énoncés servant de base à l'analyse linguistique » *(Lar. encyclop.).* Au sens 2, empr. au lat. class. *corpus juris,* v. *corps.* Sens 3 et 4 développés en fr. à partir de 2. Le sens 1 est empr. au lat. chrét. *corpus Domini, corpus Christi,* désignant l'Eucharistie.

Nous constatons que le terme « corpus » apparaît dans les domaines religieux (XIII^ème^ siècle), juridique (XIX^ème^ siècle), littéraire[145] (XX^ème^ siècle) et enfin linguistique en 1961. Jusqu'aux années 1950, les corpus sur lesquels travaillaient les linguistes n'étaient pas nommés ainsi ; par exemple, les conversations qu'enregistrent Gougenheim *et al.* pour élaborer *Le Français fondamental* sont appelées « enquêtes ». Laks fait remonter le terme jusqu'à Justinien (527-565) qui fit compiler le *Corpus Juris Civilis* :

> Recueil à vocation exhaustive qui contenait les constitutions impériales, un manuel de droit et l'ensemble de la jurisprudence commentés. En rappelant que le corpus de Justinien faisait pendant au *Corpus Juris Canonici,* on se souvient de ce que la notion de corpus doit à la pensée théologique, au moins dans les religions du Livre. (Laks : 4)

Évoquant la Torah, les Évangiles dits *synoptiques* ou les Hadiths en Islam, l'auteur les considère en tant que « vastes corpus structurés, clos, stables et publiquement acceptés, certes

[144] Le TLF ne définit pas les mots à partir de la sémantique, ni de l'étymologie ; il entend répertorier les significations d'usage à partir d'exemples attestés dans la littérature française des origines modernes à nos jours.

[145] Rastier (2005) remarque avec pertinence que les domaines de la philologie et de l'herméneutique, qui ont élaboré la notion de corpus, ont été « des disciplines injustement oubliées, du moins dans le domaine des Traitements automatiques du langage ». Ceci alors que l'évolution du TAL, comme nous l'avons vu, fut étroitement liée à la linguistique de corpus.

à vocation herméneutique ou religieuse ». En citant plusieurs exemples dont la table de Mendeleïev (tableau périodique des éléments) ou la taxinomie réalisée par Darwin suite à son voyage, Laks considère le corpus en tant que manifestation d'une méthodologie observationnelle (soit une méthodologie corpus-based), qui fait de la constitution d'un corpus et son analyse, la base de l'élaboration d'une théorie scientifique :

> De l'histoire naturelle de Buffon aux grands classements de Linné, l'accumulation des faits, des données et des descriptions est constitutive d'un classement raisonné qui fonde une première théorisation et une première modélisation. Adossée à d'énormes compendiums, la Science se dégage alors comme un raisonnement sur l'organisation des données, comme une contemplation, une θεωρία (*théoria)* conduite par la structuration interne des données (…) Le compendium s'analyse désormais en taxinomies raisonnées lesquelles constituent le socle même de toute théorisation et modélisation scientifique (…). On voit ainsi se dégager une ligne méthodologique très forte dans l'histoire des sciences : constitution de l'observable, taxinomie et systématique, théorisation et modélisation. Dans cette méthodologie *bottom-up*, la construction d'un corpus factuel joue, on le voit, un rôle considérable. (Laks, 2008 : 4-5)

Cette vision de la science, basée sur l'observation, le classement et l'analyse de faits – linguistiques si la science en question est l'étude du langage – est à opposer à celle de Chomsky qui ne voit pas dans l'histoire des sciences l'omnipotence des corpus :

> If you want to understand how bodies fall, Galileo would not have been interested in videotapes of leaves falling and balls going around and rocks rolling down mountains and so on and so forth. What he was interested in is the highly refined abstract conception of a ball rolling down a frictionless plane, which doesn't even exist in nature. (Chomsky, 2004 : 97)

Nous adhérons en partie avec la vision de Laks qui voit dans toute méthodologie scientifique la présence d'un ensemble de données constituantes d'un observatoire ; en partie car l'intuition scientifique et l'observation des faits ne sont pas des phénomènes exclusifs l'un de l'autre. Il en demeure que ce que Laks appelle « compendium », à savoir une somme de faits ou de connaissances, a toujours existé ; ces compendiums sont les ancêtres des corpus actuels, quels qu'ils soient. Les sous-parties qui suivent se proposent d'examiner les critères qui font d'une somme de faits linguistiques oraux, un corpus oral répondant aux exigences de la scientificité moderne.

En ce qui concerne le corpus linguistique, John Sinclair (J. Sinclair, Jones, Daley, & Krishnamurthy, 2004 : xix) rapporte l'anecdote suivante : au milieu des années 1960, Nelson Francis rendit visite à Randolph Quirk, à Londres. Francis était chargé du matériel volumineux contenant le Brown Corpus et dit en le déposant sur le bureau de Quirk : « Habeas corpus ». Léon rapporte que c'est plutôt Quirk et son entourage qui utilisèrent le terme dans sa dimension linguistique[146]. Dans les deux cas, il semble que le corpus linguistique ait été nommé ainsi vers le début des années 1960. En France, l'acceptation linguistique du terme apparaît dans les années 1970 ; Blanche-Benveniste (1997 : 87) rapporte que le terme est qualifié de « pédant » par Robert-Léon Wagner en 1973, dans son ouvrage *La Grammaire française*.

Mais qu'en est-il des représentations du terme ? Comme nous venons de le voir, le corpus a concerné une multitude de domaines scientifiques et, comme le note Schaeffer-Lacroix (2009 : 19), « l'apparition des corpus dans des domaines variés entraîne une hétérogénéisation et complexification du terme de 'corpus' ». Il est donc évident qu'il n'y a pas, et qu'il n'y aura pas, de consensus sur la définition du terme. Chaque ouvrage, chaque auteur et chaque projet propose une définition convergente avec les finalités de ses propres recherches, de manière explicite ou pas. La diversité des définitions provient de deux facteurs, le premier est d'ordre théorique, lié à la diversité des linguistiques de corpus ; chaque définition contiendra les visions théoriques des auteurs, qui sont plurielles. Charaudeau dit à ce propos :

> Les problèmes que pose la notion de corpus sont relativement bien connus pour avoir été longuement discutés dans le champ des sciences du langage, mais ils n'ont toujours pas donné lieu à un consensus qui aurait permis de s'en remettre à une définition faisant autorité et à laquelle on se référerait chaque fois que l'on aurait à justifier un corpus d'analyse. Cela est peut-être le symptôme de ce que le corpus n'existe pas en soi, mais dépend (…) du positionnement théorique à partir duquel on l'envisage. (Charaudeau, 2009 : 37)

[146] Léon (2008 : note 13) écrit en effet : « L'anecdote rapportée par Svartvik (2005) et reprise par Sinclair (2004), selon laquelle c'est Nelson Francis qui a utilisé le premier le terme de 'corpus' s'avère donc erronée. Outre l'expression latine « corpus inscriptionum » utilisée par Firth (1968 [1956]), c'est probablement Quirk qui, le premier, utilise le terme dans le sens qui nous intéresse ici. Il faut de plus souligner que le terme, déjà en usage chez les néo-bloomfieldiens, est disponible dans la communauté des linguistes et pour Quirk en particulier ».

Le second facteur est d'ordre pratique : outre la théorie, deux corpus ne seront jamais similaires en raison des divers contenus, locuteurs, types de corpus, époques concernées etc. Cette diversité se retrouve dans une définition proposée par Fisher *et al.* :

> Un corpus est un ensemble de textes (corpus textuel), un ensemble de mots (corpus lexical), ou un ensemble de phrases (corpus phrastique). Les textes sont sélectionnés selon des critères prédéfinis pour représenter, tant que possible, une langue, une variété d'une langue, un genre, un domaine de discours, un auteur, ou un sujet (…). On parle, par exemple, de : « corpus sémantique », « corpus en phonologie », « corpus en traductologie », « corpus électronique », « corpus littéraire », « corpus de sciences sociales », « corpus bilingues », « corpus en langues anciennes » pour ne reprendre ici qu'un échantillon. Les textes peuvent être analysés à un moment donné ou sur une période de temps (analyse synchronique vs. diachronique). (Fisher *et al.*, 2009 : 28)

Nous ne proposerons donc pas de définition du terme de « corpus », mais tenterons d'examiner quelques définitions dans la littérature afin d'y retrouver des critères que nous soumettrons à analyse.

Tout d'abord, un corpus n'est pas un ensemble de textes ou de productions orales compilés aléatoirement. Plusieurs critères font la différence entre « archives », « base de données », « banque de données » ou « portail » en ce qui concerne le WEB d'une part, et un corpus d'autre part. Nous reprendrons ici l'expression de Rastier (2005) pour qui un corpus n'est évidemment pas « un sac de mots ». Nous différencions l'accumulation de données du corpus à deux niveaux. Au niveau de la finalité de la compilation des données, et au niveau des critères de collecte de ces données. En ce qui concerne la finalité de la recherche, Leech (1991b : 11) et Kennedy (1998 : 4) avancent que la systématicité, la planification et la structuration de la collecte et des données ont pour objectif d'octroyer à une accumulation aléatoire de données une représentativité, afin d'en faire un corpus linguistiquement exploitable. Contrairement au corpus, une archive est, selon Kennedy (1998 : 4), « un répertoire de textes, souvent considérable et providentiellement compilé, habituellement non structuré ».

Ainsi, la tradition d'archivage des données sonores qui a commencé avec Ferdinand Brunot, a donné lieu tout au long de ces cent dernières années à une quantité d'archives sonores qui ne sont pas encore des corpus. Elles nécessitent un travail de tri, de documentation, de numérisation dans certains cas afin d'en faire un échantillon représentatif analysable linguistiquement. En outre, les données orales nécessitent la transcription, voire l'annotation

des données. Les critères qui différencient une accumulation de données d'un corpus constitué sont selon nous les suivants :

- le critère de la collection des données ;
- le critère du numérique ;
- le critère de la représentativité des données ;
- le critère de l'annotation des données ;
- le critère de la documentation du corpus.

L'ensemble de ces critères constitue un ensemble de protocoles dits protocoles de constitution. Ces protocoles sont le second niveau qui différencie une accumulation de données d'un corpus, et nous tenterons de démontrer pourquoi, selon une citation de Laks (2008) « lorsque l'on dispose d'une base de données sonores, le corpus correspondant reste à construire ».

2.4 Critère de la collection

Le premier critère qui apparaît régulièrement dans la littérature est celui de « collection » ou « ensemble », notion qui sous-entend que la constitution est une démarche scientifique qui implique l'établissement de protocoles de collecte indiquant systématiquement des objectifs linguistiques, si ce n'est une vision du langage. En ceci un corpus diffère des archives qui, pour reprendre l'expression de Kennedy ci-dessus, sont « providentiellement compilées ». Comme exemple, nous proposons les archives numériques audiovisuelles de l'Institut national de l'audiovisuel (INA), dont la numérisation a été entamée en 2006 en vue de sauvegarder les archives magnétiques de la dégradation. Les archives numériques ont été créées par nécessité et non selon des critères de collecte scientifiques ; si leur consultation en a été facilitée, elles n'en constituent pas pour autant un corpus exploitable, mais une source de constitution de corpus qui seront aussi variés que les protocoles de constitution qui les régiront selon les objectifs du corpus. En ce qui concerne les corpus oraux, ils devront détailler deux points principaux : les locuteurs cibles à enregistrer, ainsi que le choix du type de données à collecter. Delais-Roussarie rapporte à ce propos :

> La communauté linguistique considère, à la suite de Sinclair (1996), qu'un corpus est « une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon de langage ». D'après cette

> définition, un ensemble de données collectées ici et là sans réflexion préalable sur ce qui motive le rassemblement des documents n'est pas un corpus. (Delais-Roussarie, 2002 : 3)

Ces choix sont motivés par la structuration du corpus, et ce point sera discuté dans la section consacrée à la représentativité des corpus, car le choix des locuteurs et le choix des données sont deux facteurs principaux dans l'ensemble des paramètres qui influent sur le potentiel et la qualité de la représentativité d'un corpus.

2.5 Critère du numérique

La norme actuelle, en termes de constitution de corpus, est d'en effectuer la collecte sous forme numérique, ou de les convertir au numérique afin que l'archivage et l'analyse soient effectués par la machine informatique. Si le critère numérique est une norme (Kennedy, 1998 : 3), il ne fut pas, historiquement, un critère *sine qua none* du corpus oral ou écrit. Nous citons à ce propos Francis, dont le titre de la conférence qu'il présenta en 1991, « Language corpora BC » (corpus langagiers pré-informatiques), indique en lui-même que l'existence des corpus ne fut pas liée à l'émergence de l'outil informatique :

> I will confine myself to corpora accumulated before B.C., i.e. before the use of computers (...). Some seem to believe that there were no corpora before that. The truth is that many important corpora of English were assembled long before the computer was invented. (Francis, 1991 : 17)

Kennedy, un peu plus tard dans *An Introduction to Corpus Linguistics*, présentait, comme nous l'avons vu plus haut, le format numérique des corpus en tant que norme, mais ne pose pas non plus cette condition comme essentielle :

> Historically it is not even the case that corpora are necessarily stored electronically so that they can be machine readable, although this is nowadays the norm. (Kennedy, 1998 : 3)

La collecte sous forme numérique, ou la numérisation des données sont donc des étapes qui facilitent grandement leur manipulation, leur archivage, leur analyse et leur diffusion ou partage, mais un corpus peut être sous format papier pour les données textuelles ou format magnétique pour les données audiovisuelles. Cela constituerait néanmoins une perte de temps considérable au niveau de la constitution, mais causerait surtout des limitations réelles quant à

la pérennité et aux possibilités d'analyse[147] ; non seulement les corpus actuels sont majoritairement numériques, mais la tendance est également à la numérisation des corpus antérieurs à l'informatique. La raison pour laquelle des auteurs comme Kennedy ou Francis insistent sur le fait que les corpus ne sont pas exclusivement électroniques est leur volonté de ne pas déchoir les corpus pré-informatiques de leur qualité de corpus, et de les reconnaître en tant que tels.

Il semble néanmoins qu'une définition des corpus contemporains se doive d'inclure le critère numérique car les données non numériques ne sont pas encore obsolètes, qu'il s'agisse de données écrites ou orales. Dans la plupart des centres de langues, les productions écrites des apprenants restent des productions manuscrites, et certaines d'entre elles offrent des bases de données considérables potentiellement numérisables[148]. Les bases de données orales ne sont pas toutes numériques mais parfois effectuées à l'aide d'outils d'enregistrement magnétiques. Outre les limitations dans la réécoute, l'archivage et l'analyse des données auxquelles sont sujettes les données textuelles, les bandes son magnétiques courent le risque de détérioration. La numérisation offre ainsi des avantages considérables par rapport aux corpus papier ou magnétiques, que nous résumons ainsi :

- limitation des erreurs ;
- recherches d'occurrences et de concordances fiables et presque instantanées ;
- reproductibilité précise des analyses effectuées ;
- possibilité de manipuler des quantités considérables de données.

Nous terminerons en précisant que la démocratisation massive des ordinateurs avait poussé Leech, qui avait noté la relation étroite entre « Corpus Linguistics » et informatique, à proposer le terme de « Computational Corpus Linguistics » (1991a : 106).

[147] Par ailleurs, l'archivage de données sur papier ou sur bande magnétiques pose un problème réel d'un point de vue physique. L'archivage de données considérables nécessiterait des locaux, sans oublier les contraintes (coûts, sécurité, transport etc.) qui en découlent.

[148] Par exemple, le Département de Lettres à l'Université de Damas possède des milliers de copies d'apprenants syriens en anglais, français ou japonais, provenant de leurs évaluations durant leur formation. Les fiches des épreuves et les formats d'évaluation pourraient contribuer à la constitution de protocoles de constitution. Les sélections nécessaires à la constitution d'un corpus, voire de corpus longitudinaux, ne pourrait se faire sans l'aide de l'outil informatique.

2.6 Critère de la représentativité

La représentativité des données se résume par le potentiel d'un échantillon du langage, en l'occurrence un corpus, à représenter des vérités linguistiques générales. Afin de définir en détail ce critère, nous tenterons de comprendre l'importance de cette notion et de sa récurrence au sein de la littérature traitant de la linguistique de corpus. Nous distinguerons ensuite les différentes méthodologies employées dans le but d'atteindre la représentativité dans la constitution de corpus. Les deux principaux courants méthodologiques que nous examinerons sont ceux de la « stratification en amont » représenté par Biber (1993a, 1993b) pour le premier, et celui des « monitor corpus » représenté par Sinclair (1991, 1996, 2004) pour le second. Nous nous intéresserons en détail à la question de la taille des corpus, et nous conclurons par une revue rapide de la situation actuelle, accompagnée de quelques recommandations destinées aux compilateurs de corpus constitués ou futurs.

Dès 1982, Francis (1991 : 17)[149] définit un corpus comme « un ensemble de textes censé être représentatif d'un langage, d'un dialecte ou autre sous-catégorie du langage donnés, et destiné à être analysé linguistiquement »[150]. Comme le note Pearson, la prise en considération du critère de la représentativité est croissante depuis la définition de Francis :

> Francis' definition, expressed in 1982, would now be considered to be too vague because it is not sufficient to state that texts are "assumed" to be representative. If representativeness is considered to be an important criterion, then the means of achieving it should be explicit. (Pearson, 1998 : 43)

Le critère de la représentativité est donc ensuite évoqué par la quasi-totalité des ouvrages et articles de référence sur la linguistique de corpus[151]. Les raisons de la systématicité de la notion de représentativité tiennent dans les objectifs de cette dernière ; Leech (1991b, n. 9) ou les statisticiens Manning & Schütze (1999 : 119) suggèrent qu'un corpus est représentatif si

[149] La définition est reprise par Francis dans son propre article de 1991. En ce qui concerne l'article de 1982, voir note n° 131.

[150] Elena Tognini-Bonelli (2001) adopte la définition de Francis dans *Corpus linguistics at work*.

[151] Cf. Aijmer & Altenberg, 1991; Biber, 1993a; Fillmore, 1992; Francis, 1991; Benoît Habert, Nazarenko, & Salem, 1997; Kennedy, 1998; Leech, 1991b; McEnery & Hardie, 2011; McEnery & Wilson, 2001; Meyer, 2002; J. Sinclair, 1991; Teubert, 2010; Tognini-Bonelli, 2001.

les conclusions basées sur l'analyse du corpus peuvent être généralisées à l'ensemble du langage étudié ; cette idée se retrouve chez Sinclair qui évoque les corpus de référence :

> A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials. (Sinclair, 2004)

La représentativité est donc, avant tout, liée à un souci de scientificité des corpus : résultats fiables, pouvant être exploités et généralisés à l'ensemble du langage. Leech résume l'importance de la notion de la représentativité de la sorte :

> Without representativeness, whatever is found to be true of a corpus, is simply true of that corpus – and cannot be extended to anything else. (Leech, 2006 : 135)

Or ce constat rejoint directement les premières critiques envers la linguistique de corpus que nous avons analysées précédemment, il convient donc d'évoquer la faisabilité de l'objectif. La question de la représentativité peut en effet être rattachée aux critiques de Chomsky qui voyait en tout corpus, fussent-ils aussi grands que le BNC, l'ANC ou le BoE[152], une portion infime du langage et donc une représentation biaisée de celui-ci. Comme nous l'avons vu, ces critiques sont légitimes et ne stoppèrent pas les recherches sur corpus, et ne furent pas non plus ignorées de la part de l'ensemble des théoriciens. La réponse quasi-unanime à ces limitations pourrait tenir en un seul mot : l'échantillonnage, procédé inhérent à la création de corpus, tel que le formule Sinclair :

> Everyone seems to accept that no limits can be placed on a natural language (…). Therefore no corpus, no matter how large, how carefully designed, can have exactly the same characteristics as the language itself. Fine. So we sample, like all the other scholars who study unlimitable phenomena. We remain, as they do, aware that the corpus may not capture all the patterns of the language, nor represent them in precisely the correct proportions. (Sinclair, 2004)

C'est donc dans l'échantillonnage que réside l'accès à la représentativité, que nous illustrerons sur deux axes :

[152] Cf., respectivement, 2.9.2, 2.9.3 et 2.10.1 pour ces trois corpus.

1) un axe horizontal, qui concerne les représentations du langage : représentativité des trois médiums, écrit, audio et audio-visuel ; représentativité des types de discours ; représentativité des variations sociolinguistiques ; représentativité des langues dans le corpus (productions de locuteurs natifs, d'apprenants, d'enfants, de langue pathologique)[153]. Les critères cités donnent lieu à des catégories pouvant être affinées elles-mêmes en sous-catégories selon des méthodologies que nous allons présenter ;

2) un axe vertical, qui concerne la représentativité induite par un nombre suffisant de tokens et de mots-types, soit la taille du corpus.

Ainsi la représentativité ne dépend pas uniquement du nombre de mots, mais également des catégories choisies, du nombre d'échantillons au sein de chaque catégorie et de la taille de chaque échantillon.

Un manque de représentativité sur l'axe horizontal induit ce que Biber (1993b : 219) nomme « bias error », et que Habert (2000) traduit par « déformation ». Une déformation survient quand les caractéristiques linguistiques du corpus ne correspondent pas à celles de la population visée. Par ailleurs, si le corpus n'est pas suffisamment représenté verticalement, Biber parle alors de « random error », « incertitude » selon Habert. L'incertitude est due au fait que le corpus est trop petit pour que les conclusions qui en sont tirées soient généralisables.

La représentation verticale n'est concernée que par la taille du corpus. Biber (1993a : 243) souligne que l'idée première des chercheurs qui abordent la notion de représentativité concerne la représentation verticale, à savoir le nombre de textes ou d'échantillons à inclure afin que le corpus soit représentatif. Or, Biber rapporte que la représentation verticale n'est pas la considération la plus importante dans le processus de sélection des échantillons : les questions de définitions de la nature des textes en ce qui concerne les corpus écrits, et des populations cibles en ce qui concerne les corpus oraux, sont des considérations non seulement

[153] Ceci bien que ces corpus soient considérés par Sinclair (1996) en tant que « corpus spéciaux », et ne contribuant pas à la description du « langage ordinaire », car contenant trop de faits inusuels : « Corpora of the language of children, geriatrics, non-native speakers, users of extreme dialects and very specialised areas of communication (…) should also be designated special corpora because of the unrepresentative nature of the language involved. ».

plus importantes, mais également plus difficiles à prendre en compte. En effet, considérons l'exemple de deux corpus fictifs :

Corpus 1 : Horizontalement représentatif, nombre de strates protocolairement délimité à x. Le corpus 1 est lacunaire verticalement ; autrement dit, le nombre d'échantillons au sein de chaque strate n'est pas suffisant. Concrètement, il pourrait s'agir d'un corpus visant à représenter la langue parlée en France, et qui inclurait un seul enregistrement pour chaque catégorie : une seule représentation par catégorie socioprofessionnelle, par sexe, par âge, par éducation linguistique etc. Au sein de ce corpus le nombre de locuteurs serait équivalent au nombre de strates x.

Corpus 2 : Verticalement représentatif, le nombre de locuteurs est suffisamment important pour que les conclusions tirées de l'échantillon soient généralisables à ce que représente l'échantillon. Le corpus 2 est lacunaire horizontalement ; il ne représente qu'une seule strate de la population. Concrètement, le corpus 2 pourrait être un corpus de x enregistrements d'une seule catégorie de personnes : elles auraient toutes le même âge, seraient du même sexe, de la même catégorie socioprofessionnelle etc.

Il est évident que dans une telle configuration, la remédiation à la représentativité globale du corpus 1, qui possède déjà ses catégories, est plus aisée que la remédiation à la représentativité globale du corpus 2. Dans le premier, il suffirait d'augmenter le nombre d'échantillons, alors que le second nécessite le travail de stratification de la population étudiée en catégories sociolinguistiques qui ne font pas consensus.

Du point de vue de la représentativité, le processus de constitution d'un corpus peut être conduit selon deux méthodologies différentes : soit les catégories sont prédéfinies *a priori*, selon des segmentations que d'aucuns qualifient d'arbitraires, soit il est considéré que ces catégories ne peuvent être déduites que justement grâce à un corpus. Ce sont ces deux méthodologies que nous allons détailler; la première est celle décrite dans un article de Biber abondamment cité, « Representativeness in corpus design » (1993a), la seconde est la méthodologie décrite généralement dans les travaux de Sinclair (1996, 2004). Nous terminerons ensuite par la question du deuxième axe, soit la représentativité liée à la taille du corpus, puis en dressant un état des lieux actuel.

2.6.1 Stratification en amont

Comme nous l'avons évoqué, Biber (1993a) propose une constitution de corpus se basant sur une catégorisation en amont, puis la constitution du corpus selon la catégorisation obtenue. Or la stratification d'un corpus ne repose pas sur un type de paramètres unique, mais sur une liste plus ou moins modifiable de paramètres qui peuvent par ailleurs s'entrecroiser. Le schéma de catégorisation requiert donc l'inventaire de ces paramètres, ainsi qu'une hiérarchie entre eux. La population étudiée est donc divisée en catégories (strates), et les catégories choisies sont pourvues. Les problématiques concernant le nombre de catégories et la quantité de données à pourvoir au sein de chacune d'entre elles trouvent leurs réponses avant la constitution du corpus. Biber (1993a : 245) propose comme exemple une hiérarchie d'échantillonnage, qu'il qualifie comme « a reduced set of sampling strata, balancing operational feasibility with the desire to define the target population as completely as possible » :

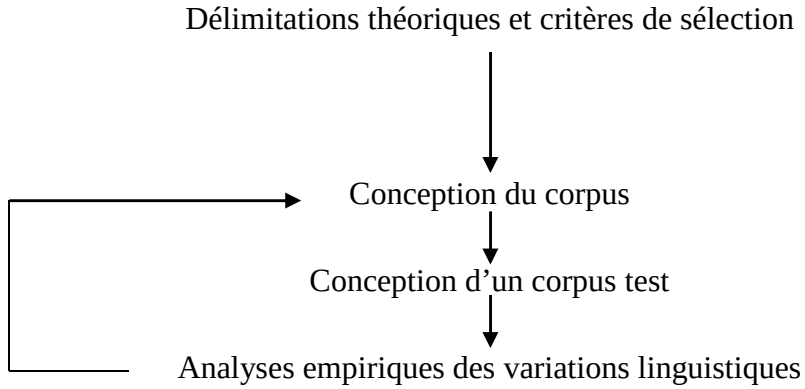
Figure 1 : Hiérarchie d'échantillonnage des corpus proposée par Biber[154]

1. Choix du médium : écrit, oral / oral / écrit oralisé
2. Format : publié / non publié
3. Cadre : institutionnel / public / privé-personnel
4. Destinataires : (a) indéfinis / numérables / individuel / monologue
   (b) présence / absence
   (c) non interactifs / peu interactifs / interactifs
   (d) connivence : discours général / spécialisé / personnel
5. Auteurs / Locuteurs : (a) variations démographiques : sexe / âge / occupation etc.
   (b) locuteurs mandatés, individuels, institutionnels etc.
6. Factualité : factuel-informationnel / semi-factuel / imaginaire
7. Objectifs : persuader, divertir, informer, instruire, expliquer, narrer, décrire, tenir des registres, révéler, exprimer des attitudes, des opinions ou des émotions etc.
8. Thématiques

C'est dans la richesse d'une telle liste que réside la difficulté de la construction d'un corpus. Bien que fournie, d'aucuns pourraient la considérer comme réduite et les possibilités de la détailler sont extrêmement variées. Nous poursuivrons pour l'instant l'examen de la méthodologie de Biber, pour revenir sur les critiques ensuite.

[154] La traduction est la nôtre.

Un corpus de petite taille est constitué, et ce « corpus pilote » est ensuite testé afin de vérifier s'il contient toutes les variations linguistiques prévues. Les résultats du test sont alors pris en compte dans l'étape d'ajout de données qui suit, puis la manœuvre est répétée de façon continue ; dans cette optique, le corpus de Biber n'est jamais terminé, mais augmente sa représentativité dans le temps. Dans sa conclusion, Biber (1993a : 256) résume le processus selon le schéma suivant :



Biber propose donc la constitution d'un corpus ne se basant pas sur la réalité du langage, mais sur des critères d'importance ne pouvant être que subjectifs. Il rejette explicitement (Biber, 1993a : 247) une représentation en strates proportionnelles en avançant le raisonnement suivant : l'échantillonnage proportionnel peut être soit proportionnel aux différents registres du langage, soit proportionnel aux réalités démographiques de la population étudiée. Puisque les linguistes n'ont pas les moyens de déterminer *a priori* les proportions des registres linguistiques au sein d'une population, la seule proportionnalité apte à être opérationnelle sera une proportionnalité démographique. Un tel corpus représentera les registres linguistiques selon leur utilisation réelle, et Biber suppose que ce type de corpus devrait contenir, selon ses estimations, 90% de langue orale, 3% de lettres et de notes et 7% de registres recouvrant les reportages de presse, d'écrits de magazines populaires, de prose académique, de fiction, de conférences, de communiqués et d'écrits non publiés[155]. Un tel corpus, quand bien même il serait constitué, ne susciterait que peu d'intérêt aux yeux de Biber :

> These kinds of generalizations, however, are typically not of interest for linguistic research. Rather, researchers require language samples that are representative in the

[155] Nous rappelons que l'article de Biber date de 1993, à une heure où le volume des données linguistiques d'Internet était moindre que de nos jours.

> sense that they include the full range of linguistic variation existing in a language. (Biber, 1993a : 247)

Que ce soit la proportionnalité relative aux registres du langage ou la proportionnalité démographique, l'échantillonnage d'auteurs ou de locuteurs selon leur représentation dans des registres ou catégories démographiques est donc rejetée par Biber. Il poursuit en arguant que la seule proportionnalité numérique d'un corpus est une représentativité quantitative qui ne reflète pas l'importance de certains registres qui n'ont pas d'importance numéraire. Selon Biber, les livres ou les journaux sont plus influents que ce que leur représentation numéraire laisse croire, et devraient de fait être surreprésentés dans un corpus par rapport à leurs proportions dans le langage, et il préconise donc une stratification non proportionnelle précédant la constitution du corpus. Nous illustrerons cette méthodologie de délimitation des strates du langage en amont par deux exemples : le Brown pour l'écrit, et le BNC pour l'oral.

Comme nous le verrons lorsque nous le présenterons en détail[156], le Brown est un corpus censé représenter l'anglais américain écrit du début des années 1960, et il contient pour ce faire 500 échantillons de 2000 mots chacun, pour un total d'un million de mots, catégorisés comme suit[157] :

[156] Cf. 2.10.1.

[157] Du manuel du Brown Corpus, disponible ici :
http://khnt.aksis.uib.no/icame/manuals/brown/

Tableau 4 : Détail des échantillons du Brown Corpus

374 échantillons de prose informative	**126 échantillons de prose de fiction**
• 44 échantillons de presse d'information • 27 de presse éditorialiste • 17 de presse culturelle • 17 d'écrits religieux • 36 de textes traitant de passe-temps et loisirs • 48 d'écrits populaires • 75 de littérature, 30 d'écrits neutres tels les rapports, catalogues ou manuels • 80 d'écrits académiques	• 29 échantillons de romans non catégorisés • 24 de romans policiers • 6 de science-fiction • 29 de romans d'aventures • 29 de romans sentimentaux • 9 d'écrits humoristiques

Or les critères ayant permis ces catégorisations qualitatives et quantitatives n'apparaissent ni dans le manuel du Brown, ni autre part ; il est simplement précisé que la procédure de sélection se déroula en deux phases : une première classification « subjective », selon les termes des auteurs, durant laquelle le nombre d'échantillons pour chaque catégorie fut décidé ; puis une seconde sélection durant laquelle les catégories furent choisies « aléatoirement »[158]. Si nous savons le contexte, nous ignorons les critères de choix et les motivations et justifications des auteurs. Biber (1993a), en citant le Brown et le LOB[159] qui fut constitué sur le même modèle, prône néanmoins leur modèle de constitution sans offrir un protocole de stratification rigoureux. Il en résulte que des corpus constitués selon cette méthodologie peuvent être soumis à la critique, du point de vue de la représentativité. Citant le Brown corpus, Váradi affirme que pour être représentatif de la population étudiée, un échantillon doit suivre le principe de la proportionnalité : les différentes catégories du corpus doivent être échantillonnées selon leur même ratio au sein de la langue en général. Váradi donne pour exemple :

> For the BROWN corpus to qualify as a representative sample of the totality of written American English for 1963 for humorous writing, it would have to be established that

[158] Ces décisions furent prises lors d'une conférence tenue à la Brown University en février 1963, et à laquelle participèrent Caroll, Francis, Gove, Kučera, O'Connor et Quirk (d'après le manuel du corpus, note 4).

[159] Cf. 2.10.1.

> humorous writings did make up 1.8 % of all written texts created within that year in the US. This single requirement serves to illustrate the enormous difficulty if not impossibility of the task. (Váradi, 2001 : 590)

Passons maintenant à notre second exemple, la partie orale du BNC. Burnard (1995 : 20-25) détaille la procédure comme suit : l'équipe du BNC sélectionna 124 personnes, de sorte qu'il y ait un nombre égal d'hommes et de femmes, un nombre égal de locuteurs dans six tranches d'âge prédéfinies et un nombre égal de locuteurs dans 4 classes sociales prédéfinies. Il fut demandé aux 124 locuteurs d'enregistrer leurs conversations privées, de manière discrète, durant une semaine, ce qui permit de rassembler des données orales d'un volume de quatre millions de mots.

Là encore, l'échantillon démographique du BNC ne peut être considéré comme représentatif au sens proportionnel du terme. La stratification reposa sur une répartition en catégories égales et non selon les réalités démographiques de la société anglaise. D'autre part, une distribution proportionnelle aurai nécessité la consultation des données démographiques de la société anglaise ; il ne s'agit certes pas d'une procédure banale, mais il nous semble que la représentation proportionnelle sur critères démographiques demeure un objectif plus accessible que la représentation proportionnelle des catégories du langage. Ceci est sans doute la raison pour laquelle l'absence de consultation des données démographiques de la société anglaise fut qualifiée de « laxisme méthodologique » par Váradi. Burnard (1995 : 20) rapporte que l'enregistrement d'un nombre égal de locuteurs au sein de chaque catégorie était bien un objectif revendiqué, en reprenant pour justification des raisons similaires à celles avancées par Biber, à savoir que la représentation proportionnelle n'est pas justifiée en raison de situations d'énonciation où un nombre restreint de locuteurs produit un nombre restreint d'énoncés mais qui seront destinés à un grand nombre de récepteurs.

2.6.1.1 Critiques de la stratification non proportionnelle

Biber (1993a) a distingué trois approches possibles en ce qui concerne la stratification du langage en vue de constitution d'un corpus : en se basant soit sur la production du langage, soit sur la réception du langage, soit sur le langage lui-même. Les deux premières approches requièrent le recours à des informations extratextuelles sur les auteurs ou les locuteurs, démarche qui – comme nous l'avons vu – fut rejetée par Biber, qui réfuta le principe même de la représentativité proportionnelle. Selon lui, cette dernière n'est qu'un indicateur numéraire

des fréquences des registres et ne pourvoit pas de représentation des registres « importants » ou « influents », tels les livres ou les journaux. Dans la sélection des textes, Biber se base donc sur un ensemble de critères principalement liés aux textes eux-mêmes, indépendamment des auteurs et récepteurs ; ceci implique, comme nous l'avons vu, un jugement que Váradi qualifie de subjectif et qu'il critique de la sorte :

> One of the fundamental aims of Corpus linguistics as I understand it is to show up language as is actually attested in real life use. However, Biber seems to argue that in designing a corpus one should apply a notion of importance that is derived from a definition of culture. For lack of any means of operationalizing this criterion of relative importance in culture, this throws the door wide open to subjective judgment in the compilation of the body of data that is expected to provide solid empirical evidence for language use. (Váradi, 2001 : 591-592)

Leech (2006 : 7) évoque certains corpus construits sur ce type de critères évaluatifs qui, selon l'auteur, considéraient les quotidiens de presse d'information plus importants, ou plus influents que la presse à sensation, qu'un roman ayant obtenu un prix littéraire était supérieur à un roman populaire et que les locuteurs des classes économico-sociales les plus élevées étaient les plus dignes d'être représentés dans les corpus. Il va de soi que ces démarches que l'auteur qualifie d'élitistes sont sans fondement scientifique dans le cadre d'une analyse linguistique. Cependant Leech ne réfute pas pour autant la représentativité proportionnelle, et son article nous intéresse pour ce qu'il propose comme moyen d'y accéder autre qu'une liste hiérarchique préétablie similaire à celle proposée par Biber. En effet, Leech considère, et ce en réponse à Váradi, que l'importance n'est pas impérativement une valeur subjective, et qu'elle peut être calculée. Pour ce faire, Leech (2006 : 6) propose la notion du « Atomic Communicative Event » (ACE). Selon lui, la représentativité doit se baser sur le nombre de locuteurs/auteurs, ainsi que sur le nombre de récepteurs. Tout binôme encodage/décodage consiste en un ACE : un message d'une personne à une seule autre constituera en un seul ACE, et une émission radio écoutée par X personnes constituera X ACEs, bien que l'énoncé reste unique. Leech propose donc de se baser sur le nombre de ACEs pour justifier de la représentativité des textes dans un corpus :

> Thus a radio programme that is listened to by a million people should be given a much greater chance of being included in a representative corpus than a conversation between two people, with only one listener at any one time. (Leech, 2006 : 6)

Leech a bien conscience de la difficulté extrême d'un calcul des ACEs exact, mais sa proposition n'est pas à être concrétisée : Leech ne la proposait qu'en réponse à la critique de Váradi qui affirmait que le critère d'importance ne pouvait être que subjectif. Toutefois, même sur le plan théorique, nous pensons que le calcul des ACEs ne pourrait aboutir à des résultats concrets ; quand Leech oppose plusieurs centaines de milliers de ACEs d'une émission radio au ACE unique d'une conversation privée, il nous semble que cette mise en balance est erronée : à l'instant T où une émission radio est diffusée, il y a X ACEs, X était un nombre allant de plusieurs dizaines de milliers à plusieurs millions de récepteurs. La théorie de Leech est de mettre en balance ces X ACEs par rapport à l'ACE unique d'une conversation privée se déroulant au même instant T. Or à l'instant T, il ne se déroule pas une mais Y conversations privées, et Y est un nombre que l'on peut supposer très grand. Il ne s'agit certes pas du même message, mais tous les ACEs de cette catégorie, aussi variables soient-ils, appartiennent à la catégorie des conversations privées et partagent donc les spécificités de cette dernière. Nous estimons donc que s'il devait y avoir un calcul rigoureux de l'importance des énoncés se basant sur les ACEs, celui-ci devrait prendre en compte l'ensemble des ACEs d'une catégorie dans le calcul et non pas un cas isolé de chacune des catégories comme le propose Leech.

Ceci n'est pas une simple divergence dans la méthodologie, car s'il semble possible de calculer, ou d'estimer les ACEs d'une émission radiophonique, le calcul de l'ensemble des ACEs d'une langue, toutes catégories confondues, à un instant donné T ou au cours d'une période délimitée, ne relève plus de la difficulté mais de l'impossibilité au vu des moyens technologiques actuels. Concernant ce dernier point et toujours sur le plan théorique, la proposition de Leech semble prendre le problème à l'envers : en effet, le jour où il sera possible de calculer l'ensemble des ACEs d'une langue, il nous semble que ce calcul devra se faire sur base d'un corpus représentatif, or c'est justement en vue de constituer ce corpus représentatif que Leech propose le calcul des ACEs. À l'heure actuelle, Váradi est donc sans doute dans le vrai dans son affirmation qu'une pré-catégorisation en vue de constituer un corpus représentatif est erronée, car cette pré-catégorisation ne remplira pas les objectifs initiaux de la représentativité :

> Surely, it is simply not feasible to put a figure on the amount of text within the various genres in the totality of texts produced by a speech community. Yet, this is what the statistical concept of a representative sample calls for. (Váradi, 2001 : 590)

Dans son article, Váradi (2001 :592) visait essentiellement à « mettre l'accent sur les incertitudes, les inconsistances et les raccourcis méthodologiques au sein de la linguistique de corpus » et en appelle à davantage de rigueur ; la méthodologie alternative est à rechercher du côté de Sinclair, comme nous le allons le voir.

2.6.1.2 Équilibre d'un corpus

Avant de poursuivre, il nous apparaît nécessaire de dire quelques mots d'une notion étroitement liée à celle de la représentativité dans la littérature : la notion de corpus équilibré. Généralement, un corpus est dit équilibré quand la taille de ses sous-catégories (genres, registres etc.) est proportionnelle à leurs fréquences d'occurrence au sein du langage général. En d'autres termes et selon Leech (2006 : 137), l'équilibre d'un corpus est synonyme de la proportionnalité que nous avons discutée ci-dessus. Nous rappelons que Biber rejeta cette proportionnalité pour prôner l'équilibre entre les catégories elles-mêmes. Nous abondons dans le sens de Leech qui considère un corpus équilibré au sens de Biber – soit un corpus dont les différentes sous-catégories seraient identiques en volume – un objectif déviant de celui de la représentativité :

> Perhaps Biber's method is just another way of achieving balance. It will mean that language varieties are to be represented in the corpus in proportion to their heterogeneity, rather than in proportion to their prevalence of use in the whole textual universe. Arguably, this is not representativeness, but another corpus desideratum: heterogeneity. (Leech, 2006 : 139)

L'équilibre d'un corpus peut donc être synonyme soit de proportionnalité, soit de corpus dont les strates sont équilibrées selon la vision de ses compilateurs. Il convient donc de convenir des réalités suivantes : dans le premier cas d'un corpus proportionnel, l'équilibre d'un corpus est un objectif inatteignable en raison de l'impossibilité actuelle à constituer un corpus dont les structures refléteraient effectivement les réalités du langage ; dans le second cas d'un corpus dont l'équilibre serait concrétisé par une stratification en catégories prédéfinies, l'équilibre du corpus est alors tout autant exposé aux mêmes critiques que nous pouvons légitimement émettre envers la stratification choisie par les compilateurs.

2.6.2 Monitor corpus

La seconde grande approche de la représentativité des corpus, par rapport à la stratification en amont, est principalement représentée par John Sinclair. Sinclair propose l'échantillonnage en

tant que solution à la problématique de la représentation du langage. En ceci, sa démarche ne diffère pas de celle employée dans la représentativité proportionnelle puisque cette dernière repose sur l'échantillonnage également. Néanmoins c'est dans leurs conceptions de l'échantillonnage que les deux méthodologies divergent, comme nous allons le voir.

Comme nous l'avons vu ci-dessus, la démarche de l'échantillonnage en vue d'une représentativité proportionnelle se base sur la délimitation de genres *a priori*, démarche à laquelle s'oppose Sinclair. Nous avons déjà discuté de la vision des corpus de Sinclair et de son appartenance à l'école de Firth, pour laquelle « le sens des textes » est l'objectif primordial de la linguistique de corpus (Teubert, 2001). Or, l'approche de Sinclair est effectivement une approche probabiliste se basant sur le sens des textes, que Léon décrit en ces termes :

> L'approche probabiliste du sens, qu'il [Sinclair] partage avec Halliday, le conduit à considérer qu'en établissant des patterns de collocations à partir de grands corpus de textes, on peut établir le sens d'une expression non de façon absolue mais plutôt comme une tendance probable. Ceci aura des conséquences sur la constitution d'un corpus, toujours augmentable et jamais fini. C'est pourquoi, dès les années 1960, Sinclair est opposé à la méthode d'échantillonnage et aux genres *a priori* ; d'ailleurs, dès le rapport OSTI, il entrevoit la possibilité d'établir une typologie des textes à partir de traits linguistiques sur des données textuelles de grande taille, au lieu de travailler à partir des genres. (Léon, 2008 : 19)

En d'autres termes, la finalité des corpus serait de dresser une liste du ou des sens d'un item lexical (Teubert, 2010 : 7) à partir des collocations de cet item. Et ce sont les impératifs de cette dernière opération qui impliquent la notion d'un corpus « toujours augmentable et jamais fini » que Sinclair nomme « monitor corpus », au sein duquel la probabilité de cerner le sens d'un item lexical est proportionnelle au volume des données. Les monitor corpus, ou corpus de référence dont nous allons détailler la constitution sont définis par Sinclair de la sorte :

> A general reference corpus is not a collection of material from different specialist areas – technical, dialectal, juvenile, etc. It is a collection of material which is broadly homogeneous, but which is gathered from a variety of sources so that the individuality of a source is obscured, unless the researcher isolates a particular text. (1991 : 17)

Ce type d'approche est résumé par Habert (2000) ainsi :

> La conviction sous-jacente est que l'élargissement mécanique des données mémorisables (les centaines de millions de mots actuelles deviendront à terme des

> milliards) produit inévitablement un échantillon de plus en plus représentatif de la langue traitée. Si l'on n'arrive pas à cerner précisément les caractéristiques de l'ensemble des productions langagières, il ne reste qu'à englober le maximum d'énoncés possibles. À terme, la nécessité de choisir finirait par s'estomper.

Comme le note Léon (2008 : 19), ces choix théoriques et méthodologiques exigent des moyens de stockage et de traitement informatiques encore indisponibles dans les années 1960. Sinclair mettra en pratique ses théories dans les années 1980 avec le projet COBUILD qui reposa sur la constitution d'un corpus de grande taille de textes authentiques et intégraux, conformément à sa conception probabiliste des corpus. C'est donc ici que Sinclair s'écarte de la méthodologie se basant sur une stratification en amont de catégories égales selon le modèle du Brown Corpus:

> It should, however, be realized that this feature is just a remnant of the early restraints on corpus building and it confers no benefit on the corpus. The use of samples of a constant size gains only a spurious air of scientific method. (Sinclair, 1996)

La méthodologie d'échantillonnage de Sinclair est donc autre, et détaillée dans par exemple « Developing Linguistic Corpora : a Guide to Good Practice. Corpus and Text – Basic Principles » (Sinclair, 2004). Elle se résume à un procédé devant prendre en considération les trois points suivants :

1. l'orientation des textes ;
2. les critères selon lesquels les échantillons seront choisis ;
3. la taille et la nature des échantillons.

En premier lieu, l'orientation des textes est le choix du type de textes à inclure dans le corpus. En guise d'exemple, Sinclair (2004) cite le Brown Corpus en tant que « corpus à visée normative » ; cet objectif de recherche de la norme ou de standardisation du langage entraîne selon Sinclair une désélection de la plupart des variétés du langage[160]. Sinclair argue que les premiers corpus, mais également la plupart des corpus récents, sont construits sur le même modèle :

> Most of the large reference corpora of more recent times adopt a similar policy; they are all constructed so that the different components are like facets of a central, unified whole. Such corpora avoid extremes of variation as far as possible. (Sinclair, 2004)

[160] En ce qui concerne le Brown, ses concepteurs ne retinrent que des textes publiés.

Il est nécessaire ici d'expliciter la terminologie de Sinclair, en ce qui concerne le terme « component » (composant). Sinclair propose les catégorisations suivantes : un corpus peut être divisé en sous-corpus (subcorpora). Les corpus et leurs sous-corpus sont hétérogènes et les sous-corpus possèdent toutes les propriétés du corpus auxquels ils sont rattachés. Les corpus et les sous-corpus sont constitués de composants. Les composants sont des ensembles de textes ou d'énoncés sélectionnés selon des critères qui font d'un composant un ensemble de données qui est au moins homogène du point de vue du critère de sélection du composant. Alors que les corpus et les sous-corpus sont hétérogènes, à l'image du langage, les composants tendent à illustrer de manière homogène un type de langage. En tant qu'exemple, voici les sous-corpus du corpus BoE[161] :

Ex 1

Sous-corpus	Nombre de mots en millions
Australian news	5.3
UK ephemera	3.1
UK magazines	4.9
UK spoken	9.3
US ephemera	1.2
BBC World Service	2.6
National Public Radio	3.1
UK books	5.3
US books	5.6
Times newspaper	5.7
Today newspaper	5.2

Ces sous-corpus se divisent ensuite en composants selon des critères de sélection (deuxième point de la méthodologie), dont Clear (1992) distingue deux types en jeu dans le choix des textes d'un corpus. Le premier regroupe des considérations essentiellement linguistiques et représente « les critères internes » : la catégorisation d'un texte sur bases de critères syntaxiques et lexicaux sera ainsi une catégorisation basée sur des critères internes. Le second type de critères, « les critères externes », concernent les informations métalinguistiques du texte, comme l'âge ou le sexe de l'auteur ou du locuteur. Sinclair considère que la

[161] Du site du corpus BoE : http://www.titania.bham.ac.uk/docs/svenguide.html#The%20Corpus%20Access%20Tool:%20a%20brief%20introduction

catégorisation d'un texte selon des critères externes délimitera les critères internes du texte, et non l'inverse. C'est donc uniquement sur bases de critères externes que Sinclair préconise l'échantillonnage, sur les recommendations de Clear :

> A corpus selected entirely on internal criteria would yield no information about the relation between language and its context of situation. A corpus selected entirely on external criteria would be liable to miss significant variation among texts. (Clear, 1992 : 29)

Sinclair donne comme exemple un ensemble de critères pouvant être pris en compte dans la sélection des textes[162] :

1) le mode du texte : écrit, oral, électronique.
2) le type du texte : par exemple dans le mode écrit, s'il s'agit de livres, de journaux ou de lettres ;
3) le domaine du texte : académique ou populaire ;
4) le langage employé dans les textes ;
5) le lieu des textes : par exemple l'anglais du Royaume-Uni ou l'anglais d'Australie ;
6) la date des textes.

Outre les exemples ci-dessus, Sinclair préconise l'emploi d'un nombre restreint de critères, clairement définis et établis de manière à parachever ou parfaire la représentativité du corpus. Que l'on suive la méthodologie des méga-corpus de Sinclair ou non, nous sommes en accord avec lui lorsqu'il affirme que la représentativité dépend uniquement du choix des critères externes lors de la constitution :

> For a corpus to be trusted, the structural criteria must be chosen with care, because the concerns of balance and representativeness depend on these choices. (Sinclair, 2004)

La constitution d'un corpus selon des choix métalinguistiques implique la présence de données métalinguistiques au sein ou en annexe du corpus. L'ensemble de ces données métalinguistiques constitue la documentation du corpus[163].

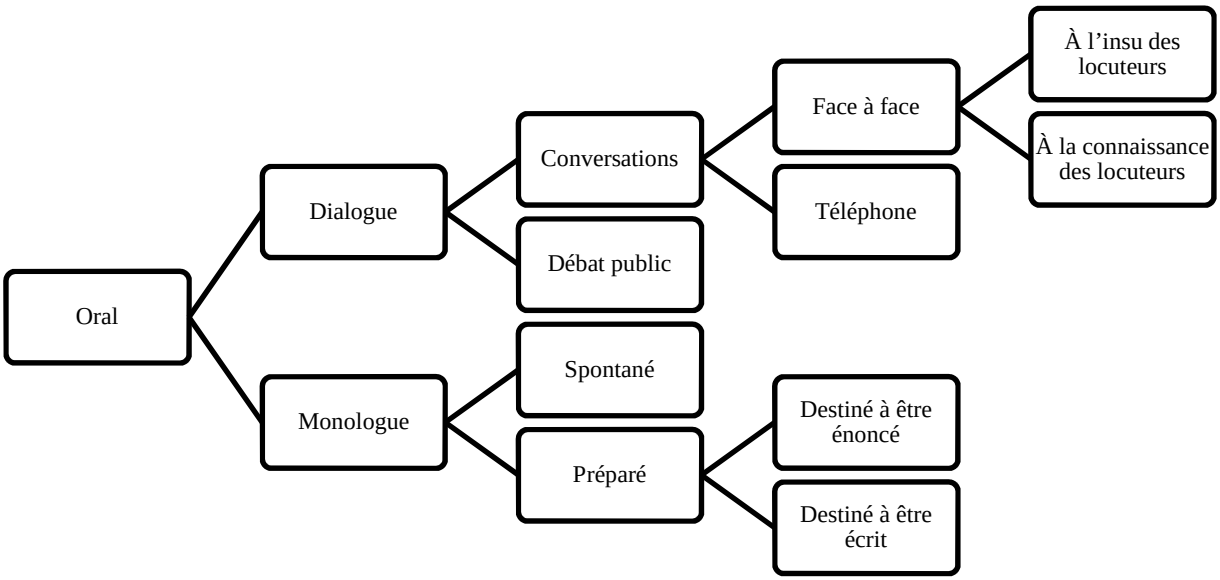
[162] Du site « Guides to good practice » (AHDS Literature, Languages and Linguistics) : http://ota.ahds.ac.uk/documents/creating/dlc/chapter1.htm#section3
Les choix terminologiques et les exemples sont ceux de Sinclair, la traduction est la nôtre.

[163] Cf. 2.8.

Nous avons vu que les corpus et les sous-corpus sont divisés en composants selon des critères externes. Nous ajoutons que Sinclair utilise les terminologies de « texte » pour un document isolé et de « cellule » pour les ensembles de données hiérarchiquement inférieurs aux composants. La troisième et dernière question qui se pose est l'échantillonnage, soit en d'autres termes, le nombre de composants, le nombre de cellules au sein d'un corpus, ainsi que le volume des données au sein de chaque cellule.

La méthodologie proposée par Sinclair est une classification par binômes. Nous prendrons l'exemple de la partie orale du Lund Corpus. La classification binaire implique la division du corpus oral en deux catégories : dialogue et monologue. Chaque catégorie sera ensuite divisée en deux parties comme présenté dans le schéma ci-dessous[164] :

Figure 2 : Stratification de la partie orale du London-Lund Corpus



Sinclair impose ensuite une condition : non pas des catégories strictement égales à l'image des catégories du Brown Corpus, mais des cellules qui posséderaient un nombre minimum de

[164] Le London-Lund Corpus of Spoken English est un corpus regroupant les corpus de deux projets : le corpus du SEU et le corpus du Survey of Spoken English, projet entamé à la Lund University en 1975 en tant que projet jumelé au SEU. Le schéma que nous présentons provient du manuel du London-Lund Corpus, disponible ici : http://khnt.hit.uib.no/icame/manuals/londlund/index.htm :

données[165]. À chaque niveau supplémentaire, le nombre de mots du corpus se voit théoriquement au minimum doubler. C'est pourquoi Sinclair préconise l'emploi d'un nombre restreint de critères pour des raisons pratiques.

En ce qui concerne le nombre minimum de mots au sein de chacune des cellules, Sinclair dit que c'est une décision qui dépend principalement du type de recherche que l'on mènera sur le corpus, mais stipule que le volume des données se doit d'être « substantiel » afin que le chercheur puisse en tirer des conclusions scientifiques viables. Ainsi Sinclair donne, à titre d'exemple, le nombre minimum d'un million de mots par cellule. C'est ici que l'on comprend les raisons de sa préconisation d'un nombre restreint de critères : dans l'exemple du London Lund Corpus ci-dessus, le nombre d'un million de mots par cellule implique un corpus oral de 16 millions de mots. Le rajout d'un seul niveau de critères aux cellules inférieures ferait de ce corpus un corpus de 32 millions de mots au minimum. Nous rappelons que la partie orale du BNC ne contient « que » 10 millions de mots, alors que le plus grand corpus oral français (corpus DELIC) ne dépasse pas 1 400 000 mots, et qu'il n'est d'ailleurs pas mis à disposition.

La vision de Sinclair est donc la constitution de corpus toujours augmentables et jamais finis au sein desquels, selon sa vision probabiliste des corpus, la représentativité serait proportionnelle au volume des données. La fiabilité, l'usabilité et donc la pertinence des méga-corpus ne fait néanmoins pas unanimité parmi les théoriciens de la linguistique de corpus, et les raisons des réticences à la constitution de méga-corpus ne sont pas uniquement d'ordre pratique[166], comme nous le détaillerons ci-dessous.

[165] La raison principale pour laquelle Sinclair rejette la création de cellules strictement égales est son refus de tronquer les textes ou les enregistrements oraux, d'où des échantillons pouvant être extrêmement variables en termes de volume des données (Sinclair, 2004) : « Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size. »

[166] Váradi (2001, n. 1) rapporte une anecdote dont il n'y a pas de trace écrite : il y eut un débat à l'Université d'Oxford entre les plus célèbres représentants des deux courants majeurs dont nous avons parlé : Quirk (SEU Project) et Leech défendaient la faveur des corpus ouverts et jamais finis. Toujours selon Váradi, la tradition veut que l'équipe de Sinclair ait été celle position des corpus fermés et équilibrés sur le modèle du Brown, tandis que Sinclair et Meijs argumentaient en qui présenta les arguments les plus convaincants.

Maintenant que nous avons présenté les différentes théories de la représentation horizontale, nous allons discuter des problématiques liées à la représentation verticale des données, soit à la question de la taille des corpus.

2.6.3 Taille des corpus

La question de la taille des corpus devrait être formulée plus précisément par la taille de chaque échantillon au sein du corpus, puisque la question du nombre d'échantillons et de leurs tailles respectives concerne la représentation horizontale. Toutefois, nous discuterons également de la taille absolue des corpus pour deux raisons :

1) tous les théoriciens de la linguistique de corpus n'ont pas différencié la représentation horizontale de la représentation verticale, et certains traitent le corpus en tant qu'entité ;
2) l'existence de corpus dit spécialisés, tels les corpus de personnes âgées, corpus d'acquisition, corpus historiques ou des corpus d'apprenants à l'image de celui que nous avons constitué dans le cadre de cette étude.

Nous formulerons la problématique de la taille des corpus de la sorte : quel doit être le nombre d'occurrences d'un terme au sein d'un corpus, afin que les conclusions tirées à partir de ce nombre d'occurrences soient généralisables ? Soit, en termes statistiques, comment déterminer la taille de l'échantillon (le corpus) afin que celui-ci soit conforme aux conditions de validité d'un échantillon statistique, et que la marge d'erreur inhérente au processus d'inférence statistique soit minimalisée ?

Avant d'apporter divers éléments de réponse, il est nécessaire de préciser que la question ne peut être résolue uniquement sur des critères statistiques. Deux points sont à prendre en compte, en amont et indépendamment des principes linguistiques et mathématiques :

1) les limitations d'ordre matériel ;
2) les visées linguistiques du corpus.

Par exemple, les premiers corpus numérisés tel le Brown Corpus et le LOB durent faire face à certaines difficultés logistiques induites par la numérisation de données sur papier. Ces difficultés furent surmontées avec l'utilisation de scanners à reconnaissance optique des

caractères[167], mais le problème persiste en ce qui concerne la collecte des données orales, la reconnaissance vocale n'étant pas aussi performante que la reconnaissance optique[168]. Les corpus oraux impliquent donc toujours l'enregistrement de données et la transcription de celles-ci, deux étapes lourdes en temps et en ressources humaines en raison du travail de terrain nécessaire à la première, et de la transcription manuelle des données[169]. Meyer (2002 : 32) rapporte ainsi que le Santa Barbara Corpus of Spoken American English nécessita six heures de travail pour la transcription et l'annotation d'une minute de parole, et que « ces faits logistiques expliquent pourquoi 90% du BNC est écrit et que seulement 10% est de la langue parlée ». Ces difficultés matérielles impliquent des considérations triviales quant à la taille désirée du corpus :

> To determine how long a corpus should be, it is first of all important to compare the resources that will be available to create it (e.g. funding, research assistants, computing facilities) with the amount of time it will take to collect texts for inclusion, computerize them, annotate them, and tag and parse them. (Meyer, 2002 : 32)

Ce n'est qu'après avoir pris en considération ces contraintes qu'une estimation scientifique de la taille idoine d'un corpus pourra être effectuée, en prenant en compte également le second point que nous avons évoqué, les objectifs d'exploitation du corpus. Par exemple, un corpus à visées lexicographiques se devra d'être suffisamment grand pour pouvoir créer un dictionnaire, alors qu'un corpus plus modeste pourra rendre compte des variations régionales d'un pays. Comme le remarque Granger (2007 : 1), un corpus de 200.000 mots sera considéré comme grand dans le domaine de l'acquisition des langues, mais minuscule s'il s'agit d'un corpus littéraire ou le recours à des corpus de centaines de millions de mots est devenu la norme[170].

[167] Pour davantage de détails concernant les types de scanners et leur rôle dans la collecte des données écrites, voir Meyer (2002, section 3.7.).

[168] D'autre part, le problème de la numérisation des données écrites concerne en grande partie uniquement les textes antérieurs à l'ère informatique et principalement le Web ; la plupart des textes actuels étant disponibles sous forme numérique, voire directement crées sous cette forme.

[169] Manuelle ou semi-automatique dans certains cas, le problème sera abordé en détail en 2.7.2.

[170] Si cela est devenu possible, c'est évidemment grâce aux progrès de l'informatique. Avant l'ère des corpus numériques, les corpus constitués étaient de taille relativement restreinte en comparaison aux corpus électroniques actuels. L'ordre de grandeur peut être beaucoup plus grand que ce que nous avons indiqué, voir à ce sujet la note n° 259.

La question des objectifs linguistiques du corpus est même primordiale pour Kennedy qui doute de la nécessité des méga-corpus – nous reviendrons sur ce point ci-dessous – et propose de privilégier la qualité des données plutôt que leur quantité :

> Rather than focusing so strongly on the quantity of data in a corpus, compilers and analysts need also to bear in mind that the quality of the data they work with is at least as important. (Kennedy, 1998 : 68)

La qualité des données signifie précisément pour Kennedy la prise en considération des objectifs linguistiques du corpus : Kennedy poursuit en mentionnant quelques exemples, selon lesquels la quantité de données nécessaire à une analyse prosodique serait de 100 000 mots, sous la condition que ces données soient du type « parole spontanée » ; ou qu'une étude sur la morphologie des formes verbales nécessiterait entre 500 000 et un millions de mots. De son côté, en 1992, Leech entrevoyait pour 2021 la possibilité d'un ensemble de corpus de mille milliards de mots[171] :

> Machine-readable text collections have grown from one million to almost a thousand million words in thirty years, so it would not be impossible to imagine a commensurate thousand-fold increase to one million million word corpora before 2021. (Leech, 1991b)

Mais que l'on ne s'y trompe pas, Leech n'effectue ces calculs que pour souligner la relative importance de la taille d'un corpus, jugée comme « critère naïf », ceci pour quatre raisons :

1) l'accumulation, ou la compilation de textes numériques n'en fait pas un corpus, car ces accumulations passent outre les critères de la représentation horizontale ;
2) l'impossible homogénéité entre représentation de la langue écrite et de la langue orale, pour des raisons liées à la relative difficulté de constitution de bases de données orales que nous avons évoquées ci-dessus ;
3) le retard institutionnel, incapable de suivre les évolutions technologiques aussi rapidement qu'elles se développent. Leech mentionne notamment le problème des droits d'exploitation ;
4) l'insuffisance technologie des outils destinés à l'analyse des corpus, tels les concordanciers.

[171] Leech n'explique pas pourquoi la courbe du volume des données pourrait être exponentielle.

Nous remarquons que les trois premiers arguments de Leech sont toujours d'actualité de nos jours, et que si le dernier peut paraître obsolète en raison des multiples outils à disposition, le problème aujourd'hui est quelque peu différent mais de même ordre : l'étal conséquent d'outils mis à disposition est en soi une avancée, mais voit ses possibilités d'exploitation limitées en raison du manque de coordination de méthodologies et de standards entre équipes.

En ce qui concerne la taille des corpus proprement dite, nous n'avons, là encore, retrouvé que deux véritables méthodologies clairement énoncées, par les mêmes auteurs qui énoncèrent les méthodologies de représentation horizontale d'un corpus : Biber et Sinclair.

Biber propose dans son étude (1993a : 248) un calcul strictement mathématique de la taille d'un corpus pour que celui-ci soit verticalement représentatif. L'équation sur laquelle Biber repose son étude permet de calculer l'erreur type (σ_m) d'un échantillon, qui est égal à l'écart type (σ) divisé par la racine carrée de la taille de l'échantillon (n), soit[172] : $\sigma_m = \sigma / n^{1/2}$

Il semble qu'une telle méthodologie ne soit pas appropriée pour une raison que Biber discute lui-même : calculer le potentiel de représentativité d'un échantillon suppose avoir à disponibilité *a priori* d'un échantillon représentatif. Or il n'existe pas, en ce qui concerne par exemple le français, un corpus oral scientifiquement reconnu comme représentatif de la langue (et encore moins en ce qui concerne l'interlangue). L'application des calculs de Biber nécessite pourtant de tels corpus :

> Present-day researchers on English language corpora are extremely fortunate in that they have corpora such as the Brown, LOB, and London-Lund corpus for pilot investigations, providing a solid empirical foundation for initial corpus design. (Biber, 1993a : 256)

Nous remarquons néanmoins les points suivants : le calcul de la taille d'un échantillon de manière aussi rigoureuse, reste tributaire des biais de représentation au sein des corpus de

[172] L'écart type indique la « variation » ou la « dispersion » des données d'un échantillon par rapport à la moyenne. Un écart type faible indiquera que les données tendent vers la moyenne, et vice versa. Par exemple, un échantillon à l'écart type faible indiquera la même fréquence d'utilisation de tel ou tel terme. L'erreur type indique la marge d'erreur que contient un échantillon restreint par rapport à l'écart type d'un échantillon représentatif. Dans le cas de deux échantillons de tailles égales pris dans un corpus, plus ces échantillons seront petits, plus leurs données seront dispersées par rapport à la moyenne du corpus, et seront donc imprécis à indiquer la fréquence adjectivale par exemple. Le calcul de cette imprécision est l'écart type.

référence cités. Cette remarque n'est pas une critique envers la structure de ces corpus, dont certains, comme le Lund, font crucialement défaut en langue française, mais force est de constater que malgré leur importance et leur utilité, ces corpus peuvent être critiqués et ne sauraient être des corpus de référence garantissant une représentation fiable à 100%. Là encore, nous ne prétendons pas la nécessité d'une représentation parfaite, mais c'est ce que laisse supposer l'emploi de calculs aussi précis que celui que nous avons présenté. Nous rappelons que la structure du Brown a été élaborée grâce à l'intuition de ses auteurs ; cette intuition a certes fait l'objet de critiques, comme celles de Váradi, mais elle n'est ni à exclure, ni à être considérée comme une référence absolue en matière de constitution de corpus. Enfin, nous n'avons pas connaissance de corpus ayant été constitués selon la méthodologie de Biber, qui remonte à plus de vingt ans.

2.6.3.1 Les méga-corpus

La nécessité de corpus de très grande taille est défendue par Sinclair, qui considère que « grand » est la valeur par défaut de la quantité de données (Sinclair, 1996), et préconise un corpus « aussi grand que possible », se basant en cela sur la loi de Zipf :

> The only guidance I would give is that a corpus should be as large as possible, and should keep on growing. This advice is based on the pattern of word occurrence in texts, first pointed out by Zipf (…). In order to study the behaviour of words in texts, we need to have available quite a large number of occurrences (…). This is why a corpus needs to contain many millions of words. (Sinclair, 1991 : 18)

Toujours en vertu de la loi de Zipf, Sinclair (2004) se justifie par le besoin d'un nombre minimal d'occurrences d'un phénomène donné pour que celui-ci puisse être étudié. Il avance que la récurrence d'un phénomène représente une fréquence au moins double, et qu'une récurrence stricte (le phénomène apparaît seulement deux fois au sein du corpus) n'est pas suffisante pour l'étude du phénomène. Sinclair poursuit en arguant qu'un chercheur doit se fixer un taux de fréquence minimal en-dessous duquel l'occurrence ne peut être un objet d'étude.

Il y a moins deux obstacles à ce type de méthodologie : en premier lieu, que le linguiste prévoie le taux de fréquence minimal d'un phénomène suppose que tout compilateur de corpus sait, dès la constitution du corpus, ce à quoi il servira, ce qui n'est pas le cas et ce qui ne devrait de toute façon pas être le cas. En second lieu, dans le cadre d'un corpus de

référence, la quantité des données nécessaire à l'obtention la double occurrence de certains phénomènes relève de l'inaccessible, *a fortiori* quand il s'agit de données orales. À titre d'exemple, Geyken (2008 : 82) rapporte les résultats d'une étude selon laquelle « il faudrait disposer d'un corpus représentant cinquante années du journal *Le Monde* si l'on voulait « voir apparaître au moins une fois tous les mots composés recensés selon les critères définitoires du lexique-grammaire ». En nous basant sur les chiffres d'une autre étude menée sur un corpus de presse compilé à partir de *Le Monde*[173], nous avons estimé qu'un tel corpus représenterait environ un milliard de mots. Sachant que si une occurrence apparaît dans un corpus de n mots, elle n'apparaîtra pas obligatoirement une seconde fois dans le même corpus doublé à 2 x n mots[174], sachant également qu'il y a sans doute des faits linguistiques encore plus rares que les mots-composés cités dans l'exemple ci-dessus, nous pouvons en conclure qu'un corpus horizontalement représentatif au sens entendu par Sinclair[175] serait un corpus d'au moins x milliards de mots.

La théorie de Sinclair est toutefois en phase avec ses applications pratiques, puisque le corpus BoE sur lequel il travailla, comporte environ 650 millions de mots. D'autre part, non seulement les projets de Sinclair ont permis la création du premier dictionnaire entièrement constitué sur corpus, mais Kennedy (1998 : 70) rapporte que l'analyse des collocations au sein de grands corpus a permis l'identification de nouvelles variantes, de nouveaux modèles de construction et voire de nouvelles caractéristiques grammaticales. Kennedy illustre son propos en citant une étude de Sinclair[176] (1989) sur la préposition *of*, dans laquelle il démontre que la description linguistique de l'une des plus fréquentes prépositions de la langue anglaise dans les grammaires antérieures ne correspond pas totalement à ce qu'il a pu constater dans les corpus.

[173] L'étude s'intitule « La famille *laïcité* dans la base LM10 (*Le Monde* 10 ans - 91-00) ». Comme indiqué dans le titre, le corpus représente dix années du journal, soit 200 millions de mots. Les chiffres que nous avons utilisés proviennent de l'adresse suivante :
http://www.tal.univ-paris3.fr/plurital/cours2-2004/Corpus_Laicite_LM10.html

[174] Il peut même se produire un paradoxe dans une comparaison inter-corpus : Geyken (2008 : 86) a vérifié qu'une occurrence apparaissait 86 fois dans un corpus d'un milliard de mots, et que la fréquence descendait à 54 dans un autre corpus deux fois plus grand.

[175] Les objectifs d'un tel corpus, nous le rappelons, seraient principalement lexicographiques via le calcul des collocations.

[176] Nous n'avons pu retrouver cette étude, nous citons ce qu'en rapporte Kennedy.

2.6.3.2 La nécessité de corpus moins grands, plus spécifiques

Qu'un corpus soit plus grand qu'un autre, ou enrichi avec le temps, peut paraître indiscutablement positif. Mais les auteurs sont nombreux à plaider pour des corpus plus spécifiques, mieux construits, plus accessibles et surtout plus adaptés aux besoins du linguiste[177]. C'est la position de Cappeau & Gadet, qui rappellent que si l'évolution de l'informatique a permis la constitution et l'exploitation de corpus de grande taille, cette évolution doit rester pour les linguistes « une condition nécessaire, mais non suffisante, pour espérer disposer d'un recueil exploitable. » Les auteurs poursuivent en mettant en garde contre les expansions injustifiées d'un corpus :

> En contrepartie, on peut craindre que le linguiste ne s'enivre d'une accumulation de données, avec l'idée implicite que plus il y en a, mieux c'est : cent mille mots, c'est forcément mieux que cinq mille, même si ces cinq mille-là devaient bouleverser la représentation d'un champ. (Cappeau & Gadet, 2007a : 101)

Il est nécessaire de préciser que ces réflexions concernent l'état actuel des choses, à savoir la relative abondance des données textuelles, mais que la difficulté à obtenir des données orales transcrites reste d'actualité. Comme exemple de ces « dérives » qui ne concernèrent que les données textuelles, nous citerons Burnard :

> Durant les années 1990, même si toutes sortes de supports textuels faisaient l'objet d'une maquette au format numérique avant leur impression sur papier, l'idée que cette forme numérique elle-même pouvait avoir une quelconque valeur n'était pas monnaie courante. En outre, la numérisation était à cette époque précédant le commerce électronique encore loin d'être une constante, tant dans sa couverture que dans les formats. La conséquence était que les chercheurs succombaient à une tendance, bien compréhensible, de sauter sur n'importe quel texte électronique disponible sans prendre plus en compte leur statut spécifique par rapport à la langue en générale. Pour prendre un exemple célèbre, une grande partie du Wall Street Journal était alors déjà disponible sous forme numérique et guettait le danger suivant : celui de s'appuyer sur cet unique type d'écrit, d'un registre bien spécifique, pour servir de base à la linguistique computationnelle naissante et en déduire ainsi des généralisations abusives à l'ensemble de la langue. (Burnard, 2007 : 21)

Pour résumer la question de la taille adéquate des corpus, nous dirons que bien qu'il faille des corpus de très grande taille pour des études lexicographiques, le calcul des collocations ou une

[177] Par exemple Williams (2005 : 14), Habert (2000), Kennedy (1998 : 68).

description exhaustive de la langue qui reposerait uniquement sur des données attestées, la qualité des résultats de dépend pas du volume des données, et pourrait même en pâtir. En effet, plusieurs études montrent que plus de 90% des phénomènes langagiers apparaissent dans des corpus restreints[178] ; d'une part le gain obtenu grâce aux méga-corpus ne compensera pas les efforts matériels et humains fournis pour la constitution de tels corpus, et d'autre part la manipulation d'un volume de données largement inaccessibles à l'esprit humain risque grandement d'être erratique. Bergounioux *et al.* notaient bien avant la démocratisation massive des outils informatiques et d'Internet, l'importance des ressources infinies d'un corpus restreint par rapport à un méga-corpus :

> Un corpus bien fait est par définition inépuisable et la modestie de l'enquêteur revient à accepter l'idée que ce qui sera lu dans cent ans sur ses documents n'est probablement pas imaginable. (Bergounioux *et al.*, 1992 : 89)

Déterminer la taille d'un corpus amènera donc le chercheur à définir ses besoins, prendre en compte les possibilités humaines et matérielles entrant en jeu ainsi que les outils d'exploitation dont il dispose. Même dans une configuration logistique idéale, la linguistique de corpus n'est pas – encore ou ne le sera-t-elle jamais, nous l'ignorons – une science exacte. À ce propos, nous conclurons par cette phrase de Kennedy (1998 : 68) : « At this stage we simply do not know how big a corpus needs to be for general or particular purposes ».

2.6.4 Conclusion et représentativité participative des corpus oraux

Aux problématiques de la représentativité, il n'y a malheureusement que de faibles éléments de réponse. Cela vient du fait que les méthodologies employées pour représenter le langage dans son ensemble ne font pas encore consensus. La linguistique de corpus semble tellement peu capable de répondre à ces questionnements que beaucoup de corpus sont encore constitués sans même que la question de la représentativité ne se pose. Váradi (2001) en tire la conclusion pessimiste selon laquelle un corpus dont la représentativité parfaite n'a pas été attestée, ne peut fournir des informations linguistiques sur le langage, mais sur ce qu'il contient uniquement. Par ailleurs, Leech (2006) montre bien qu'à l'heure actuelle, aucun corpus n'est effectivement reconnu comme représentatif du langage. D'ailleurs, Cappeau &

[178] Cf. 1.7.2.1.

Gadet (2007a: 108-109) remettent en question l'idée même d'un grand corpus qui serait représentatif de tous les aspects du langage.

Nous avons également constaté que la problématique de la représentativité diffère selon le type de corpus ; elle est d'ordre horizontal quant aux données écrites en raison de la relative facilité à regrouper des textes écrits, et la représentativité verticale concerne davantage les données orales du fait des moyens matériels et humains considérables à engager pour l'obtention de données orales retranscrites et constituées selon des protocoles scientifiques. Enfin, il a été montré que la représentativité d'un corpus ne peut être évaluée que grâce aux critères externes dont parlait Sinclair, soit grâce aux métadonnées du corpus.

En raison de ces trois points (la représentativité dépend de la documentation des corpus, un seul corpus ne saurait être représentatif du langage, le manque de données concerne les corpus oraux), il s'est développé une certaine idée que nous appelons « représentativité participative des corpus oraux » : il ne s'agit plus de prétendre représenter le langage en un seul corpus, mais le corpus isolé participerait à la représentation du langage grâce à un regroupement des corpus constitués indépendamment les uns des autres. Cela suppose l'utilisation du corpus par des chercheurs qui ne l'ont pas constitué, et bien que cette idée paraisse évidente, la plupart des corpus oraux constitués en France n'ont malheureusement servi que leurs propriétaires. Nous utilisons le terme de « propriétaire » à dessein, car la question des droits d'exploitation est un obstacle majeur à la diffusion des corpus, sans être le seul. À ce propos, nous citerons Cappeau & Gadet qui résument la situation :

> Tout le monde reconnaît que la constitution de corpus consomme un temps considérable, et tout le monde n'a pas le goût d'y consacrer son temps et ses forces. De fait, nombre de linguistes se demandent où « trouver » des corpus, comme s'il s'agissait d'une opération purement utilitaire (…). Une autre possibilité envisagée récemment serait d'œuvrer pour un regroupement des corpus d'équipes différentes. Ce rapprochement de disparates permettrait de disposer rapidement de « gros » corpus oraux. L'état des besoins, la demande de chercheurs qui souhaitent étendre leurs investigations aux corpus oraux rend cette solution attrayante. Elle pose néanmoins des questions qui ont été indiquées dans ces pages, et elle repose sur des mythes - qui arrivent trop tardivement pour être considérés comme fondateurs : l'illusion d'une transcription minimale fructueuse pour toutes les disciplines intéressées à l'oral, l'illusion que toutes les situations présentent le même intérêt pour toutes les études, bref que tout corpus serait bon pour tous et pour tout ! (Cappeau & Gadet, 2007a : 108-109)

Ainsi, si l'on accepte l'idée de l'impossibilité d'un corpus représentatif au sens absolu du terme, la possibilité d'une représentativité participative de l'oral reste limitée par les difficultés suivantes :

1) droits d'exploitation et de diffusion ;
2) la transcription minimaliste ou standardisée des données ;
3) la documentation du corpus.

En ce qui concerne le premier point, il peut s'agir de corpus dont les droits d'exploitation n'ont pas été clairement établis, ce qui rend leur diffusion difficile comme cela est le cas pour le corpus du GARS. Il se peut également que les propriétaires ne désirent pas la mise à disposition de leurs corpus. Dans ce cas également, il s'agira de « corpus fantômes » selon l'expression de Baude & Abouda (2006 : 3), qui considèrent qu'un corpus non disponible (pour des questions de droits d'exploitation ou pour d'autres) n'existe pas. Quant au second point, quel que soit la neutralité affichée du transcripteur, une transcription de données orales implique inévitablement une interprétation. Or dans un cadre de représentativité participative où des données sont par définition destinées à être en quelque sorte uniformisées, le protocole de transcription employé reste de loin le problème le plus ardu à résoudre. Enfin, un corpus représentatif est au moins un corpus correctement documenté. Une documentation correcte signifie la disponibilité des métadonnées et la possibilité de les inclure dans les requêtes lancées sur le corpus. La question des métadonnées fait partie d'un ensemble plus vaste qui constitue un critère constituant des corpus : l'annotation des corpus.

Nous discuterons plus en détail de ces trois points qui sont des critères constituant du corpus. Pour conclure sur la représentativité, nous préconisons non pas la recherche d'une représentativité absolument rigoureuse, mais une représentativité significative au sens suivant : un corpus doit être suffisamment documenté afin que sa représentativité puisse être correctement évaluée. Si, certes, certains corpus ne pourront fournir des informations généralisables à l'ensemble du langage, la documentation annexe pourra au moins permettre aux utilisateurs de savoir ce qu'elles représentent, quand bien même cette représentation serait limitée aux données du corpus.

2.7 Annotation de corpus

L'annotation d'un corpus signifie l'ajout au corpus de données qui ne sont pas explicitement présentes lors de la compilation des données. Nous adhérons à la qualification qu'emploie Leech, qui qualifie l'annotation de « valeur ajoutée ». En ceci, l'annotation en tant qu'ajout d'information ne pourra se faire qu'en interprétant les données brutes et l'annotation comporte et la valeur interprétative lui est inhérente[179]. McEnery *et al.* (2011 : 31) précisent que l'annotation en termes d'ajout de données ne signifie pas que des données nouvelles ont été ajoutées ; elles ont été formalisées, représentées graphiquement ou constituent des interprétations. Les données brutes d'un corpus (textes bruts ou enregistrements) peuvent donc être enrichies d'informations aussi variées que les différentes branches de la linguistique : annotation morphosyntaxique (art of speech annotation), qui consiste à étiqueter les mots d'un corpus ; lemmatisation ; analyse syntaxique ; annotation évaluative ; documentation des données ; annotation sémantique ; annotation coréférentielle ; annotation stylistique ; annotation des phénomènes phonétiques, phonologiques et prosodiques. Par ailleurs, les corpus oraux actuels tendent beaucoup de nos jours à un type d'annotation que nous qualifions de paralinguistique, qui concerne l'annotation de la gestuelle des locuteurs, des expressions faciales et des postures du corps, donnant lieu aux corpus multimodaux.

La première question qui se pose est de savoir l'intérêt du processus d'annotation. Sur un corpus non annoté, les seules requêtes possibles seront celles de type orthographique. En revanche, un corpus annoté permettra des requêtes selon l'annotation effectuée. L'annotation augmente le potentiel d'exploitabilité du corpus en conséquence, Delais-Roussarie dit à ce propos :

> Tout d'abord, l'annotation des corpus est intéressante dans la mesure où elle permet de proposer des corpus étiquetés réutilisables et sur lesquels il est possible de faire des requêtes plus fines qu'en recherchant des simples chaînes de caractères avec un concordancier. Ensuite, il est important de comprendre que *ces étiquetages ne visent pas à proposer des analyses qui supplantent le linguiste*, mais seulement à lui faciliter

[179] Il est ici nécessaire de différencier entre « l'annotation » qui est par définition interprétative, et du « balisage » : le balisage des données orales consiste à l'ajout de balises qui nous donnent des informations qui concernent par exemple la prise de parole (qui parle ?), le temps de parole (combien de temps ?) ou autre type d'informations non interprétatives.

> le travail, notamment s'il veut extraire des énoncés pour travailler sur des constructions ou des points particuliers. (Delais-Roussarie, 2008 : 160-161)

Cette section commencera donc par une description du processus d'annotation et une présentation des différentes méthodologies d'annotation. Nous détaillerons ensuite les annotations que nous jugeons les plus pertinentes quant à notre propre corpus. Nous aurons auparavant discuté d'une annotation que nous n'avons pas citée plus haut : la transcription. Nous ne l'avons pas citée car le processus de transcription n'est pas considéré comme un type d'annotation par l'ensemble des théoriciens ; cette problématique et les autres annotations seront discutées ci-dessous.

2.7.1 Le processus d'annotation

Les conditions de l'annotation : En raison de la valeur interprétative de l'annotation et si le corpus est destiné à la diffusion, il est nécessaire que l'annotation soit différenciée des données brutes. La valeur interprétative ne pouvant être évitée, il s'agit de permettre aux chercheurs de la situer afin qu'elle n'altère pas les données elles-mêmes. Pour ce faire, Leech (2004) préconise les directives suivantes :

1) le processus d'annotation se doit d'être réversible ; à partir du texte annoté, un chercheur tiers devra pouvoir récupérer le texte vierge où la transcription brute ;
2) il doit être possible d'extraire l'annotation afin de pouvoir la sauvegarder de manière indépendante ;
3) l'annotation ne doit en aucun cas aboutir à une perte d'une partie des données brutes ;
4) le schéma d'annotation (voir ci-dessous) doit être disponible aux utilisateurs du corpus ;
5) il doit être indiqué le type d'annotation effectué, le nombre et l'identité des annotateurs, ainsi que les outils d'annotation utilisés ;
6) l'annotation devra rester ce qu'elle est : un outil supplémentaire ayant un caractère interprétatif ;
7) les schémas d'annotation doivent reposer sur les théories les plus neutres et les plus consensuelles possibles (voir ci-dessous) ;
8) aucun schéma d'annotation ne devra être présenté en tant que standard d'annotation.

Le schéma d'annotation : Un schéma d'annotation est l'ensemble des protocoles utilisés pour l'annotation du corpus. Il se doit de contenir les bases de données utilisées, la typologie des phénomènes annotés ainsi que toute information concernant l'analyse linguistique des

phénomènes annotés. Le schéma d'annotation inclut également le schéma des codes utilisés pour l'annotation en elle-même. Des exemples de ces schémas seront donnés dans les parties qui détailleront certains types d'annotation.

Considérons ici la préconisation essentielle de Leech : les schémas d'annotation doivent reposer sur des théories linguistiques neutres. Le schéma d'annotation suppose en effet une typologie, soit un système de classification des phénomènes annotés. En raison du fait que la linguistique est une science où les consensus sont rares et où les terminologies et les définitions diffèrent énormément, Leech suggère que les théories linguistiques sur lesquelles reposent les schémas d'annotation se doivent d'être des théories faisant plus ou moins consensus, tout en gardant à l'esprit le degré de subjectivité qu'implique un tel conseil :

> However, looking at linguistics more carefully, we can usually observe a certain consensus: examining a text, people can more or less agree which words are nouns, verbs, and so on, although they may disagree on less clear cases. If this is reasonable, then an annotation scheme can be based on a 'consensual' set of categories on which people tend to agree. (Leech, 2004)

Nous pouvons donc en conclure que l'élaboration d'un schéma d'annotation est une démarche qu'il serait préférable d'effectuer en équipe, l'annotation étant un acte linguistique interprétatif.

Enfin, un schéma d'annotation reste et doit rester une proposition d'analyse pour deux raisons :

1) les possibilités d'annotation sont multiples : un seul énoncé peut supporter plusieurs types d'annotation à la fois (annotation morphosyntaxique, évaluative et gestuelle par exemple, chacune d'entre elles possédant son propre schéma), et chacune d'entre elles sera différente selon le schéma d'annotation choisi ou constitué ;
2) les chercheurs n'ont pas à disposition une langue universelle d'annotation. Les schémas d'annotation proposés n'ont pas pour objet la création d'un standard.

Les méthodologies d'annotation : Pour que les informations ajoutées lors du processus d'annotation aient une valeur réelle et fassent des données annotées des corpus annotés plus intéressants à exploiter, l'outil informatique a compensé le coût du processus d'annotations en moyens humains. Outre ce que préconise Leech ci-dessus, l'annotation se doit d'être de

qualité, autrement dit constante, cohérente, reproductible selon le schéma d'annotation et donc documentée. Il est possible d'annoter un corpus de différentes manières :

1) manuellement ;
2) automatiquement ;
3) semi-automatiquement ;
4) en ayant recours au « crowdsourcing ».

Une annotation entièrement manuelle requiert des moyens humains considérables ainsi que les dangers des incohérences et des inévitables erreurs humaines. De par sa valeur interprétative, une annotation manuelle reste aléatoire en raison des variations de jugement des annotateurs. Le gain de temps et la cohérence des machines font donc de l'outil informatique un outil indispensable[180]. D'autre part, bien que la machine soit cohérente dans tous les cas[181], le taux d'erreur de 0% n'a jamais été atteint, quel que soit le type d'annotation. Les chercheurs tendent à recourir au compromis de l'annotation semi-automatique qui peut alors se manifester de trois manières :

1) L'annotation manuelle précède l'annotation automatique. Nous avons pu constater que ce type de méthodologie concerne des phénomènes très peu cernables par la machine, comme par exemple l'annotation de la référence et des chaînes de référence. Dans le cas d'une étude sur la référence qui a été menée dans le groupe de travail « coréférence » du laboratoire LaTTICe[182], Landragin (2011 : 8) explique en effet que « l'annotation des expressions référentielles, l'annotation des éléments coréférentiels et la construction de chaînes de coréférence ne peuvent pas être automatisées, ou alors au prix de beaucoup d'erreurs ». Pour cette raison, la méthodologie employée par les auteurs de l'étude fut la suivante : en premier lieu, la préparation du corpus, le repérage des expressions référentielles, le repérage des éléments coréférentiels, la création d'une chaîne de coréférence et la création des relations d'appartenance entre référents se faisaient manuellement. En second lieu, l'annotation des propriétés de la chaîne de coréférence, des propriétés des expressions référentielles et des

[180] Nous citerons toutefois le projet PFC dont les annotations sont entièrement manuelles.

[181] Quand un logiciel annote tel ou tel phénomène de manière erronée, la systématicité de l'outil informatique permet de corriger la lacune en une fois. Ceci n'est pas le cas en ce qui concerne l'erreur humaine qui ne survient pas systématiquement à chaque configuration semblable.

[182] Laboratoire mixte CNRS / ENS / Université Paris 3, UMR 8094.

éléments coréférentiels ainsi que l'annotation de syntagmes supplémentaires pouvaient ensuite être conduits de manière automatique[183].

2) une première annotation automatique est lancée, un chercheur ou un groupe de chercheurs interviennent *a posteriori* pour vérifier et corriger le taux d'erreur de la machine. Cette méthodologie de la plupart des corpus annotés semi-automatiquement ;

3) dans le cas de très grands volumes de données, il se peut qu'il n'y ait pas de vérification humaine systématique. Cependant Véronis (2000, n. 5) donne l'exemple de la Bank of English dont plusieurs centaines de millions de mots ont été étiquetées morphosyntaxiquement sans vérification avec un taux d'erreur avoisinant les 5% ; nous reprenons les termes de Véronis qui affirme « que la phase manuelle existe bien même dans ce cas : c'est l'utilisateur qui l'applique à chaque requête ». La part d'automatisation n'est donc jamais totale, ni jamais totalement absente. Elle peut varier et Véronis (2000 : 3) la voit comme « un continuum de possibilités entre l'annotation purement manuelle et l'automatisation complète ».

Considérons la quatrième méthode d'annotation, le « crowdsourcing »[184], qui nous semble prometteuse. Le crowdsourcing, ou « externalisation ouverte »[185] en français, est une technique qui consiste à faire effectuer une tâche par un nombre relativement important de personnes (le terme « crowd » signifie « foule »). Ces personnes peuvent effectuer la tâche en parallèle du projet initial ou en faire partie intégrante. Elles peuvent être rémunérées, mais dans la plupart des cas, il s'agit de professionnels ayant recours au savoir-faire conjugué d'un grand nombre de personnes, ou au « talent latent de la foule », selon les termes de Howe (2006 : 2). L'idée principale de la technique trouve également nombre d'autres formulations, comme « intelligence collaborative » ou « collaboration de masse ».

Le crowdsourcing est directement lié aux évolutions des nouvelles technologies de l'information et de la communication qui ont permis d'accélérer et d'outiller la logistique du crowdsourcing, principalement via Internet. Il ne concerne évidemment pas exclusivement les

[183] La méthodologie ainsi que les outils utilisés sont détaillés dans l'article cité.

[184] Le terme est un néologisme proposé par Howe (2006), la technique en elle-même remonte au début des années 1990.

[185] La traduction est celle proposée par Lebraty (2007).

corpus ou même la linguistique, le recours au crowdsourcing pouvant avoir lieu quelle que soit la science. Pourtant l'un des premiers exemples de crowdsourcing fut la compilation d'un corpus, bien avant toute technologie informatique ; il s'agit de l'édition du Oxford English Dictionary. En 1864, Frederick Furnivall, l'un des éditeurs de l'OED, fonde « The Early English Text Society » dont l'objectif était la publication de vieux manuscrits. Le recrutement de 800 volontaires pour lire et relever des citations de ces manuscrits aboutit à deux tonnes de papiers de citations, dont la plupart n'étaient pas exploitables en raison du manque de formation et de l'inconsistance des volontaires. Quand James Murray devient éditeur de l'OED en 1879, il fouilla et réorganisa la collection de Furnivall pour constater que la plupart des citations concernaient les termes rares et délaissaient l'usage commun. C'est alors qu'il lança un appel dans les journaux et les librairies, dans lequel il demanda à tout volontaire désireux de le faire de lui envoyer une citation pouvant potentiellement constituer une entrée dans le dictionnaire :

> Make a quotation for every word that strikes you as rare, obsolete, old-fashioned, new, peculiar, or used in peculiar way (...). Make as many quotations as you can for ordinary words, especially when they are used significantly, and tend by the context to explain or suggest their own meaning. (Murray, 2001 : 178)

Cette mise à contribution de tiers dont la démarche était volontaire et non rémunérée fit en sorte que Murray recevait environ 1000 participations par jour, et leur nombre s'éleva à 3 500 000 en 1882.

C'est néanmoins depuis une vingtaine d'années que le crowdsourcing se répand, parfois dans des projets de très grande envergure comme par exemple l'encyclopédie libre, collaborative et multilingue Wikipédia, ou le projet Stardust@home[186] de la NASA, qui permet à des volontaires ayant accès à Internet de rechercher des échantillons de poussière interstellaire (matières solides de l'extérieur du système solaire) grâce un outillage logiciel disponible sur le site du projet. La société Amazon a d'ailleurs développé une plateforme Internet de crowdsourcing sous le nom de Amazon Mechanical Turk[187], où il est possible d'effectuer

[186] Site du projet :
http://stardustathome.ssl.berkeley.edu/

[187] Site du MTurk :
https://www.mturk.com/mturk/welcome

toute sorte de requête inopérable pour un ordinateur, et qui sera effectuée par les internautes volontaires[188] (contre rémunération ou non).

En ce qui concerne les corpus et plus particulièrement l'annotation de corpus via le crowdsourcing, nous citerons une expérience d'Evanini & Zechner (2011), qui utilisèrent les compétences d'une équipe de deux annotateurs professionnels et de onze annotateurs non formés (dont les annotations sont nommées « naive annotations » par les auteurs) pour l'annotation prosodique d'un corpus d'oral spontané d'anglophones non natifs. L'expérience montre (Evanini & Zechner, 2011, sect. 5) que le recours au crowdsourcing, et donc à des annotateurs non formés « est raisonnable et rentable », certains des annotateurs naïfs ayant pu obtenir un degré d'accord avec les annotateurs professionnels convenable. Le résultat le plus intéressant est que le degré d'accord peut être significativement amélioré grâce à un entraînement concis des annotateurs naïfs. Le recours au crowdsourcing pose donc la question du degré d'accord entre annotateurs, qu'ils soient formés ou non. Le degré d'accord est mesuré grâce au coefficient de Kappa[189].

[188] Nous avons par exemple pu retrouver des requêtes de transcription de données orales. Pour accéder à la tâche, il faut d'ailleurs obtenir une qualification « Audio Transcript Verification » via un test. L'accès aux services du MTurk est pour l'instant réservé aux résidents des États-Unis. La méthode MTurk pour la transcription a d'ores et déjà porté ses fruits dans le milieu académique, comme le prouvent plusieurs études ayant eu recours à ce type de crowdsourcing (Keelan Evanini & Zechner, 2011; K Evanini, Higgins, & Zechner, 2010; Gruenstein, McGraw, & Sutherland, 2009; Marge, Banerjee, & Rudnicky, 2010)

[189] Dans une configuration idéale de schéma d'annotation parfait sur des phénomènes parfaitement cernables, les annotations effectuées par n annotateurs devraient être les mêmes. Il est évident que cette configuration n'existe pas ; le problème réside dans les décisions à prendre quant aux annotations délivrées, lesquelles faut-il modifier, garder ou supprimer ? Pour ce faire, il est d'usage de mesurer le taux d'accord entre les annotateurs. Le coefficient de Kappa (Cohen, 1960) permet de mesurer le degré d'accord entre deux personnes lors d'un codage, soit la mesure de la variabilité inter-annotateurs (inter-rater agreement). Il s'agit du Kappa de Cohen. Pour une mesure de l'accord entre plus de deux personnes, on utilise le Kappa de Fleiss. L'équation ainsi que la méthodologie de la mesure du Kappa sont détaillés dans par exemple « Measuring nominal scale agreement among many raters » (Fleiss & Cohen, 1971), où le Kappa (κ) sera de 1 en cas d'accord total et inférieur à 0 en cas de désaccord. Les fluctuations du κ entre 0 et 1 sont corrélées au degré d'accord, plus k est grand et se rapproche de 1, plus le degré d'accord est grand. Une fois le κ mesuré, il permet de répondre ou de résoudre les problématiques suivantes: 1) mesurer l'accord permet de juger de la pertinence de la tâche d'annotation, ainsi que de la qualité du schéma d'annotation. En cas de κ trop faible, il se peut que la tâche soit à revoir, ou que le schéma ne soit pas assez fourni ou rigoureux ; 2) les annotateurs obtenant un κ trop faible peuvent être détectés

Sur le processus d'annotation, nous terminerons en rappelant que l'annotation des corpus ne constitue pas une étape obligatoire, et peut être même considérée comme une étape à éviter. Nous avons vu que ce point de vue était défendu par les représentants de l'école firthienne. Sinclair (1991 : 21) dit à ce propos que « the safest policy is to keep the text as it is, unprocessed and clean of any other codes ». Les raisons qu'avance Sinclair sont les suivantes : le processus d'annotation nécessite un ensemble de codes et de conventions afin que les annotations puissent être exploitées en tant que partie intégrante du corpus ; or il n'y a pas de standards en linguistique et les différentes méthodologies de catégorisations ne font pas consensus. Pour cela, l'annotation d'un corpus reflète la vision des annotateurs et ne sera utile que dans des cas particuliers. Pour les autres cas, l'annotation risque de mener à des biais linguistiques qui altéreraient la qualité des analyses en empêchant les chercheurs de découvrir par eux-mêmes d'autres faits linguistiques que ceux soulignés lors de l'annotation. Les positions de Sinclair et de l'école firthienne sont des choix méthodologiques qui indiquent de fait des positionnements théoriques. Si l'on ne peut que convenir que l'annotation « teinte » le corpus par les théories linguistiques sur lesquelles elle a été élaborée, nous rappelons que les directives de Leech préconisent la prudence grâce à des annotations séparables des données. D'autre part, nous avions également évoqué le cas des corpus oraux qui restent très peu traités par l'école firthienne. Les outils d'annotation ne donnent pas les mêmes résultats selon qu'ils sont appliqués sur des corpus écrits ou des corpus d'oral spontané ; en ce qui concerne par exemple la transcription, les logiciels de reconnaissance vocale sont loin de fournir des résultats aussi concluants que les logiciels de reconnaissance optique des caractères en raison de certaines des spécificités de l'oral, tels les répétitions, les amorces, les tics de langage, l'autocorrection etc.[190]. Ces mêmes phénomènes altèrent également les capacités de la machine à par exemple procéder à un étiquetage morphosyntaxique du corpus.

et les raisons des divergences peuvent être résolues en mieux les formant ; 3) le calcul du κ en supprimant respectivement chaque annotateur permet de détecter les annotateurs consensuels (si le κ baisse en supprimant un annotateur, c'est un annotateur consensuel). Les annotations des annotateurs consensuels peuvent permettre la création d'une annotation de référence.

[190] Les phénomènes de l'oral spontané et les méthodologies que nous avons employées pour notre corpus seront détaillées dans le troisième chapitre.

2.7.2 Transcription des données

Comme le résument Cappeau & Gadet (2010), « tout le monde semble désormais s'accorder sur le fait qu'il il est cognitivement difficile, voire exclu (…) de faire l'économie d'une représentation écrite (une transcription) pour travailler sur des séquences orales et pour développer à partir de là une analyse linguistique. » Nous n'avons effectivement pas retrouvé, ni dans les théorisations en linguistique de corpus, ni dans les travaux sur corpus, de projets de corpus oraux destinés à être analysés ou diffusés sans que ces corpus ne soient transcrits, et le débat semble tranché[191] : les exploitations d'une masse de données orales sans représentation graphique sont fortement limitées.

Traditionnellement, avant l'avènement des outils informatiques contemporains, la transcription des données se faisait au moyen d'une pédale de transcription. Ce type d'appareil permettait de contrôler le flot sonore avec les pieds, et de transcrire avec les mains sans effectuer d'allers-retours chronophages ; toute personne s'essayant à la transcription découvre très vite que le débit moyen de la parole spontanée est trop rapide pour une transcription instantanée. Ceci est d'autant plus valable pour une raison que détaille Blanche-Benveniste (2010 : 16) : la linguistique ne se préoccupe pas de ce qu'un énoncé « veut dire », mais de ce qu' « il dit littéralement »[192], et les linguistes ne peuvent se permettre de transcrire ce qu'ils ont compris, mais doivent transcrire ce qu'ils ont perçu[193] (nous reviendrons ci-dessous sur la notion de perception en transcription). De nos jours, la transcription au moyen de logiciels informatiques semble s'être démocratisée. Comme tout outil informatique, les logiciels de transcription permettent un gain de temps considérable ; en outre, ils permettent

[191] Les auteurs semblent d'accord sur le fait que l'écriture est le seul moyen d'analyser la langue au moins depuis l'ouvrage de l'anthropologue Jack Goody (1979), dans lequel il soutient que la représentation graphique de l'oral permit l'archivage des informations et la possibilité de les manipuler sans restrictions cognitives ; Goody démontre ainsi comment l'écriture fut ainsi directement liée au développement des sciences en général. À propos de l'histoire de l'écriture et de la relation entre écriture, oral et traditions orales, il conviendra de consulter Auroux (1994), Olson (1996) ou Descamps *et al.* (2005) ; Blanche-Benveniste (2010 : 14-19) offre un résumé de quelques pages de la problématique ainsi qu'une bibliographie détaillée.

[192] Blanche-Benveniste reprend cette distinction à la suite de Olson (1996) ; en anglais la distinction est entre les verbes « to mean » et « to say ».

[193] Citant Descamps *et al.* (2005), Blanche-Benveniste (2010 : 17) évoque les travaux des documentalistes qui, lors des transcriptions d'entretiens pour constituer des archives sonores, « négligent généralement une partie de la forme pour retenir l'essentiel de l'information ».

l'alignement des transcriptions. En France, la possibilité d'aligner les transcriptions au signal sonore est évoquée dès 1997 par Blanche-Benveniste (1997 : 90), qui parle de « transcriptions couplées avec les enregistrements ».

Par ailleurs, depuis l'article systématiquement cité de Ochs, « Transcription as theory » (1979), aucun auteur ne défend l'idée d'une transcription neutre et ne reflétant aucun positionnement théorique. Dans son article, Ochs (1979 : 44) écrit: « Transcription is a selective process reflecting theoretical goals and definitions ». Nous proposons dans cette section de développer la notion de transcription avec ce qu'elle implique de démarches dans la constitution d'un protocole de transcription. Pour ce faire, nous proposons de considérer la transcription comme un type d'annotation. Il faut à la fois démontrer ce point de vue et expliquer pourquoi nous l'adoptons. Comme nous l'avons évoqué, la transcription est une théorie, et un protocole de transcription matérialise des positionnements théoriques préétablis : la transcription n'est pas une étape préparatoire des données pour l'analyse mais une analyse en elle-même qui interprète les données et conditionne les possibilités d'analyse futures. Ainsi nous considérons la transcription comme une annotation pour les deux raisons suivantes :

1) il s'agit d'un ajout d'informations à la masse de données initiale ;
2) la valeur de ces données est une valeur interprétative.

En ceci, la transcription en tant que représentation graphique du signal sonore ne diffère pas des autres processus d'annotation. Bien que cette vision des choses soit répandue et incontestée (la transcription est un ajout d'informations interprétatives), nous ne retrouvons pas dans les inventaires d'annotations de la littérature la transcription en tant que processus d'annotation revendiqué. En ce qui concerne la littérature académique française[194], la transcription n'est pas assimilée à l'annotation, et le terme n'apparaît pour ainsi dire jamais en parallèle avec le terme transcription. Seul le glossaire de l'IRCOM définit la transcription comme « une instance particulière de l'annotation, distincte de la glose »[195]. Bien que la

[194] Par exemple, le vol. 14 de « Recherches sur le français parlé » (1997), ou les actes du colloque « Les enjeux de la transcription » (Cahiers de l'université de Perpignan, n° 37/2008, Bilger (2008))

[195] Cf. site de l'IRCOM :
http://ircom.huma-num.fr/site/glossaire.php#t

valeur analytique de la transcription ne soit pas niée, elle est le plus souvent différenciée du processus d'annotation :

> Une transcription orthographique (complétée éventuellement par divers systèmes d'annotations). (C. Blanche-Benveniste, 1997 : 88)

Blanche-Benveniste (1997 : 92) poursuit :

> Elles [les annotations] commencent avec les marques de diverses sortes que l'on porte sur le texte transcrit.

Les experts européens d'EAGLES font de même (EAGLES, 1996a : 4) :

> This transcription is afterwards enriched using different annotation systems aiming at reflecting all the important events that take place in the process of speech production.

Enfin, la situation est similaire dans les ouvrages de référence sur les corpus tels ceux de McEnery *et al.* (McEnery & Hardie, 2011; McEnery & Wilson, 2001), Kennedy (1998) ou Meyer (2002).

Il ne s'agit pas de simplement dénommer un processus dont la valeur interprétative fait consensus. Tout choix terminologique implique des contraintes méthodologiques et théoriques qui constituent le cœur de notre propos : si l'on considère la transcription comme un type d'annotation, le processus se doit alors de suivre les directives proposées par Leech et de fait, les conventions de transcription font office de schéma d'annotation qui doit remplir les conditions suivantes :

1) les codes de transcription créés pour un corpus particulier et qui n'appartiennent pas aux codes orthographiques usuels se doivent d'être amovibles et indépendants de la transcription reposant sur un système neutre tel l'API ou l'orthographe traditionnelle ;
2) les conventions de transcriptions doivent être disponibles aux utilisateurs du corpus ;
3) le nombre et l'identité des transcripteurs doivent être indiqués, ainsi que leurs méthodologies et les outils utilisés ;
4) les conventions de transcription doivent reposer sur les théories les plus neutres et les plus consensuelles possibles ;
5) aucun protocole de transcription ne devra être présenté en tant que standard de transcription.

Si ces conditions étaient systématiquement suivies, nous pensons que les corpus oraux gagneraient en diffusibilité, en lisibilité et la valeur interprétative des transcripts se verrait minimisée. Pour l'exemple, nous donnons ici un extrait d'annotation provenant de la plateforme CLAPI[196] :

Ex 2

FAB c'est finela
JEB c'é[tait] [(finela) biscuit/]
SOP [t` imagines]
FAB [tout c` qui attrait autour]
(0.1)
FAB [ouais]
SOP [tu te] DÉ- déculpabilise euh
(0.2)
JEB pa`ce que là
SOP ((rire)) [(pour manger un gâteau) ah non c'est génial]
JEB [x je on va faire l` pavé pourquoi alors pourquoi y a lance]ment alors i` nous a bien dit pour dynamiser l` marché/
(0.3)

Nous constatons que les directives d'annotation ne sont pas respectées au moins sur les points suivants :

1) la langue de transcription n'est pas neutre en raison d'aménagements graphiques ;
2) les codes annotatifs spécifiques au protocole de transcription sont intégrés aux données et ne sont pas facilement suppressibles.

Nous discuterons donc ici de notre positionnement théorique vis-à-vis de la transcription en général ; nous nous baserons et défendrons les points de vue suivants :

1) la transcription est un processus d'annotation ;
2) la transcription se doit d'être minimale : le terme minimal signifie pour nous qu'une représentation exhaustive de tous les phénomènes oraux n'est ni possible, ni souhaitable ;
3) la transcription ne doit pas altérer les habitudes de lecture, d'une part pour la lisibilité des données, mais surtout pour permettre aux logiciels d'effectuer des recherches qui s'avéreraient impossibles en cas d'aménagements graphiques ;

[196] Réunion de travail entre publicitaires - Lyon Saxe / Enregistrement LSG 35. Disponible sur Clapi : http://clapi.univ-lyon2.fr/

4) le manque à gagner lors d’une transcription minimale est compensé par l’alignement des textes sur le son.

Nous nous contenterons ici de discuter des généralités de la transcription. Les difficultés, choix et spécificités de la transcription de notre propre corpus seront détaillées dans le troisième chapitre.

Enfin, une précision terminologique est nécessaire : Gadet (2008, n. 5) souligne qu’il n’est pas dans les habitudes terminologiques françaises de distinguer entre l’action de transcrire (transcription) et le produit de l’acte, appelé « transcripts ». Nous emploierons désormais le terme au sens défini par Gadet.

2.7.2.1 Reconnaissance automatique de la parole

Avant de discuter des points théoriques de la transcription et puisque nous avons évoqué l’outillage de la transcription, nous tenons à dire quelques mots sur la reconnaissance vocale automatique. Il s’agit d’une opération (aujourd’hui logicielle) grâce à laquelle la parole est analysée puis représentée de manière à pouvoir être lue sur une machine (aujourd’hui sur ordinateur). La reconnaissance automatique de la parole est évoquée pour la première fois dans un article qui date de 1952 (Davis, Biddulph, & Balashek, 1952) et les logiciels se sont aujourd’hui démocratisés et disponibles au grand public.

Depuis une dizaine d’années, plusieurs études ont démontré que la fiabilité des transcriptions logicielles était significativement en-deçà des performances humaines pour deux raisons principales : d’une part, les difficultés de la transcription, comme les variances phonétiques, le débit de parole plus ou moins rapide, les bruits, les disfluences, les agrammaticalités, les mots inventés, les accents etc. ne sont pas encore gérées par les logiciels de reconnaissance de la parole et nécessitent une intervention humaine ; d’autre part, les logiciels les plus répandus parmi le grand public sont élaborés de manière à voir leur performances s’améliorer avec le temps, à condition d’entraînement et d’utilisateur unique, ce qui est loin d’être une option possible pour la constitution de corpus qui comptent généralement plusieurs locuteurs.

Un document du NIST[197] détaille l'évolution du taux d'erreur de mots[198] en reconnaissance vocale depuis 1988 jusqu'en 2009, et ce sur plusieurs types d'oral (le taux d'erreur approximatif est indiqué, ainsi que les évolutions au fil des années) :

1) oral lu : (1988 : 30%), (1991 : 3,8%) ;
2) oral des médias : (1996 : 40%), (2005 : 9 – 20%) ;
3) standards téléphoniques : (1992 : 90%), (2003 : 30%) ;
4) conférences : (2002 – 2010 : 40 – 60%) ;
5) oral conversationnel : (1995 : 80%), (2005 : 50%).

Ces taux sont à comparer avec le taux humain d'erreurs de mots, qui se situe entre 2 et 4%. L'observation du document nous montre deux points essentiels : seule la lecture de textes parvient à des taux d'erreurs comparables aux taux d'erreurs humains ; ensuite, si les performances se sont améliorées durant les premières années de manière parfois fulgurante grâce aux évolutions des technologies et des bases de données, il semble que la situation soit étale depuis quelques années. Nous ne sommes pas en mesure de savoir si ces stagnations sont définitives, mais elles montrent que le recours à la reconnaissance vocale pour la constitution de corpus oraux ne semble pas à l'ordre du jour[199].

2.7.2.1 Fiabilité des perceptions

[197] National Institute of Standards and Technology. Le document s'intitule « The History of Automatic Speech Recognition Evaluations at NIST », et est disponible ici :
http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html

[198] Le taux d'erreur de mots, ou word error rate (WER) en anglais, est l'unité de mesure la plus usitée pour l'évaluation de la performance d'un logiciel de reconnaissance automatique de la parole. Le WER est proportionnel au nombre de mots omis, au nombre de mots substitués et au nombre de mots ajoutés par rapport au nombre de mots de référence.

[199] Les logiciels de reconnaissance vocale sont toutefois exploités dans nombre d'autres domaines, comme par exemple dans le sous-titrage pour sourds et malentendants à la télévision française, obligatoire depuis la loi sur le handicap de 2005 (Loi n° 2005-102). À ce propos, le site Rue89 publiait en 2010 un article sur le sujet, où les exemples d'erreurs sont nombreux : « je suis contre les 17 heures » pour « je suis contre les baby-sitters », ou « as tonne du Cher » pour « Aston Kutcher ».
Nous avons eu connaissance de l'article :
http://www.rue89.com/2010/03/28/le-massacre-des-sous-titrage-pour-les-sourds-144363
grâce à la thèse de Bazillon (2011 : 14), qu'il conviendra de consulter pour de plus amples détails sur la reconnaissance automatique de la parole.

D'après Billières et Gaillard, un interlocuteur type produit entre 100 et 200 mots par minute, soit l'émission de 3 à 4 syllabes ou l'actualisation de 12 à 20 phonèmes par seconde :

> Cela signifie que nous identifions un mot toutes les 400 millisecondes environ en allant le récupérer au sein d'une structure mentale hypothétique dénommée lexique mental comprenant quelque 60 000 unités pour un individu normal. (Billières et Gaillard, 2008 : 174)

Comme nous l'avons dit ci-dessus, la transcription en linguistique a pour objectif la transcription de ce qui a été dit, et non pas de ce qui a été compris. Néanmoins, la réception d'un énoncé oral, que ce soit au cours d'une conversation ou au cours d'une transcription, est toujours l'effet conjugué de deux exercices : la perception et le décodage du message. La transcription n'est pas une activité de transposition de codes ; Bilger *et al.* disent à ce propos :

> Il est bien connu que le linguiste ne recueille pas les sources orales à la manière dont un magnétophone enregistrerait des séquences sonores, mais est constamment amené à faire certaines hypothèses sur ce qui a « vraiment » été dit par le locuteur dont on souhaite consigner les paroles. (Bilger *et al.*, 1997 : 57)

La question de la perception est certes liée à la qualité des enregistrements sonores, mais ceci est loin d'être le seul critère : à l'excellente qualité sonore que l'on peut aujourd'hui obtenir et la possibilité de réécoute indéfinie de courts passages, est couplée la potentialité des logiciels à supprimer ou amplifier certains sons ; certains logiciels comme PRAAT offrent même l'éventualité d'analyse du son par spectrogramme. Ces avancées pourraient laisser croire que la transcription tendrait vers l'infaillibilité ; or si les possibilités du traitement ont facilité la perception du son, la transcription ne s'en tient pas à la représentation graphique des phénomènes perçus. Comme le notent Bilger *et al.*, la transcription ne saurait être indépendante du processus humain de compréhension et d'interprétation du son :

> La transcription d'un énoncé oral est toujours le résultat simultané d'une activité de signification arrimée sur une activité de perception. En effet, un énoncé même correctement perçu peut être transcrit de façon erronée. (Bilger *et al.*, 1997 : 59)

Les évolutions technologiques comportent même certains risques. Mondada (2000 : 2) remarque que « le fait de réécouter de façon indéfiniment répétée des fragments d'oralité exerce un effet de loupe agrandissant des phénomènes qui autrement échappent à l'ouïe, les faisant littéralement émerger comme tels », et rapporte des travaux ayant évoqué les effets de « décontextualisation et d'extraction de l'oral par rapport à son contexte singulier de

production », ou de « fixation du flux dynamique ». Ces remarques signifient que « l'effet de loupe » risque de faire de la transcription un acte de calque de ce qui a été perçu, et le transcript s'éloignerait alors de la représentation de ce qui a été dit (voir Ex 7 qui illustre notre propos d'une transcription se rapprochant davantage de ce qui a été perçu de ce qui a été dit).

En termes du schéma de Jakobson, la transcription est la représentation graphique d'une communication entre un locuteur et un destinataire, et les six fonctions énoncées par Jakobson pour assurer la réussite d'une communication humaine entrent en jeu lors d'une activité de transcription[200]. Ainsi et malgré la possibilité de représenter les sons perçus très rigoureusement, la part inhérente d'interprétation liée à la transcription fait que nous pouvons être confrontés à une multitude de transcriptions pour le même énoncé, ceci étant encore plus flagrant pour les corpus d'apprenants, d'enfants ou de parole pathologique[201].

Afin de s'en rendre compte, nombre d'auteurs se sont essayés à la vérification de la transcription des corpus (par exemple Giovannoni & Savelli (1990), ou Cappeau (1997) qui dresse et discute d'une liste d'erreurs récurrentes dans les transcriptions de personnes peu expérimentées). Dans une étude détaillée, Pallaud (2002 : 288) a effectué une vérification de la transcription de 12 corpus de français parlé, vérification qui a révélé « un certain nombre d'erreurs très variables en quantité et en types d'un corpus à l'autre ». La majorité des erreurs concernent des éléments oubliés, et l'ajout d'éléments est très rare : « on retrouve la très grande prédominance des oublis dans les pauses silencieuses et remplies ainsi que dans les

[200] Nous citerons ici – pour l'exemple – Descamps *et al.* (2005: 437), qui préconisent que le transcripteur le mieux à même de transcrire un entretien serait celui qui a mené l'enquête car « il connaît les différentes phases de l'entretien et peut anticiper son déroulement, il est capable de restituer les propos inaudibles, de reconstituer les questions trop elliptiques ou abrégées, d'orthographier correctement les noms propres des personnes ou des institutions ; il connaît la voix de son interlocuteur et de mémoire, peut retranscrire les inflexions, interpréter les émotions ou les différentes tonalités. Il sait « de quoi il est parlé » et transcrit beaucoup plus vite que celui qui ignore ce dont il est question. Les risques d'erreur, de mécompréhension ou d'affabulation sont beaucoup moins importants ». Cette suggestion peut être rattachée à la fonction référentielle du schéma du Jakobson.

[201] Notre propos ici se limite à souligner la complexité du processus perception-décodage-compréhension dans le cadre de la transcription. Ce processus suscite l'intérêt de nombreuses sciences en linguistique, sciences sociales ou en médecine, et nous ne le détaillerons pas ici. Nous renvoyons à par exemple l'article « Approche pluridisciplinaire de la perception de la parole » (Billières & Gaillard, 2008) qui constitue une porte d'entrée à la question, et dans lequel les auteurs assurent le rôle des facteurs extralinguistiques du processus, sans que la science soit en mesure, à l'heure actuelle, d'en démontrer le poids et les fonctions exacts.

tours de parole. De même, les éléments lexicaux, onomatopées et interjections, beaucoup plus souvent oubliées qu'ajoutées, constituent le tiers des modifications apportées aux textes ». Que ce soit des erreurs de transcription ou des transcriptions différentes bien que conformes aux conventions mises en place, les raisons tiennent dans des perceptions et des interprétations différentes.

2.7.2.2 Conventions de transcriptions

Les conventions de transcription (ou protocole de transcription) sont un ensemble de choix que nous répartissons sur trois problématiques qui constituent le cœur de la question de la transcription :

1) le type de transcription ;
2) la segmentation des données ;
3) le choix des données à représenter.

Les choix effectués lors de l'élaboration du protocole dépendent des objectifs d'exploitation du corpus à constituer. Dans une rétrospective analytique des conventions de transcription des principaux corpus oraux de France, Cappeau & Gadet (2010) mettent l'accent sur la diversification des solutions proposées aux problématiques de la transcription au sein des conventions respectives du CFPP2000 et de CLAPI, et les auteurs considèrent que ces différences ne sont pas surprenantes, « vu la relation entre conventions de transcriptions et objectifs d'exploitation des corpus ». La transcription sera orthographique ou en API selon que le corpus est destiné à des fins syntaxiques ou phonétiques ; la segmentation du flux sonore différera dans les corpus d'analyse conversationnelle de celle effectuée dans les corpus de dialectologie ; enfin, chaque discipline choisira de représenter tel ou tel phénomène de l'oral par lequel elle est concernée en délaissant les autres car, comme le dit Blanche-Benveniste (1997 : 88) :

> Lorsqu'il s'agit de représenter par écrit des données orales, ce respect des données exige qu'on fasse des choix (on ne peut pas être fidèle à tous les phénomènes en même temps).

Les transcripts d'un enregistrement sonore sont donc subordonnés au protocole de transcription, lui-même tributaire des objectifs du corpus. Mondada résume :

> La transcription n'est pas autonome par rapport aux conventions auxquelles elle obéit : sans elles, elle n'est pas lisible ; avec elles, elle est située dans un réseau beaucoup plus large, de communautés scientifiques, de présupposés spécifiques, d'autres textes produits par les mêmes conventions et donc de phénomènes rendus reconnaissables et comparables à travers les corpus. (Mondada, 2008 : 80)

Dans cette section, nous allons discuter des trois problématiques énoncées ci-dessus ; elles seront discutées au vu de ce que nous venons d'exposer, à savoir le lien entre choix protocolaires et objectifs du corpus.

Types de transcriptions

À la question de savoir quel type de langue graphique de base utiliser pour représenter l'oral, le choix se situe entre soit l'orthographe standard, soit la transcription phonétique en utilisant par exemple l'Alphabet Phonétique International. Le choix d'une de ces deux représentations n'implique pas le suivi des standards respectifs à la lettre : les conventions peuvent recommander le respect de l'orthographe standard comme c'est le cas pour les conventions de transcription du GARS[202] qui préconisent une transcription « sans aucun trucage orthographique. Les grands dictionnaires servent de référence pour les mots de la langue, les noms propres, les interjections et les onomatopées », mais d'autres projets qui se basent sur une transcription orthographique ne suivent pas tous la démarche de l'orthographe standard référencée dans les dictionnaires. Afin d'illustrer notre propos, nous allons voir ce qu'il en est de ce point dans trois conventions de transcription du français qui se basent sur la transcription orthographique[203].

Les données de la plateforme CLAPI, majoritairement transcrites selon les conventions ICOR, permettent d'adapter l'orthographe lorsque les caractéristiques phoniques le requièrent, comme par exemple « mouais » pour « ouais » ; des symboles sont insérés au sein des mots

[202] Les conventions de transcription du GARS sont disponibles ici : http://sites.univ-provence.fr/delic/corpus/index.html

[203] Pour une vue d'ensemble plus détaillée des conventions de transcription, voir les travaux de Cappeau & Gadet (Cappeau & Gadet, 2010, 2012).

pour signaler l'allongement des sons, et les majuscules sont utilisées pour marquer l'accentuation, comme dans l'exemple suivant[204] :

Ex 3

SIM h (0.3) moi/ hm c'est **PLAte l'éco:le** criss que c'est d[ull]

Les élisions sont notées :

Ex 4

SIM qu'est ce c'est ça [**c`t` affaire** là]
SIM [pas de me v]oir mais de **t` marier** ((rires))
MAN **pauv`** p`tit gars

ainsi que les aphérèses :

Ex 5

MAN =of course celui là i` i` répandrait plusieurs millions d'exemplaires **`fin** ça changerait

D'autres conventions, comme celles de Valibel[205], ajoutent à la transcription orthographique une transcription en SAMPA[206] lorsque la forme « fait l'objet d'une prononciation particulière (ressentie comme marquée par le transcripteur) » (protocole Valibel : 10), comme dans les exemples suivants :

Ex 6

je vais toujours voir le football [fOtbal] avec le voisin
il m'a fait une scène [se~n]
il faut qu'il soit [swaj] là
ça m'a coûté cent euros [sa~z2RO]
il a eu un infarctus [e~fRaktys]
on va manger des courgettes [guRZEt]?

[204] Tous les exemples concernant les conventions ICOR proviennent du corpus 2/43, « Appels téléphoniques ~ conversations entre amis - call friends », disponible sur la plateforme CLAPI : http://clapi.univ-lyon2.fr

[205] Centre de recherche sur les Variétés linguistiques du français en Belgique. Les conventions de transcription sont disponibles ici : http://www.uclouvain.be/cps/ucl/doc/valibel/documents/conventions_valibel_2004.pdf

[206] SAMPA: Speech Assessment Methods Phonetic Alphabet. Il s'agit d'une alternative à l'API que Valibel a abandonné en raison de l'incompatibilité de la police avec le format de leurs documents.

> il a dégringolé [dedRe~gOle] dans les escaliers ?
> c'est une faute de construction [kRo~stRyksjo~]

En ce qui concerne les conventions du projet CFPP2000, les auteurs ont noté certaines élisions comme « j'sais » ou « t'arrives », jugeant qu'elles étaient « bien répandues, gagnant peu à peu la bande dessinée, le roman, les blogs » (protocole CFPP2000 : 13), mais se refusaient aux transcriptions « i » ou « iz » pour « il » ou « ils », parce qu'elles n'appartiennent pas aux « habitudes orthographiques du français[207] » (protocole CFPP2000 : 13).

Ce rapide aperçu de quelques conventions quant au respect de l'orthographe nous a permis de voir que quand bien même le choix d'une transcription orthographique est effectué, les transcripts des différentes équipes n'en sont pas pour autant similaires, loin de là ; Cappeau & Gadet (2010) disent que la revue des conventions de transcription qu'ils ont effectuée « frappe par la diversité des pratiques ». Ainsi le choix de l'orthographe comme langue graphique de transcription ne constitue pas un critère majeur de différenciation entre les transcripts, alors que l'on aurait pu s'attendre à des transcripts homogènes puisque l'orthographe est stable. Nous verrons que les deux autres problématiques de la transcription que nous discuterons ci-dessous (segmentation et chois des données à représenter) modifient énormément les habitudes orthographiques de l'écrit lors des transcriptions. Il faut également ajouter à ceci le nombre importants de corpus où des balises, ou des codes d'annotations n'appartenant pas aux standards orthographiques sont insérés dans les transcripts (voir Ex 3 ou Ex 7).

Le choix de la langue de transcription dépendra, là encore, des objectifs du corpus. L'API convient aux études phonétiques, phonologiques ou prosodiques, et le recours à l'orthographe a lieu pour les études syntaxiques, lexicales et discursives. Blanche-Benveniste (1997 : 97) souligne cette idée en mettant l'accent sur la terminologie anglaise, ici légèrement plus précise que le français :

> On se sert de l'Alphabet Phonétique International pour les études consacrées spécifiquement au signal sonore (*Speech Research*). Mais il est très rare de rencontrer des transcriptions de langue parlée (*Spoken Language*) faites systématiquement en

[207] À ce propos, Cappeau & Gadet rapportent que la prononciation [i] pour « il » est d'ailleurs attestée depuis des siècles ; les auteurs (2010, n. 6) citent également Damourette & Pichon qui remarquèrent et commentèrent cette prononciation au sein de la bourgeoisie cultivée de Paris.

> API. Cette différence entre *Speech Research*, portant sur l'aspect phonique du langage, et *Spoken Language Research*, (difficile à rendre en français), portant sur l'étude grammaticale, lexicale ou discursive de productions orales, a représenté pendant longtemps une frontière majeure entre deux disciplines.[208]

En 1996, Llisteri diffuse au sein d'EAGLES[209] les recommandations en ce qui concerne les corpus oraux, dans lequel nous retrouvons un tableau comparatif entre « spoken research » et « speech research », mais où l'on remarquera que le terme « corpus linguistics » correspond pour l'auteur au « spoken language » (EAGLES, 1996a : 5) :

Tableau 5 : Différences d'approches entre « spoken research » et « speech research »

	Corpus Linguistics	**Speech Research**
Materials	Unprepared, unelicited speech	Controlled, elicited speech
Scope	Discourse, dialogue	Utterance
Recordings	Natural environment	Controled environment
Transcription	Orthographic enriched	Phonetic and orthographic
Oriented	Symbolical categorical	Speech signal, temporal

La transcription orthographique, largement plus répandue au sein des corpus français (cf. l'inventaire de Cappeau & Seijido (2005), est pourtant soumise à maintes critiques ; elle est accusée de déformer les spécificités de l'oral car normée pour l'écrit, et ne remplirait plus ses objectifs primaires. Si nous ajoutons à cela le potentiel plus rigoureux d'une transcription phonétique, il semble paradoxal que la plupart des choix soient portés sur une transcription orthographique, même quand la matière phonique est l'objectif ou l'un des objectifs du corpus. Indépendamment des objectifs du corpus, nous retrouvons dans les diverses conventions de transcription trois arguments récurrents en faveur de la transcription orthographique qui pourraient expliquer sa prédominance :

[208] Campione & Véronis (2001 : 1) proposent, avec réserves, les termes de « oral spontané » et de « parole de laboratoire ».

[209] EAGLES : Expert Advisory Group on Language Engineering Standards. Il s'agit d'une initiative de la Commission européenne pour la mise en place de standards pour la constitution de bases de données linguistiques, ainsi que de leurs de constitution et d'exploitation. L'ensemble des documents EAGLES sont disponibles sur le site officiel :
http://www.ilc.cnr.it/EAGLES/browse.html

- le coût (temporel ou humain, ou les deux) de la transcription phonétique, beaucoup plus important que celui d'une transcription orthographique ;
- la lisibilité des transcripts ;
- les possibilités de requêtes, de recherches et d'analyses limitées par une transcription phonétique.

À propos du premier point, Leech *et al.* (1997 : 90)[210] qualifient la transcription orthographique de « pseudo-procédure, dont la seule excuse est le coût prohibitif qu'il y aurait à tenter quoi que ce soit d'autre ». En effet, Véronis (2000 : 5) estime que pour la transcription d'une minute de parole, assortie à un alignement des phonèmes sur le signal, le temps nécessaire serait entre dix et quinze heures de travail. Nous ne discuterons pas plus que cela de cet aspect des choses ; si la nécessité d'une transcription phonétique est établie mais qu'elle s'avère impossible en raison des limitations de moyens matériels, cela ne relève plus de la linguistique de corpus, scientifiquement parlant.

En ce qui concerne le second point, les choix sont très différents selon les équipes, comme nous l'avons vu plus haut, quant au recours ou non aux « trucages orthographiques »[211] qui, s'ils sont utilisés, limitent les possibilités de requête et d'analyse tout comme l'API et comme nous l'évoquions dans notre troisième point. Ces trucages sont évoqués par Gadet (2008 : 39-44) qui discute des avantages et des pertes du recours à ce qu'elle appelle les aménagements graphiques. Les arguments qu'elle propose contre ces aménagements sont en premier lieu de type sémiotique : le recours à des codes graphique non habituels engendre une pénibilité dans à la fois dans l'écriture et dans la lecture des transcripts, et Gadet prouve par des exemples l'inévitable inconsistance de ces choix dans des corpus à partir d'une certaine taille. Ce point est également évoqué par Cappeau & Gadet :

> Outre la constante vigilance exigée pour la notation et l'effort requis pour le lecteur (qui ne sauraient ni l'un ni l'autre constituer une cause de rejet), on soulignera surtout la difficulté de maintenir la même exigence tout au long de la notation écrite des

[210] Cités par Campione & Véronis (2001 : 2).

[211] « Trucages orthographiques » est la terminologie de Blanche-Benveniste & Jeanjean (1987). Dister & Simon (2008, n. 5) rapportent avoir également retrouvé dans la littérature les terminologies suivantes : « bricolage orthographique », « aménagement graphique » ou « bâtards phonético-orthographiques ». Nous avons pour notre part relevé « jeux typographiques » chez Delais-Roussarie (2002 : 10).

> phénomènes. Cette représentation confronte aussi au phénomène de la catégorisation des locuteurs. (Cappeau & Gadet, 2010)

En second lieu, et son argument est ici de type sociolinguistique, Gadet discute du risque de catégorisation des locuteurs :

> Le lecteur fait immanquablement des inférences en lisant des transcriptions comme *j'vous sers un p'tit thé ?* ou *toucheu pas à mon fliqueu*, et ces inférences conduisent immanquablement à catégoriser, stéréotyper, caricaturer (populaire pour le premier exemple, méridional pour le second). (Gadet, 2008 : 41)

Cela ne signifie pas que Gadet soit contre les aménagements graphiques, car elle poursuit en discutant des cas où ils sont nécessaires, et sa conclusion essentielle est que la transcription en tant que « chaîne de gestes théoriques », ne peut atteindre l'idéal de neutralité auquel certains ont aspiré, et qu'elle se doit donc d'être consciente et responsable.

Segmentation

L'une des problématiques du travail sur des données orales par le biais d'une représentation graphique est l'inexistence de délimitations relativement nettes et consensuelles par rapport à celles qui sont consécutives de la ponctuation de la langue écrite. Le recours à la ponctuation des transcripts a donc été naturellement envisagé comme solution pour la segmentation du flux oral en unités davantage représentatives des spécificités de l'oral, et qui feraient le parallèle avec les notions de « mot », « phrase », « paragraphe » ou « texte », notions d'ailleurs elles-mêmes encore floues. Ceci bien qu'il soit maintenant de notoriété que :

> Une des notions qui « saute », c'est celle de « phrase » ; impossible de découper dans le parlé quelque chose qui corresponde à la notion de phrase pour l'écrit » (Claire Blanche-Benveniste & Jeanjean, 1987 : 89).

Avant de discuter ci-dessous de la ponctuation des transcripts ainsi que des autres procédés de segmentation en unités, il est nécessaire d'indiquer que la nature des unités minimales varie énormément selon les solutions proposées, et que cette nature est souvent tributaire, là aussi, des objectifs du corpus. Précisons que la « controverse des unités », selon le terme de Mondada (2000 : 3), est une question qui s'est posée dès la constitution des premiers corpus oraux modernes, et qui n'a toujours pas de réponse définitive ; Mondada souligne son propos en rapportant la multiplication des propositions. Ainsi retrouvons-nous plusieurs types de solutions :

- La segmentation syntaxique

Nous citerons par exemple le projet Rhapsodie[212], dont les modalités de segmentation syntaxique sont détaillées par Benzitoun *et al.* (2010). Ce type de segmentation, outre les problématiques liées à la segmentation syntaxique traditionnelle de l'écrit, est confronté aux spécificités syntaxiques de l'oral, qui sont présentées et discutées en détail par Bilger *et al.* (1997 : 74-83), et sont résumées sous la dénomination de « segments flottants », soit des séquences structurellement ambiguës en raison de :

1) impossibilité de trancher si la séquence est à rattacher à son contexte gauche ou droit ;
2) séquence dont l'appartenance syntagmatique est floue car située entre deux niveaux hiérarchiques différents ;
3) séquences dont la continuité syntaxique est brisée par une autre séquence ;
4) séquences incomplètes.

Dans le strict cadre d'une segmentation pour la transcription (et non pour une analyse syntaxique finie), ces difficultés sont en grande partie résolues grâce aux traits prosodiques des séquences, accessibles via la réécoute des passages pouvant poser problème.

- La segmentation en tours de parole

Cette segmentation concerne principalement les corpus destinés à l'analyse conversationnelle. Elle a de fait été adoptée par les premiers linguistes interactionnistes (H. Sacks, Schegloff, & Jefferson, 1974) (voir 0 et le système de transcription de Jefferson sur lequel nous reviendrons ci-dessous), et les difficultés auxquelles est confrontée une segmentation en tours de parole sont extrêmement délicats à résoudre pour un transcripteur :

1) prises de parole simultanées ;
2) tentatives de prise de parole ;
3) la présence de backchannels[213] verbaux ;

[212] Projet financé par l'Agence Nationale de la Recherche, « consacré à l'élaboration d'un corpus de référence de français parlé muni d'annotations prosodiques et syntaxiques semi-automatiques », tel que l'on peut lire sur le site :
http://rhapsodie.risc.cnrs.fr/fr/

[213] Cf. ***.

4) interruptions de la part de ou des interlocuteurs, interruptions auxquelles le locuteur initial peut réagir par le silence (cas où il faudra traiter la syntaxe non achevée), par la reprise de parole (syntaxe discontinue), par une incise (syntaxe disloquée) ou par continuer son discours sans prendre en compte l'interruption, du moins sur le plan syntaxique (dans ce dernier cas, la variation prosodique sera à interpréter par le biais de l'interruption, en plus de devoir traiter la question du chevauchement de parole).

Ces difficultés sont encore accrues par la présence de plusieurs locuteurs, comme dans le cas de transcription d'émissions audiovisuelles (voir par ex. Bazillon, 2011). Loin d'être résolues, ces problématiques « font l'objet d'analyses contradictoires dans la littérature », selon les termes de Mondada, qui poursuit en parlant de ces phénomènes :

> [Ils] demandent une analyse des unités conversationnelles qui prenne en compte les postures de « locuteur en train de parler », de « locuteur entrant en compétition pour le tour », de « locuteur qui s'apprête à prendre la parole » ou d'« auditeur attentif », où le fait de considérer le « tour » comme une unité pertinente ou non, de le traiter comme une unité structurelle ou comme un accomplissement pratique, de le définir comme une unité minimale ou comme une entité articulée en « unités de construction du tour » déclenche des interprétations très différentes des lignes ou des paragraphes de la transcription. (Mondada, 2000 : 4-5)

Dans certains cas, la transcription n'est pas confrontée à la problématique de la prise de parole : il s'agira de corpus dont les données ont été enregistrées par un enquêteur, qui n'enregistre qu'une seule personne. Dans ces cas-là, les transcripteurs tranchent plus facilement en favorisant la parole de l'enquêté ; ceci ne signifie pas que le discours du locuteur se construit indépendamment de l'interaction avec l'enquêteur. Durant nos propres enregistrements, nous avons pu constater que les simples backchannels, même les non-verbaux d'entre eux, influaient sur le cours de l'entretien et ceci se manifestait par des auto-interruptions, des reprises ou des reformulations.

Ce type de corpus constitués d'interviews a, d'après Cappeau & Gadet, longtemps été majoritaire :

> Les interviews ont longtemps constitué la situation la plus représentée (de façon écrasante), avec l'idée – soit naïve soit hâtive – qu'existeraient des positions de recueil neutres. L'accent désormais mis davantage sur les recueils dans des conditions écologiques, exploitables à des fins interactionnelles ou autres, a quelque peu changé la donne. (Cappeau & Gadet, 2010)

Si le recueil de données « dans des conditions écologiques » est effectivement en vogue, la question de la segmentation en tours de parole ne devrait donc plus concerner uniquement les interactionnistes. En ce qui concerne les corpus spécialisés de parole pathologique ou d'apprenants, nous n'avons pas connaissance de telles méthodes de recueil à l'heure actuelle, pour des raisons évidentes.

- La ponctuation

Les choix concernant la ponctuation ou non des transcripts, ainsi que des systèmes de ponctuation à utiliser sont aussi nombreux et divers qu'il y a de conventions de transcription. Les disparités concernent aussi les phénomènes qu'il faut annoter : pauses, ruptures syntaxiques, traits prosodiques etc. Cappeau & Gadet (2010) repassent en revue les systèmes de ponctuation des corpus oraux français et nous n'y reviendrons pas. Nous tenterons ici d'examiner les motivations d'une ponctuation des transcripts, et d'en tirer les conclusions idoines.

Cappeau & Gadet résument, à la suite de Catach (1980), les fonctions de la ponctuation :

> - elle participe à l'organisation syntaxique (qui se manifeste par les rôles de séparateur et d'organisateur - comme l'indication des hiérarchisations entre constituants) ;
> - elle entre en correspondance avec l'oral (indique le suprasegmental) ;
> - elle apporte enfin un supplément sémantique.

En ce qui concerne l'organisation syntaxique des transcripts, la pertinence de la ponctuation n'est pas avérée, tel que le montre une étude citée par Campione & Véronis, selon laquelle l'examen de la ponctuation d'un corpus oral a montré des ruptures syntaxiques là où il y avait continuité syntaxique, et une absence de ponctuation lors de ruptures :

> Taylor (1996) étudie la transcription orthographique ponctuée du Lancaster/IBM Spoken English Corpus (SEC), et montre que 47,2% des frontières prosodiques ne correspondent pas a une ponctuation, tandis que 17,1% des ponctuations ne correspondent pas a une frontière prosodique. (Campione & Véronis, 2001)

D'autre part, nous considérons que quand bien même la ponctuation pourrait efficacement jouer son rôle de délimiteur syntaxique, ce rôle relève d'une annotation syntaxique et non pas du rôle de la transcription. Ainsi le rôle syntaxique de la ponctuation des transcripts est à la fois inefficace et hors-sujet, s'il n'est pas clairement assumé en tant qu'annotation syntaxique.

Reconsidérons maintenant les deux seconds points : pointage suprasegmental et apport sémantique. Le hors-sujet est à notre sens encore plus flagrant car nous ne pouvons y voir qu'une analyse non revendiquée car non assumée en tant qu'annotation à part entière. Il faut également prendre en compte que ceci ouvre la porte à l'instauration de systèmes de codage exponentiels car comme le note Blanche-Benveniste, le strict système de ponctuation conventionnel est très insuffisant pour refléter l'étal des « différentes forces illocutoires » [214] :

> Le point d'exclamation, la virgule, la majuscule ou les guillemets fournissent des équivalents approximatifs de plusieurs sortes de phénomènes oraux. Mais on sait que ces équivalences sont en trop petit nombre pour pouvoir refléter le grande diversité des effets de l'oralité, comme par exemple l'accent d'insistance, l'allongement, la montée de la voix, le changement de débit et de tout ce que l'écriture est incapable de représenter, comme le ton ironique ou les différentes forces illocutoires. (Blanche-Benveniste, 2010 : 19)

D'autre part, la valeur essentiellement subjective d'un apport sémantique aux transcripts va de soi : Campione & Véronis (2001) rapportent d'ailleurs une étude selon laquelle le désaccord entre annotateurs prosodiques est de 27% des cas, les auteurs poursuivent en notant que ceci « impose des relectures par des annotateurs multiples accroissant encore le coût global de la tâche ».

Nous considérons ainsi que la ponctuation reflète la volonté implicite de transcrire le plus possible, à défaut de pouvoir tout retranscrire ; volonté sur laquelle nous partageons l'avis de Cappeau & Gadet :

> On peut se demander si les linguistes ne font pas preuve de naïveté ou d'une certaine arrogance lorsqu'ils espèrent fournir une transcription qui refléterait l'oralité et permettrait d'emblée une analyse. Ils semblent ainsi vouloir concilier des exigences

[214] D'autant que dans le jeu d'apport sémantique par un biais non alphabétique, il ne semble pas qu'il y ait de limites à la créativité humaine, comme en témoigne le nombre incommensurable d'artifices typographiques, illustratifs ou iconiques employés dans la communication virtuelle (forums, réseaux sociaux, messageries) employés pour l'expression de ce qui n'a pu, ou de ce qu'on a pas voulu exprimer selon les normes de l'écrit (voir les travaux de Marcoccia (2000) sur la typologie des émoticônes, leur caractéristiques et une analyse de leur fonctionnement). Un système de ponctuation spécialement élaboré pour un protocole de transcription ne diffère pas, selon nous, ni sur le fond ni sur la forme, des systèmes élaborés par les communautés virtuelles. Nous pensons que coupler des systèmes sémiotiques non conventionnels et non alphabétiques à la transcription transforme les conventions de transcription en systèmes binaires dont le risque qu'ils soient trop spécifiques et jamais exhaustifs, est trop grand.

> opposées et atténuent le risque qu'une transcription trop complexe fait courir : plus le nombre de phénomènes notés est élevé, plus est grand le risque d'une notation incomplète. (Cappeau & Gadet, 2010)

Nous terminerons par un rôle que nous n'avons pas évoqué : celui des la lisibilité des transcripts. Nous avons en effet remarqué qu'il était assez usuel dans la littérature, de présenter des textes non ponctués afin de prouver la difficulté de lecture dans ces cas-là. Là encore, nous renvoyons à la section 2.7.2.3 où nous discuterons de la lisibilité des transcripts.

Choix des données à représenter

Décider des données à représenter et des données à supprimer s'agit sans doute de la problématique la plus ardue à laquelle sont confrontés les transcripteurs. Les choix protocolaires quant au niveau de transcription à adopter sont encore une fois en lien avec les objectifs du corpus. Un corpus constitué dans des visées discursives ou syntaxiques sera transcrit au niveau un, voire deux (voir ci-dessous pour le détail des niveaux de transcription), et les transcriptions seront effectuées sans précisions sur la prosodie ou la longueur des pauses et des hésitations, précisions dont l'utilité se manifeste dans des corpus à visées dialectologiques ou interactionnelles, qui tendent vers des transcriptions de niveaux trois et quatre. Les transcripts des différents corpus s'avèrent donc extrêmement divers, tant au niveau de la forme qu'au niveau des phénomènes représentés, en raison de la pluralité des objectifs et des possibilités d'exploitation d'un corpus ; des études plus détaillées ont constaté la diversité des pratiques comme celle de Cappeau & Gadet, dans laquelle les auteurs résument :

> On pouvait penser que la réflexion engagée sur les corpus oraux ainsi que le regret de ne pas disposer d'un vaste corpus auraient favorisé des convergences dans les pratiques, mais les visées d'exploitation des données imposent des options souvent divergentes dans les choix de situations et dans les informations considérées comme essentielles. (Cappeau & Gadet, 2010)

Comme nous l'avons vu plus haut en citant l'étude de Pallaud (2002), la tendance va vers une épuration qui reflète un idéal incarné par la langue écrite. Blanche-Benveniste écrit à ce propos :

> La première tentation est de ramener les productions orales au format de l'écrit, en les débarrassant de certaines caractéristiques jugées négatives : impression que les locuteurs hésitent, répètent et bafouillent ; qu'ils ne savent pas bien ce que sont les

> « phrases », qu'ils ne marquent pas les délimitations syntaxiques qu'on attend d'eux et qu'ils utilisent l'intonation comme une compensation obligée aux déficits de leur syntaxe. Ces phénomènes gênent l'étiquetage morpho-syntaxique des textes et l'établissement d'arbres syntagmatiques tels qu'ils ont été conçus pour le français écrit. D'où la tentation de « normaliser » les transcriptions de l'oral, afin de les rendre plus aisément analysables : enlever les hésitations, les incomplétudes et les répétitions ; mettre de la ponctuation afin de retrouver les unités syntaxiques familières de l'écrit ; bref, réduire la langue parlée aux formats de la langue écrite, comme si seule la langue écrite normalisée pouvait recevoir une description grammaticale cohérente. (Blanche-Benveniste, 2005 : 47-48)

Néanmoins, vouloir représenter un maximum d'informations est un autre extrême dont les conséquences ne sont pas moins graves. Un énoncé est le résultat d'un processus beaucoup trop complexe, riche et subjectif pour que l'on puisse tout représenter graphiquement.

Pourtant, nombre d'études font le choix de transcrire des phénomènes inévitablement choisis subjectivement. Ainsi, pour la transcription d'un corpus oral tchèque, Psutka *et al.* (2006) décident-ils de transcrire « les bruits de langue, les toussotements, les rires, les bruits respiratoires, l'inspiration et les clapotements de lèvres ». Le système de transcription de CHILDES, base de données que nous avions évoquée en 0 et sur laquelle nous reviendrons en détail dans le quatrième chapitre, propose des codes pour la transcription de 240 « communicators »[215], dont voici quelques exemples et leurs transcriptions :

Tableau 6 : Exemples de « communicators » de CHILDES et leurs transcriptions

la joie : ah	la découverte : ahhah
la sympathie : aw	le triomphe : ha(h)
l'amusement : heehee	« je ne sais pas » : emem
le « non catégorique : nuhuh	la peur : (y)eek
la douleur : ouch	la pitié :tut

Les possibilités infinies de telles transcriptions sont à nos yeux un danger pour plusieurs raisons ; en premier lieu ce sont des conventions élaborées et nécessitent pour les utilisateurs du corpus qui n'ont pas participé à sa construction un apprentissage dont l'utilité restera ponctuelle ; en second lieu, leur valeur interprétative ne peut que marquer et orienter le

[215] Les « communicators » sont présentés dans le manuel de transcription de Childes par MacWhinney (2008 : 49-50). L'ensemble des 240 « communicators » est disponible dans le lexique pour l'anglais, joint lors du téléchargement du logiciel Clan.

corpus ; enfin, elles encombrent le corpus de sorte que même un utilisateur qui voudra se contenter des données linguistiques pourrait ne plus les discerner dans le flot typographique qui résulte de telles transcriptions.

Nous illustrerons ce dernier point par un extrait du corpus transcrit selon le système de transcription élaboré par Gail Jefferson[216]. Le système de Jefferson[217] préconise la transcription des chevauchements, des pauses brèves, des pauses en secondes, des reprises, des intonations finales montantes et descendantes, des interruptions, des accélérations et ralentissements du débit par rapport au standard du locuteur, des chuchotements, des paroles criées, de l'emphase, des inspirations et expirations audibles et laisse libre cours aux transcripteurs d'annoter les phénomènes non verbaux. Voici ce que donne un extrait de transcription[218] :

Ex 7

[216] Gail Jefferson (1938-2008) fut une linguiste, sociologue et ethnométhodologue qui contribua à l'émergence de l'analyse conversationnelle avec Sacks et Schegloff. Elle élabora un système de transcription qui posa les bases de la transcription aux États-Unis, à l'image de ce que feront Blanche-Benveniste & Jeanjean (1987) en France.

[217] Disponible ici :

http://www.transana.org/support/onlinehelp/team1/transcriptnotation1.html

[218] Fichier 4-18nh.doc, extrait du Corpus Jefferson, disponible sur la plateforme Talkbank:

http://talkbank.org/media/CABank/Jefferson/Watergate/0word/

Nixon: __|__ He's obviously on the kick iv uh*hhmhh* ((swallow)) ·tchk!
( ): | °·*he:h*°
| (0.3)
Nixon: __|__ saving 'im↑sel**f***hh*=
( ): =°*hmh*°=
Nixon: =↓$_{m}$an[d uh↓]
[ ]=
Hald: [°Yeah°]
Nixon: = *hh*mgh*hh*m (0.2) °·*hoopeh yehehaahehaa*° °°·p·t·°° *h*Uh en thė: U.S.
Att↑orney is gunnuh have=tough ↑problem*hhmh but hmh* $_{m}$I think thih
____ U.S. Attorney ul ·*huhhuhhe* my guess is will:*h*=

Pour ces raisons, les experts d'EAGLES (1996a : 15) ont défini quatre niveaux de transcription, de un à quatre, les transcriptions des niveaux supérieurs étant plus détaillées dans les niveaux supérieurs :

> Niveau 1 : la transcription est orthographique avec une ponctuation minimale, sans notation d'informations interactionnelles comme le tour de parole.
>
> Niveau 2 : la transcription est en orthographe enrichie avec des notations basiques sur les locuteurs, les tours de parole et les elements non verbaux.
>
> Niveau 3: ce niveau contient toutes les informations du niveau 2 auxquels s'ajoutent des précisions interactionnelles et intonatives, comme les intonations prosodiques et les accents toniques (entre autres). EAGLES note que ce niveau ne peut être réalisé que par des phonéticiens et que la qualité des enregistrements doit pouvoir le permettre.
>
> Niveau 4 : c'est le niveau de transcription le plus détaillé. Il inclut toutes les informations présentes dans le niveau 3 auxquelles s'ajoutent des annotations intonatives, acoustiques et phénétiques supplémentaires. Les tons et les syllabes accentuées sont transcrits phonémiquement et alignés à une représentation numérique de la vague sonore ainsi que la fréquence fondamentale et le spectrogramme de l'énoncé.

À ces niveaux-là, la difficulté de rendre des transcripts dont l'intérêt reste dans le cadre de la linguistique augmente : Sinclair parle « d'abus », et Leech[219] de « pollution » ou « corruption » des textes. Blanche-Benveniste (1997 : 90) dit à ce propos :

[219] Tous deux cités par Blanche-Benveniste (1997, n. 10 : 90), qui rapporte ce que Sinclair disait lors du colloque de Lisbonne (3/10/95), et ce que Leech disait lors du colloque EAGLES de janvier 1996.

> Les néophytes, lorsqu'ils se lancent dans l'exploitation des corpus de langue parlée, ont tendance à insister sur les aspects les plus bizarres ou les plus anecdotiques. Le résultat est souvent un intérêt assez disproportionné accordé aux éléments comme les onomatopées, les claquements de langue, les raclements de gorge, les rires, toutes choses qu'on aura du mal à intégrer dans une description linguistique. (Blanche-Benveniste, 1997 : 90)

Inévitablement, les divergences dans les transcriptions sont l'un des facteurs ayant limité la réexploitabilité et la diffusibilité des corpus ; les corpus constitués en France ont pour la plupart suivi le même parcours ; comme le notent Baude & Abouda (2006 : 3), « certains corpus ont été constitués dans le cadre d'une recherche précise et n'ont de pertinence que pour celle-ci. Les conditions de collecte ou le travail très spécifique d'annotation ne permet pas la diffusion de ces données ». La multiplicité des pratiques, la diversité des transcripts possibles pour le même segment sonore et l'éventuelle surcharge qui en résulte préoccupa les théoriciens dans les années 1990 ; Blanche-Benveniste (1997 : 91) rapporte que certains chercheurs dépensaient autant d'énergie à effacer les surcharges d'un corpus que les initiateurs en avaient dépensé lors des transcriptions. Pour ces raisons, les experts d'Eagles recommandèrent de proposer une version « nue » de la transcription. Sinclair[220], de son côté, discutait de la nécessité de s'entendre sur un « niveau zéro » de la transcription, qui serait une transcription orthographique dépouillée de tout autre indication.

Ce niveau zéro aurait permis la réexploitation d'un corpus constitué, aurait facilité son utilisation et aurait pu ouvrir la voie – en France – à l'unification des corpus oraux en un corpus qui aurait été susceptible de rivaliser avec les corpus oraux anglo-saxons. La volonté d'un protocole de transcription uni et valable pour tous s'est vue confrontée aux faits : près de cinquante ans après les premiers corpus oraux, et plus de vingt ans après les recommandations officielles d'Eagles, rien de tel n'a vu le jour. Gadet résume la situation :

> Pas plus qu'il n'y a d'analyse tous azimuts et tous objectifs, il ne saurait y avoir de transcriptions tous azimuts et tous objectifs, aussi minimale puisse-t-elle paraître ; et il faut renoncer aux rêves d'harmonisation. (Gadet, 2008 : 46)

Néanmoins, nous plaidons pour une transcription orthographique minimale, dénuée d'informations supplémentaires. Nous avons évidemment conscience de la nécessité d'annotations prosodiques ou syntaxiques selon les cas, et notre proposition ne stipule pas

[220] Lors du Colloque de Madrid (janvier 1996), cité par Blanche-Benveniste, note 11.

leur bannissement ; nous préconisons uniquement que si un ajout d'informations doit avoir lieu, celui-ci doit être revendiqué en tant que niveau d'annotation supplémentaire, indépendant du niveau d'annotation de la transcription. De fait, et c'est là le cœur de notre propos, une annotation prosodique ou syntaxique se doit alors d'être suppressible du texte initial.

La transcription que nous proposons n'est pas une volonté de standardisation des méthodologies de transcription ; nous rappelons d'ailleurs que Leech préconisait qu'aucun schéma d'annotation ne devait être présenté en tant que standard. Nous proposons ce type de transcription afin, qu'en cas de réutilisation du corpus, les transcripts puissent être remaniés selon les nouveaux objectifs. En ceci, nous nous rapprochons des points de vue défendus en France (Cappeau & Gadet, 2010, 2012; Mondada, 2000, 2008) selon lesquels le transcript, en tant que produit, n'est pas un produit fini, mais une proposition pouvant sans cesse être revue, critiquée ou modifiée. Nous rappelons que le manque à gagner d'une telle transcription est compensé par une transcription incitant à l'écoute. Nous allons développer ce point ci-dessous.

2.7.2.3 Pour une transcription incitant à l'écoute

Il serait inenvisageable, de nos jours, de proposer des transcripts non alignés au son. La revue de la littérature consacrée à la transcription, ainsi que les différents protocoles de transcription que nous avons eu le loisir de consulter ou de connaître lors de présentations ou de communications, nous laissent penser que cet outil n'est cependant pas correctement exploité. Nous constatons en effet une contradiction entre les objectifs d'un alignement et le protocole qui va avec. S'il s'agit de représenter avec exactitude l'ensemble des phénomènes visés, l'alignement au son (voire la disponibilité du son) n'a plus lieu d'être, puisque tout est représenté.

Or nous avons démontré que la volonté d'une transcription exhaustive est un fantasme ; si l'on ajoute à cela la possibilité d'un alignement des transcripts au son, nous plaidons alors pour une transcription minimaliste dont le seul rôle est de permettre à l'utilisateur du corpus de savoir, via un concordancier, quels passages écouter en y accédant directement. Mondada le résume de la sorte :

> La transcription n'est pas un objet autonome par rapport à l'enregistrement : loin de le remplacer, la transcription n'a de sens qu'autant qu'elle y renvoie comme instance de vérification, et qu'elle en permet la consultation outillée. (Mondada, 2008 : 79)

Plusieurs points restent néanmoins à éclaircir. En premier lieu, nous concédons que ce type de démarche convient davantage à des corpus à des fins syntaxiques, morphologiques et lexicales. Pourtant, Blanche-Benveniste a entrevu, dès 1997, les possibilités de l'alignement, même pour les corpus font les objectifs sont orientés vers le « speech language » :

> Un certain nombre des ambitions de fidélité aux phénomènes sonores vont tomber en raison des avancées technologiques, par exemple la possibilité d'avoir, sur un CD-rom, des transcriptions couplées avec les enregistrements. On pourra envisager, par exemple, de vérifier les phénomènes prosodiques d'un passage transcrit orthographiquement, ou les liaisons, sans dépendre d'un système de notation étranger à la recherche en cours. (Blanche-Benveniste, 1997 : 90)

Il ne semble pas que 16 ans plus tard, l'idée ait beaucoup avancé. En tout cas, les protocoles récents ne détaillent pas si les enrichissements sont présents dans les transcripts en tant qu'outils de recherche pour la réécoute, ou s'il s'agit de réminiscences du fantasme d'étudier la langue orale en se basant *uniquement* sur les transcripts, en les lisant.

Cette idée de lecture nous amène à notre second point de précision, celui de la lecture des corpus. Nombre d'études discutent de la lisibilité du corpus, y afférant par exemple la nécessité de ponctuation ou de découpage syntaxique ; on retrouve également assez souvent des échantillons de textes bruts, dans le but de prouver la pénibilité de lecture. Or nous ne pensons pas qu'un corpus oral soit destiné à être lu. Il est destiné à être écouté, la fonction des transcripts se limite à diriger l'utilisateur vers les énoncés qui pourraient l'intéresser. Cette démarche implique certes des méthodologies chronophages[221], nous la voyons néanmoins plus prudente que celle de conclusions sur la langue orale en se passant totalement d'une réécoute. La moindre des choses serait en tout cas de ne pas alourdir les transcripts pour des raisons de lisibilité.

Nous terminerons nos remarques par un dernier point ; dans les cas particuliers, mais nombreux, où les transcripteurs sont tentés, soit par les trucages orthographiques, soit par un système de codage spécifique pour signaler une mauvaise prononciation, une déformation

[221] Il n'est certes pas question d'écouter l'ensemble d'un corpus de plusieurs centaines de milliers de mots, voire de millions de mots. Tout comme il n'est pas question de *lire* l'ensemble du corpus. Voir note suivante.

morphologique ou tout autre métaplasme, nous précisons encore une fois que ces pratiques ne doivent pas être suspendues, mais qu'elles doivent relever, comme nous l'avons précisé ci-dessus, d'un système d'annotation indépendant et suppressible des conventions de transcriptions[222]. Les corpus qui résulteraient de telles pratiques n'en seront donc pas moins riches, car toute annotation reste possible. Il sera seulement plus facile pour d'autres équipes de remanier les transcripts dépouillés de conventions spécifiques et ponctuelles.

Nous avons cité dans ce paragraphe Blanche-Benveniste et Mondada, qui soulignèrent brièvement l'utilité de l'alignement. À notre connaissance, cette question n'a pas été soulevée et la problématique de la transcription est souvent discutée indépendamment de l'alignement. Nous avons cependant lu chez Cappeau & Gadet, un questionnement qui abonde dans notre sens :

> Reste enfin à s'interroger sur la pertinence d'une tentative de représentation par écrit de traits oraux à une époque où l'on dispose de l'alignement texte/son, qui tend désormais à se généraliser. (Cappeau & Gadet, 2010)

Nous résumerons comme suit : l'alignement des transcripts au son n'est pas une banale option dans un corpus oral. Elle permet au contraire de résoudre en grande partie les doutes quant à la transcription, qui devient elle-même un outil permettant d'accéder à l'essentiel contenu dans la bande sonore.

2.7.3 Annotation morphosyntaxique

L'annotation morphosyntaxique, appelée en anglais « part-of-speech annotation », « grammatical tagging » ou « morphosyntaxic annotation », consiste à assigner à chaque item du corpus un code indiquant sa fonction morphosyntaxique. Durant le processus d'annotation, les linguistes se font assister par des outils d'annotation automatique, parfois nommés

[222] Pour les études qui nécessite une masse importante de données, comme les études lexicales ou de calcul de fréquence, la notation des phénomènes particuliers de l'oral n'est pas pertinente, et de tels corpus ne peuvent être appréhendés que par des concordanciers.

étiqueteurs, qui permettent l'automatisation partielle du processus : le degré de finesse de l'annotation varie selon l'étiqueteur utilisé[223].

L'annotation morphosyntaxique est l'une des moins récentes et des plus répandues parmi celles appliquées sur les corpus[224]. La principale raison de sa popularité est la performance des logiciels actuels dans l'annotation, qui offrent la possibilité d'une annotation quasi-automatique. Les logiciels sont actuellement en mesure de reconnaître l'item et de lui assigner la fonction morphosyntaxique qui lui correspond. Certains programmes d'annotations sont capables de prendre en compte le contexte syntaxique de l'item afin de déterminer sa fonction morphosyntaxique. Il reste que dans ce cas ou dans le premier, l'intervention humaine ou du moins une vérification humaine, reste nécessaire. Celle-ci peut se faire avant ou après l'intervention de la machine.

La programmation des outils d'annotation de corpus oraux relève de la responsabilité des informaticiens, mais il y a au moins trois étapes où le linguiste intervient dans leur élaboration, que nous allons examiner :

1) à l'oral où les segments syntaxiques sont encore plus floues qu'à l'écrit, il faut définir ce qu'est un item : suite de caractères précédés et suivis par un espace ; unités lexicales ; unités sémantiques ; gestion des locutions, des noms composés, des néologismes etc. La segmentation du flux de la parole implique des choix linguistiques et techniques.
La plupart des projets d'annotations syntaxiques du français portent uniquement sur l'écrit et posent comme pré-requis de l'analyse la segmentation en "phrases" (Gendner & Adda-Decker, 2002) qui n'est autre qu'une segmentation en phrases graphiques délimitées en amont par une majuscule et en aval par une ponctuation forte (Abeillé, Clément, Kinyon, & Toussenel, 2001 : 170) avec quelques exceptions telles que les incises. En ce qui concerne l'oral, les problématiques de la segmentation du flux sonore concernent également la transcription des données ;

[223] Les experts européens d'EAGLES ont fourni des recommandations précises en matière d'annotation morphosyntaxique, sur lesquelles nous ne reviendrons pas, car largement disponibles et diffusées en ligne (EAGLES, 1996b), et d'ailleurs traduites et présentées par Delais-Roussarie (2008 : 162).

[224] Le premier corpus électronique, le Brown, était morphosyntaxiquement annoté.

2) établir une taxinomie des fonctions morphosyntaxiques de la langue du corpus à traiter, et assigner à chaque fonction un code (POS) ;

3) établir un lexique de la langue cible et assigner à chaque forme sa fonction morphosyntaxique, en prenant en compte la question des homophones et des formes qui n'apparaissent pas dans le dictionnaire.

Ces démarches ne sont pas réitérées pour chaque corpus ; des lexiques déjà constitués peuvent être utilisés, comme par exemple le DicoLPL qui contient 440 000 entrées (cf. VanRullen *et al.*, 2005). Néanmoins, ce lexique doit être adapté aux outils informatiques utilisés par le chercheur ayant décidé de le réutiliser.

Comme exemple d'annotation morphosyntaxique, nous citerons encore une fois le CFPP2000. Les données ont été annotées avec TreeTagger[225], en utilisant une base d'annotation adaptée pour le français oral et développée par Christophe Benzitoun. La liste des étiquettes utilisées (POS) pour l'annotation est la suivante :

Ex 8

MLT	multi-transcription (/x,y/, (n'))
NAM	nom propre
NOM	nom commun
NOM|NAM:sig sigle	
NUM	numéral
PRO:clo	clitique objet
PRO:cls	clitique sujet
PRO:clsi	clitique sujet impersonnel
PRO:dem	pronom démonstratif
PRO:ind	pronom indéfini
PRO:int	pronom interrogatif (comment, où, quand, quoi, etc.)
PRO:pos	pronom possessif (mien, tien…)
PRO:rel	pronom relatif
PRO:ton	pronom tonique
PRP	préposition
PRP:det	préposition+déterminant (au, du, aux, des)
PRP :int	particule interrogative (est-ce que, est-ce qui)
SYM	symbole
TRC	troncation quand on n'arrive pas à déterminer la base du mot tronqué
VER:cond	verbe au conditionnel
VER:futu	verbe au futur
VER:impe	verbe à l'impératif
VER:impf	verbe à l'imparfait
VER:infi	verbe à l'infinitif
VER:pper	verbe au participe passé

[225] Disponible ici :

http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

VER:ppre	verbe au participe présent
VER:pres	verbe au présent
VER:simp	verbe au passe simple
VER:subi	verbe au subjonctif imparfait
VER:subp	verbe au subjonctif présent
étiquette:trc	mots tronqués (à la place de l'étiquette, mettre la partie du discours correspondante)
PUN	ponctuation
PUN:cit	guillemets
SENT	fin de phrase

L'annotation sur un tour de parole donne ceci[226]:

Ex 9

> **spk3** [459.438] : j'sais pas pour vous pour vous dire + moi je sais pas on a des abonnements euh + opéra théâtre euh concerts euh euh pour l'année qui vient
>
> <speaker="05-01_christophe"> j'#je#PPER1S sais#savoir#VINDP1S pas#pas#ADV pour#pour#PREP vous#vous#PPER2P pour#pour#PREP vous#vous#PPER2P dire#dire#VINF +#+#COO moi#moi#PPER1S je#je#PPER1S sais#savoir#VINDP1S pas#pas#ADV on#on#PPER3S a#avoir#VINDP3S des#un#DETDPIG abonnements#abonnement#NCMP euh#euh#INT +#+#COO opéra#opéra#NCMS théâtre#théâtre#NCMS euh#euh#INT concerts#concert#NCMP euh#euh#INT euh#euh#INT pour#pour#PREP l'#le#DETDFS année#année#NCFS qui#qui#PRI vient#venir#VINDP3S

Une fois un corpus annoté morphosyntaxiquement, il peut être lemmatisé. La lemmatisation consiste à assigner à chaque item du corpus sa forme canonique. Soit, pour une langue en partie flexionnelle comme le français, annoter l'ensemble des formes d'un paradigme donné avec le lemme du paradigme ; il s'agira de l'infinitif pour les verbes, du masculin singulier pour les noms, en ce qui concerne les deux grandes catégories de flexion. Les difficultés concernent par exemple l'attribution ou non du statut de locution à certains syntagmes, de la lexicalisation des locutions et des choix de traitement en ce qui concerne les mots-composés.

La lemmatisation est une annotation principalement utile en lexicographie et en sémantique. Néanmoins, les études sur la réelle utilité et la fiabilité de la lemmatisation des corpus sont

[226] Exemple tiré du corpus CFPP2000 :

CFPP2000 [05-01] Christophe_Andre_H_62_Marie_Anne_Andre_F_63_5e

Disponible ici :

http://cfpp2000.univ-paris3.fr/Corpus.html

encore imprécises, voire contradictoires (voir Lemaire, 2008 : 7) ; par exemple, Lemaire rapporte une étude selon laquelle « aucune différence entre un corpus basque non lemmatisé et sa version lemmatisée, tous deux traités par LSA[227], pour mesurer la cohérence et de la compréhension des textes chez des apprenants. » La lemmatisation des corpus sert cependant les desseins d'institutions ayant besoin d'extraire les informations jugées pertinentes d'un corpus oral, comme par exemple l'analyse des lemmes de conversations téléphoniques entre agent et client dans des standards tels ceux d'EDF, ou pour des conversations relatives à des incidents survenus en mer au standard du Centre de Recherche de la défense canadienne (voir Cailliau & Poudat, 2008 : 270-271 pour les deux études citées).

Nous donnerons comme exemple de corpus oral lemmatisé les travaux présentés par Gendner & Adda-Decker (2002 : 16), qui procédèrent à la lemmatisation d'un corpus écrit et oral de journaux écrits et d'émissions journalistiques audiovisuelles, diffusées entre 1997 et 2000. Voici ce que donne une lemmatisation :

Forme fléchie	**Lemme**	**POS**
notre	notre	D.s
sentiment	sentiment	Nms
c'	ce	P..
est	être	Vmip-s
que	que	C—
si	si	C—
les	le	D.p
syndicats	syndicat	Nmp
signent	signer	Vmip-p
ça	ça	P.s

Quand un corpus est morphosyntaxiquement annoté, il est également possible d'effectuer une annotation d'un niveau plus complexe nommée « parsing ». Un corpus « parsé » est un corpus arboré, du terme anglais « treebank ». Le procédé consiste à établir la liste des liens syntaxiques entretenus entre les éléments du corpus. Habert *et al.* (1997 : 43) le résument ainsi : « Il s'agit en effet de délimiter des groupes, de les nommer (les catégoriser), et de statuer sur leurs relations ».

[227] LSA : Latent Semantic Analysis. Il s'agit d'une technique de traitement des langues naturelles qui extrait et analyse les concepts qui apparaissent dans un ensemble de documents pour afin de spécifier ces documents ; la LSA se base sur le postulat que les mots proches en signification apparaissent des documents similaires.

Or, Habert *et al.* poursuivent sur l'obstacle commun de ces trois niveaux : la multiplicité des points de vue quant à la catégorisation des constituants. S'il est déjà difficile, mais relativement possible de trouver des accords sur l'étiquetage des occurrences, la segmentation en ensembles d'énoncés est loin de faire l'unanimité. Ceci est d'autant plus flagrant à l'oral : alors que l'écrit possède la phrase graphique, les unités de traitement de l'oral peuvent être multiples : - tour de parole ? - silence ? - pause longue ? - pause brève ? - unité intonative ?

En ce qui concerne le français, seuls deux corpus arborés sont disponibles : le corpus MCVF[228] (Modéliser le changement : les voies du français) de l'université d'Ottawa, constitué de textes d'ancien français, et le French TreeBank[229], de l'université Paris 7. Il n'y a pas, à notre connaissance, de corpus oraux arborés pour le français, nous donnerons donc un exemple tiré du French Treebank de Paris 7, dont voici la liste des étiquettes[230] :

Ex 10

- AP (syntagme adjectival)
- AdP (syntagme adverbial)
- COORD (syntagme coordonné)
- NP (syntagme nominal)
- PP (syntagme prépositionnel)
- VN (noyau verbal)
- VPinf (proposition infinitive)
- VPpart (proposition participiale)
- SENT (phrase indépendante)
- Sint, Srel, Ssub (propositions : conjuguée interne, relative, subordonnée)

L'exemple présenté est suivi de l'annotation morphosyntaxique, de sa lemmatisation ainsi que de son parsing[231] :

Ex 11

Il est entendu que les fonctions publiques restent ouvertes à tous les citoyens.

```
<SENT>
 <VN> <w lemma="il" ei="CL3ms" ee="CL-suj-3ms" cat="CL" subcat="suj" mph="3ms">Il</w>
 <w lemma="être" ei="VP3s" ee="V–P3s" cat="V" subcat="" mph="P3s">est</w>
 <w lemma="entendre" ei="VKms" ee="V–Kms" cat="V" subcat="" mph="Kms">entendu</w>
 </VN>
 <Ssub>
 <w lemma="que" ei="CS" ee="C-S" cat="C" subcat="S">que</w>
```

[228] http://www.voies.uottawa.ca/voies_fr.html

[229] http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php

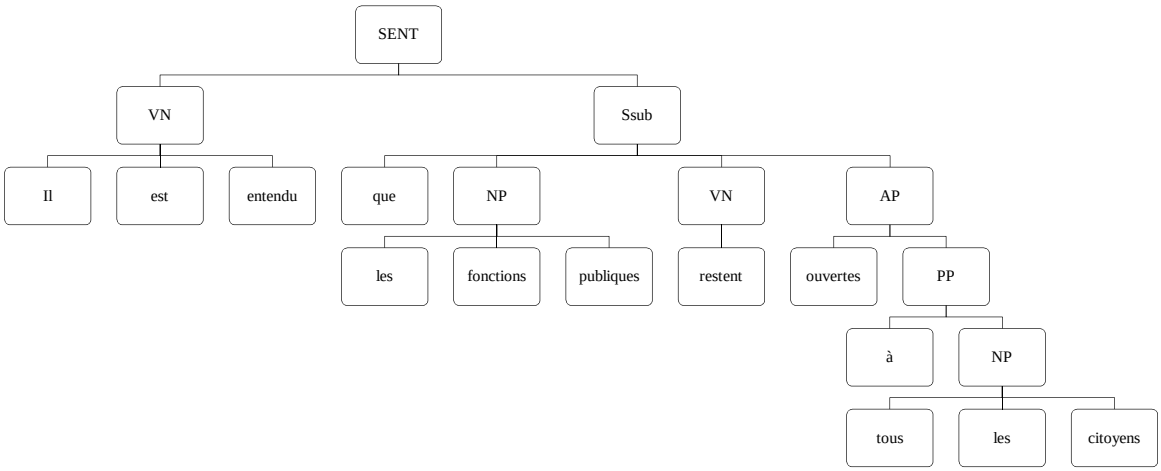
[230] Les étiquettes sont présentées en anglais, la traduction est celle de Delais-Roussarie (2008 : 169).

[231] Emprunté à Abeillé *et al.* (2001 : 42).

```
<NP> <w lemma="le" ei="Dfp" ee="D-def-fp" cat="D" subcat="def" mph="fp">les</w>
<w lemma="fonction publique" ei="NCfp" ee="N-C-fp" cat="N" subcat="C" mph="fp">
<w catint="N">fonctions</w> <w catint="A">publiques</w> </w>
</NP>
<VN> <w lemma="rester" ei="VP3p" ee="V–P3p" cat="V" subcat="" mph="P3p">restent</w> </VN>
<AP> <w lemma="ouvert" ei="Afp" ee="A-qual-fp" cat="A" subcat="qual" mph="fp">ouvertes</w>
<PP> <w lemma="à" ei="P" ee="P" cat="P">à</w>
<NP> <w lemma="tout" ei="Amp" ee="A-ind-mp" cat="A" subcat="ind" mph="mp">tous</w>
                <w lemma="le" ei="Dmp" ee="D-def-mp" cat="D" subcat="def" mph="mp">les</w>
                <w lemma="citoyen" ei="NCmp" ee="N-C-mp" cat="N" subcat="C"
        mph="mp">citoyens</w>
</NP></PP></AP>
</Ssub>
<w lemma="." ei="PONCTS" ee="PONCT-S" cat="PONCT" subcat="S">.</w>
</SENT>
```

et la forme arborée sera représentée ainsi :

Ex 12



2.7.4 Annotation évaluative

L'interlangue est terme que l'on doit à Selinker (1972), qui désigne la structure et la nature spécifiques du système d'une langue en cours d'acquisition par un apprenant à un stade donné. L'interlangue des apprenants d'une langue étrangère est un système intermédiaire qui possède ses propres spécificités, et Granger, qui aborde le sujet d'un point de vue des corpus d'apprenants, en discute ainsi :

> La langue de l'apprenant diffère de la langue maternelle tant quantitativement que qualitativement. Elle se manifeste par des fréquences de mots, d'expressions et de structures très différentes, certains éléments étant surutilisés et d'autres considérablement sous-utilisés (…). Elle est également caractérisée par un taux élevé

> d'usages impropres, à savoir des erreurs orthographiques, lexicales et grammaticales. (Granger, 2007 : 1)

Nous parlerons dans cette section d'un type d'annotation visant à cerner les spécificités de l'interlangue, il s'agit de l'annotation évaluative. Nous sommes directement concernés par ce type d'annotation, pour avoir constitué un corpus oral d'apprenants ; nous n'avons cependant pas eu le loisir de procéder à une annotation évaluative exhaustive. Les erreurs des apprenants ayant retenu notre attention tout au long de notre parcours (recueil des données, transcription, rédaction de nos travaux de thèse), nous discuterons toutefois de l'annotation évaluative : dans un corpus d'apprenant, les erreurs produites sont indiquées et classées selon une typologie des erreurs que les annotateurs auront auparavant dressée. Comme pour l'annotation morphosyntaxique, l'objectif des chercheurs est de pouvoir procéder à ce type d'annotation via l'aide d'un logiciel, et non pas manuellement. Là encore, les logiciels de ce type sont davantage destinés à l'annotation des corpus écrits et non pas des corpus oraux. Ainsi, pour les corpus d'apprenants en FLE, le seul projet d'annotation évaluative d'envergure dont nous avons eu connaissance concerne un corpus de productions écrites d'apprenants, qui est le corpus FRIDA[232] et il n'existe pas, à l'heure actuelle, de corpus oral de FLE dont les erreurs sont annotées. FRIDA contient 450 000 mots dont les deux tiers ont été annotés selon les étapes suivantes :

> 1) correction manuelle du corpus de FLE ;
> 2) élaboration d'un système d'étiquetage d'erreurs pour le FLE ;
> 3) insertion d'étiquettes d'erreurs et de corrections dans les fichiers textes ;
> 4) extraction de listes de types d'erreurs spécifiques et de statistiques d'erreurs ;
> 5) analyse linguistique basée sur la concordance des principaux types d'erreurs. (Granger, 2007 : 2)

S'inspirant de travaux antérieurs, le schéma d'annotation de FRIDA se veut une synthèse entre une taxinomie des erreurs basée sur les catégories linguistiques (morphologie, lexique et erreurs grammaticales), et une autre axée « sur la manière dont les structures de surface ont été altérées » (soit la syntaxe de l'interlangue : omissions, additions, ordre erroné).

Les auteurs présentent leur schéma comme « informatif mais gérable », soit suffisamment détaillé mais pas au point de devenir ingérable pour l'annotateur, « réutilisable » quel que soit

[232] FRIDA : **Fr**ench **I**nterlanguage **Da**tabase : http://sites.uclouvain.be/cecl/projects/Frida/frida.htm

la langue, « souple » soit modifiable et enfin, « cohérent ». La classification des erreurs se présente ainsi (Granger, 2007 : 3) :

Tableau 7 : Catégorisation des erreurs du corpus écrit FRIDA

Domaine d'erreurs		Catégorie d'erreurs	
<F>	Forme	<AGL> <MAJ> <DIA> <HOM> <GRA>	Agglutination Majuscule/minuscule Signe diacritique Homonymie Autres erreurs d'orthographe
<M>	Morphologie	<MDP> <MDS> <MFL> <MFC> <MCO>	Dérivation-préfixation Dérivation-suffixation Inflexion Inflexion-confusion Composition
<G>	Grammaire	<CLA> <AUX> <GEN> <MOD> <NBR> <PER> <TPS> <VOI> <EUF>	Classe Auxiliaire Genre Mode Nombre Personne Temps Voix Euphonie
<L>	Lexique	<SIG> <CPA> <CPD> <CPV> <CPN> <FIG>	Signification Complémentation adjectivale Complémentation adverbiale Complémentation verbale Complémentation nominale Figement
<X>	Syntaxe	<ORD> <MAN> <RED> <COH>	Ordre des mots Mot manquant Mot redondant Cohésion
<R>	Registre	<RLE> <RSY>	Lexique Syntaxe
<Y>	Style	<CLR> <LOU>	Obscur Lourd
<Q>	Ponctuation	<CON> <TRO> <OUB>	Ponctuation erronée Ponctuation redondante Ponctuation manquante
<Z>	Coquille		

Exemples authentiques d'erreurs annotées au sein du corpus FRIDA (Granger, 2007 : annexe A) :

<F> **<AGL>** le **portebagages** (porte-bagages) est sur le toit
<MAJ> Mme Thatcher incarnait la peur des **anglais** (Anglais)

<DIA> Il existe une **ambiguité** (ambiguïté)
<HOM> Ils **ce** (se) déshabillent
<GRA> un **labirint** (labyrinthe)

<M> **<MDP>** le sentiment de **malcontentement** (mécontentement)
<MDS> ...qui continue à **évolutionner** (évoluer)
<MFL> Elle s'occupe des enfants et des **travails** (travaux)
<MFC> Il n'a pas très bien **réussit** (réussi)
<MCO> la participation de **celles-dernières** (celles-ci)

<G> **<CLA>** L'unique chose **que** (qui) n'est pas bonne en Belgique...
<AUX> Je m'**avais** (étais) très bien amusé
<GEN> La protection sociale a été **amélioré** (améliorée)
<MOD> Bien qu'ils **sont** (soient) pressés, ...
<NBR> Elle reprit ses **esprit** (esprits)
<PER> J'espère **s'** (m') adapter rapidement
<TPS> Un sondage qui **était** (a été) publié dans le Monde montre que...
<VOI> Les éclipses **ont vues** (sont vues) comme des présages
<EUF> En prenant **ce** (cet) aspect, ...

<X> **<ORD>** Je peux **m'amuser bien** (bien m'amuser).
<MAN> Je crois **0** (qu') ici il y a beaucoup plus de soirées
<RED> le domaine social **et** (,) économique et politique
<COH> A la métropole, il existe plus d'allocations et d'aide pour les chômeurs, les handicapés et les personnes âgées, **et** (tandis que) dans les îles, il n'y a pas beaucoup de soutien.

<R> **<RLE>** **Quand même** (néanmoins), c'est une histoire lointaine
<RSY> Étant donné que **j'ai pas** (je n'ai pas) souvent présenté mes travaux...

<Y> **<CLR>** Quand la famille d'un travailleur étranger vient le rejoindre, **on** (??) est obligé d'organiser toutes les formalités avant de quitter son pays.
<LOU> mais il y a des autres choses qui **m'ont donné une très grande surprise** (qui m'ont fort surpris)

<Q> **<CON>** La langue devient plus française **-** (:) on l'appelle maintenant le créole francisé.
<TRO> Ce sont surtout les différences biologiques**,** (0) qui sont présentées dans la société avec la grande force de l'organisme humaine.
<OUB> Quand j'ai du temps libre **0** (,) je veux faire des choses reposantes...

<Z> **qunad** (quand); **ps** (pas)

Ce schéma, destiné à l'annotation de l'écrit, ne conviendrait pas à l'oral sans quelques modifications, ne serait-ce qu'en ce qui concerne les erreurs du type F (forme), mais d'aucuns pourraient suggérer, modifier ou supprimer certaines catégories. Comme pour les conventions de transcription ou n'importe quel autre schéma d'annotation, le schéma universel et standard

n'existe pas. Il nous semble que la probabilité d'un tel schéma est la plus faible dans le domaine de l'annotation évaluative, tant celle-ci est interprétative, subjective et reflétant la vision de la langue de l'annotateur.

2.8 Critère de la documentation

Nous en venons maintenant à un autre critère des corpus, fortement lié à celui de la représentativité, qui est celui de la documentation. La documentation d'un corpus peut être assimilée à la description des données. Rastier (2005) parle de la documentation qui est pour lui une conception qui « ne retient que des variables globales caractérisant les documents, sans tenir compte de leur caractère textuel, ni de leur structure. » Cette documentation diffère selon le type de corpus ; auteur, époque ou courant pour un corpus littéraire ; dates et sources journalistiques pour un corpus de presse etc. Nous intéressant aux corpus oraux, nous discuterons leur documentation en détail ci-dessous. Par ailleurs, la documentation est inhérente au processus de collecte s'il est suivi selon des protocoles établis et permet, selon Péri-Woodley (1995) la réutilisabilité du corpus, tout en augmentant son potentiel de représentativité :

> La notion de ressources réutilisables implique des corpus à géométrie variable, capables d'être adaptés à des approches et à des objectifs divers. A partir des ressources disponibles, les chercheurs devront pouvoir construire – sélectionner, réorganiser – des corpus en fonction de besoins spécifiques. Le potentiel de représentativité d'un corpus est alors étroitement lié à sa diversité et à la qualité de sa documentation. (Péri-Woodley, 1995)

Comme exemple de documentation d'un corpus oral, nous citerons le corpus ASCYNT[233], dont la documentation comprend les points suivants :

- présentation générale ;
- contenu des CDs ;
- enregistrement et numérisation des données ;
- transcriptions des données : modalités et conventions ;

[233] Le corpus ACSYNT est un corpus oral qui comprend 120 000 mots environ et qui a été élaboré par Élisabeth Delais-Roussarie, en collaboration avec J.M Tarrier, D. Bourigault, I. Choi-Jonin, C. Fabre, L. Molinu et M. Rouquier. La présentation du corpus, ainsi que sa documentation sont disponibles ici : http://www.llf.cnrs.fr/Gens/Delais-Roussarie/ACSYNT_documentation.pdf

- documentation linguistique diverse ;
- identité des locuteurs ;
    - nom, sexe, âge
    - classe socioprofessionnelle du locuteur et de ses parents
    - profil géographique du locuteur qui se déduira des lieux et régions où il a vécu, ainsi que les origines géographiques de ses parents
    - langues maîtrisées
- enquêteur.

Comme nous le constatons, une documentation contient également des informations tels les contenus des CDs, la numérisation des données et les protocoles suivis. Ces informations sont nécessaires, comme présenté par Péry-Woodley ci-dessus, à une éventuelle réutilisation partielle des données (à des fins d'analyse ou de constitution) ou de l'ensemble du corpus. Nous allons discuter, un à un, des éléments principaux d'une documentation à même de permettre une analyse pertinente, ainsi qu'une réexploitation et une représentativité des données ; éléments que nous considérerons du point de vue des corpus oraux.

Nous ne discuterons toutefois pas dans cette section de points tels les transcriptions des données, l'identité des enquêteurs ou les spécificités des enregistrements. Selon nous, ces questions ne se rattachent pas à la documentation du corpus mais sont liées aux annotations desquelles elles dépendent : par exemple, les conventions de transcription doivent figurer dans le schéma de transcription en tant que schéma d'annotation, et l'identité des enquêteurs est lié au protocole d'enregistrement. Seules les questions concernant la situation d'énonciation et l'identité des locuteurs seront traitées ici.

Avant de poursuivre, il nous semble nécessaire de citer et de commenter ici un avis de Cappeau & Gadet concernant la représentation sociolinguistique des locuteurs :

> Or la question apparaît mal posée. D'abord, c'est faire l'hypothèse qu'un locuteur puisse être dit représentatif de quelque chose. Mais représentatif de quel aspect de son identité ? Pourquoi davantage des « hommes », des « catholiques » ou des « avocats », que des « joueurs d'échecs » ? Aucun individu n'est catégorisable à partir de ses énonciations de façon rapportable à une seule dimension identitaire. Une seconde conséquence, qui découle d'ailleurs de cette première objection, est que l'on privilégie ainsi une conception du social reposant sur des catégories pré-construites, dont les locuteurs sont censés être porteurs. Une telle sociolinguistique aura-t-elle un pouvoir heuristique, si elle ne fait que confirmer ce que l'on a considéré comme pré-construit

> (Gadet 2000) ? Par exemple, qu'une femme est une femme, et qu'un jeune n'est pas un vieux : découvertes sociologiques fulgurantes, qui méritent bien un gros investissement méthodologique et analytique ! D'ailleurs, permettent-elles vraiment de telles conclusions, ou bien ne font-elles que boucler un raisonnement sur lui-même, un discours étant regardé comme représentatif de la façon dont il a été catégorisé (par exemple, un « jeune » parle comme on a cru pouvoir dire que les jeunes parlaient) ? (Cappeau & Gadet, 2007a : 104-105)

Ces remarques sont certes pertinentes, mais comme l'indique le titre de l'article des auteurs, « L'exploitation sociolinguistique des grands corpus. Maître-mot et pierre philosophale », ces problèmes sont d'ordre sociolinguistique et il n'est pas dans le cadre de notre étude, qui s'intéresse davantage aux tenants théoriques et pratiques de la constitution des corpus, que de résoudre ces problématiques. Les sociolinguistes qui voudront interroger un corpus sur base de critères sociolinguistiques seront plus à même d'effectuer leurs requêtes selon des critères de sélection qu'ils auront à justifier. Le rôle des compilateurs de corpus se limite à fournir aux chercheurs un panel d'outils de sélection, les choix d'utilisation dépendront de protocoles antérieurs à la constitution du corpus.

2.8.1 Situation d'énonciation

Considérons l'exemple d'une transcription du corpus Léonard[234], dont voici la documentation :

Ex 13

@Participants: CHI Léonard Target_Child, MOT Mother, FAT Father, PAT Family_Friend, OBS Aliyah Observer
@ID: fra|Paris-Corpus_Leonard|CHI|1;08.09|male|bi||Target_Child|||
@ID: fra|Paris-Corpus_Leonard|MOT|||||Mother|||
@ID: fra|Paris-Corpus_Leonard|FAT|||||Father|||
@ID: fra|Paris-Corpus_Leonard|PAT|||||Family_Friend|||
@ID: fra|Paris-Corpus_Leonard|OBS|||||Observer|||
@Birth of CHI: 15-OCT-1990
@Media: 01, video
@Date: 23-JUN-1992
@Time Start: 18:00
@Time Duration: 18:00-18:33
@Comment: Coder - Aliyah MORGENSTERN, Christophe PARISSE Date of transcript: Printemps 1993, Eté 1995, 2006
@Location: Léonard's home

[234] Fichier 01.cha de la plateforme CHILDES :
http://childes.psy.cmu.edu/browser/index.php?url=Romance/French/Paris/leonard/

@Situation: dans le bain. MOT est accroupie près de la baignoire.

Nous constatons que les informations données sont les suivantes :

- les personnes présentes lors de l'enregistrement, ainsi que le lien qu'entretient chacune d'entre elles avec l'enfant ;
- des informations sur la situation d'énonciation : lieu de l'enregistrement, activité de l'enfant et détails concernant la mère ;
- le nom du corpus ;
- le sexe de l'enfant, sa date de naissance, la date de l'enregistrement, et de la transcription. Le type d'enregistrement (vidéo) ;
- la durée de l'enregistrement ;

Les deux premiers points concernent la situation d'énonciation. La problématique majeure concernant les situations d'énonciation des locuteurs des corpus porte sur la représentativité de celles-ci, comme le résument Cappeau & Gadet :

> Les interviews ont longtemps constitué la situation la plus représentée (de façon écrasante), avec l'idée – soit naïve soit hâtive – qu'existeraient des positions de recueil neutres. L'accent désormais mis davantage sur les recueils dans des conditions écologiques, exploitables à des fins interactionnelles ou autres, a quelque peu changé la donne. (Cappeau & Gadet, 2010)

Nous ne reviendrons pas ici sur les situations d'énonciation du point de vue de la représentativité, notion que nous avons amplement discutée. Néanmoins, le recueil de donnée dans des conditions « écologiques » se heurte à un obstacle majeur, celui du juridique. Dans le « Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux », il est dit :

> On parle souvent de formulaires d'autorisation à soumettre aux informateurs; il est cependant important de faire dépendre cette autorisation de l'information préalable donnée aux personnes concernées: sans *information*, la *demande d'autorisation* n'a pas d'objet ni de sens. C'est pourquoi on parle de *consentement éclairé (informed consent)*, dans le sens où l'acceptation de l'enregistrement est étroitement dépendante de la compréhension des finalités pour lesquelles il est effectué. Sur certains terrains, la difficulté de faire comprendre les finalités de la recherche ne doit cependant pas inciter le chercheur à passer outre la demande de consentement, et celle-ci doit alors être formulée en accord avec le type de société dans laquelle se déroule le terrain. (Baude *et al.*, 2006 : 44)

Il apparaît donc que si « l'écrasante » majorité des situations sont des interviews, la raison est à imputer davantage à la difficulté de recueil de donnée écologiques qu'à la conviction que l'interview représente une situation de recueil neutre. Cette difficulté est d'autant plus accrue lorsqu'il s'agit de corpus spécialisés de langage en acquisition, d'apprenants ou de parole pathologique.

Ainsi comme nous l'avons évoqué plus haut, l'annotation métalinguistique n'est pas concernée par la pertinence des annotations. Son rôle est de permettre à l'utilisateur d'un corpus de comprendre la situation d'énonciation, et de formuler le jugement adéquat. Pour ceci, il est nécessaire que cet utilisateur ait à disposition les informations suivantes sur la situation d'énonciation :

1) informations sur le locuteur ;
2) les personnes en présence ;
3) le lieu d'énonciation ;
4) informations temporelles ;
5) l'activité ou les activités se déroulant lors de l'énonciation.

2.8.2 Sexe des locuteurs

Les études portant sur les variations linguistiques sexolectales ne sont pas une nouveauté et sont nombreuses[235], et parfois contestées en raison des critiques visant les démarches méthodologiques et les conceptions stéréotypées de ces études. Nous ne traiterons pas du problème ici, mais jugeons que la constitution d'un corpus se doit d'inclure les informations concernant le sexe des locuteurs, à défaut de respecter une parité locuteurs/locutrices au stade de la collecte des données.

Si les données sont écrites, il est plus difficile de prévoir un équilibre, notamment dans certains domaines où la parité hommes/femmes est faible (littérature classique, certains

[235] Pour davantage d'informations sur ce sujet, voir « La langue française au féminin: le sexe et le genre affectent-ils la variation linguistique? » (Armstrong, Bauvois, Beeching, Bruyninckx, & Gadet, 2001).

domaines scientifiques) ou d'autres domaines où les auteurs ne sont pas clairement identifiés (manuels d'utilisation, ouvrages coécrits[236]).

En ce qui concerne l'oral, Meyer (2002 : 49) préconise d'inclure des conversations entre deux femmes, entre une femme et un homme, entre une femme et deux hommes, entre deux femmes et deux hommes etc., invoquant des études selon lesquelles une femme parlera différemment selon la présence d'un ou de plusieurs autres personnes du même sexe ou de sexe différent. Meyer est cependant conscient que ceci n'est qu'une préconisation, ardue à réaliser :

> To summarize, there is no one way to deal with all of the variables affecting the gender balance of a corpus. The best the corpus compiler can do is be aware of the variables, confront them head on, and deal with them as much as is possible during the construction of a corpus. (2002 : 49)

D'autre part, l'équilibre recherchée dans un corpus peut biaiser les données, si cet équilibre n'existe pas dans la réalité, tel que le remarque Kennedy :

> In societies where the education of girls is less extensive than for boys, it may not be meaningful to compare linguistics characteristics of male and female participants. (Kennedy, 1998 : 66)

Là encore, une annotation métalinguistique à visées de documentation aura pour fil conducteur la transparence des informations, à défaut de pouvoir accéder à une représentation paritaire, qu'elle soit numérale ou non.

2.8.3 L'âge des locuteurs

Certains corpus comme ceux de la plateforme CHILDES contiennent des données de locuteurs enfants, adultes et adolescents, ceci est valable également pour la partie langue parlée du BNC, que ce soit des productions écrites ou orales[237]. Mais la plupart des corpus contiennent exclusivement des données de locuteurs adultes, pour trois raisons principales :

[236] Kennedy (1998 : 66) et Meyer (2002 : 49) ajoutent qu'à l'écrit – notamment pour les articles scientifiques ou de presse –, il est très difficile de pouvoir retracer les modifications opérées par l'éditeur.

[237] Voir à ce sujet (inclure la référence de Lou Burnard sur le BNC).

1) il est juridiquement difficile d'être dans le cadre légal quant aux données produites par des enfants, surtout en ce qui concerne les données filmées ;
2) en ce qui concerne les données écrites, la collecte de données d'auteurs adultes est plus aisée, simplement en raison du fait que les enfants ne produisent pas la plupart des types de données écrites (romans, articles, rapports, manuels etc.). Quant aux données orales enregistrées à dessein, un adulte comprend qu'il est en situation de collecte et facilite la tâche de l'enquêteur. La collecte des données d'enfants implique un surcoût matériel et humain ;
3) les données obtenues ne sont exploitables que dans certains domaines comme l'acquisition du langage, les pathologies du langage, l'acquisition en contexte bilingue etc.

Ceci dit, un corpus de locuteurs adultes se verra également dans la possibilité de spécifier l'âge des locuteurs, en raison des possibles variations diastratiques dues à la maturité ou les études du locuteur. À titre d'exemple pour les données orales, l'ICE considère qu'un locuteur est adulte à partir de 18 ans et définit pour l'âge adulte quatre catégories : 18-25 ans, 26-45 ans, 46-65 ans, 66 ans et +.

2.8.4 Niveau socioculturel ou socioprofessionnel

Le niveau socioculturel ou socioprofessionnel du locuteur fut un critère pour certains corpus, comme par exemple celui du « International Corpus of English », qui n'inclut que les locuteurs « instruits », et définit en tant que locuteur instruit une personne ayant atteint le collège au cours de sa scolarité. Meyer critique cette mesure de la sorte :

> Arguably, such a restriction is arbitrary and excludes from the ICE Corpus a significant range of speakers whose language is a part of what we consider Modern English. Moreover, restricting a corpus to educated speech and writing is elitist and seems to imply that only educated speakers are capable of producing legitimate instances of Modern English. (Meyer, 2002 : 50)

Si la question reste ouverte, il semble qu'il y ait néanmoins unanimité à du moins inclure les informations personnelles pertinentes du point de vue socioculturel ou socioprofessionnel dans la documentation d'un corpus, afin de permettre aux chercheurs d'y opérer des analyses sociolinguistiques.

2.8.5 Conclusion sur la documentation du corpus

Le rôle de la documentation d'un corpus n'est pas de faire de lui un corpus représentatif, mais un corpus dont la représentativité peut être évaluée. La documentation se devra d'inclure toutes les informations métalinguistiques du locuteur et de la situation d'énonciation, et il sera du rôle de l'utilisateur d'interpréter ces métadonnées avec discernement.

Comme pour toute annotation, la documentation doit accompagner le corpus sans y être intégrée, ou du moins en être suppressible. Les métadonnées d'un corpus ne sont donc pas à présenter aléatoirement, plusieurs possibilités s'offrent aux concepteurs de corpus. Ceci ne concerne pas uniquement la documentation, mais également les méthodologies et langages employés pour toute forme d'annotation des données. C'est ce à quoi va s'intéresser la section suivante.

2.9 Les langages informatiques et l'annotation

Nous présenterons ici, succinctement, les méthodologies et langages informatiques utilisés pour l'annotation des corpus. Nous commencerons par un bref historique de ces langages afin de mieux définir le XML, et nous terminerons par la présentation de deux corpus : le British National Corpus pour ses liens avec le XML et l'American National Corpus pour le volume de données orales disponibles en XML. Le BNC et l'ANC, ainsi que la plupart des corpus d'importance disponibles et diffusés en XML sont des corpus anglo-saxons de constitution et de contenu ; nous conclurons donc cette section par un rapide examen des alternatives françaises à de tels corpus, principalement concrétisées par l'initiative TGE-Adonis.

Dès le début de l'informatique et la manipulation de grandes quantités de données, toutes disciplines confondues, la nécessité de langages informatiques pour la description de ces données s'est avérée nécessaire pour plusieurs raisons : d'une part, le langage de description doit être un langage informatique pour qu'il puisse être interprété par les ordinateurs ; d'autre part, le recours à un langage simplifié via des codes permit la synthèse des informations au lieu du recours à des descriptions orthographiques qui ne pouvaient qu'être disparates ; enfin, les langages de description informatique se prévalent d'un caractère universel qui passe outre les spécificités de chaque langue humaine.

Un langage de définition de données (LDD) est un ainsi langage informatique de balisage, destiné à décrire les données d'une base de données, et les liens structurels entre lesdites données, quelles qu'elles soient.

2.9.1 XML et TEI

En ce qui concerne les données linguistiques, Charles Goldfarb, en 1969, était chef de projet chez IBM et fait lancer le premier LDD destiné à la description du langage, le « Generalized Markup Language » (GML), qui devient le « Standard Generalized Markup Language » (SGML) en 1986.

C'est ce format qui est utilisé pour la première édition du British National Corpus (BNC) en 1994. Cependant, et tel que le rapporte Burnard (2007 : 29), « le BNC précède l'ère du World Wide Web » et les concepteurs du corpus n'avaient pas pensé appliquer aux données un format compatible à la notion encore embryonnaire du WEB.

Les pages WEB utilisent le format de données « HyperText Markup Language » (HTML), qui n'est pas un LDD mais un format de données ; cela veut dire que le HTML ne se préoccupe pas de la description des données, mais de l'apparence de ces données quand elles sont présentées sur une page web. Il a fallu donc la création d'un LDD compatible avec les pages WEB. C'est ainsi que la naissance du « Extensible Markup Language » (XML) est justifiée en 1997:

> Son but est de permettre au SGML générique d'être transmis, reçu et traité sur le web de la même manière que l'est HTML aujourd'hui.

Le XML est donc un LDD applicable aux données linguistiques. Sa principale caractéristique est l'absence d'étiquettes prédéfinies. Cela signifie que le XML offre un panel d'étiquettes que l'utilisateur aura le loisir de définir, car il n'y a pas de standards XML. Par exemple, pour la catégorie « pronoms », les concepteurs du BNC ont choisi la liste des étiquettes suivantes : DPS, DTQ, EX0, PNI, PNP, PNQ, PNX, selon le type de pronom (2007 : 26). La nécessité de la création d'une liste d'étiquettes laisse la liberté aux annotateurs de choisir leurs annotations, tout en profitant d'un langage standardisé. Avant de détailler ce dernier point,

voici ce que donne une représentation XML d'un énoncé suivi de son annotation morphosyntaxique[238] :

Ex 14

*FLE: j' aime bien les rues.
%mor: pro:subj|je&1S
v|aimer-PRES-_1SV
adv|bien
det|les&PL
n|ru-_FEM-_PL
pro:rel|qui v:exist|être&PRES&3PV
adv:neg|pas adv|très
adj|large-_PL.

en XML :

- <CHAT xmlns:xsi="**http://www.w3.org/2001/XMLSchema-instance**" xmlns="**http://www.talkbank.org/ns/talkbank**" xsi:schemaLocation="**http://www.talkbank.org/ns/talkbank http://talkbank.org/software/talkbank.xsd**" Media="**AbdullahM**" Mediatypes="**audio**" Version="**2.0.2**" Lang="**fra**" Options="**bullets**" Corpus="**CIL**" Id="**test**" Date="**1984-01-01**">
- <Participants>
<participant id="**FLE**" role="**Student**" language="**fra**" age="**P29Y**" sex="**male**" group="**arabe**" />
<participant id="**ENQ**" role="**Investigator**" language="**fra**" />
</Participants>
- <u who="**FLE**" uID="**u0**">
- <w>
j'
- <mor type="**mor**">
- <mw>
- <pos>
<c>**pro**</c>
<s>**subj**</s>
</pos>
<stem>**je**</stem>
<mk type="**sfxf**">**1S**</mk>
</mw>
</mor>
</w>
- <w>
aime
- <mor type="**mor**">
- <mw>
- <pos>
<c>**v**</c>
</pos>
<stem>**aimer**</stem>
<mk type="**sfx**">**PRES**</mk>
<mk type="**sfx**">**_1SV**</mk>
</mw>
</mor>
</w>
- <w>
bien
- <mor type="**mor**">
- <mw>
- <pos>
<c>**adv**</c>

[238] Exemple tiré de notre propre corpus.

```
      </pos>
 <stem>bien</stem>
      </mw>
      </mor>
      </w>
- <w>
      les
- <mor type="mor">
- <mw>
- <pos>
 <c>det</c>
      </pos>
 <stem>les</stem>
 <mk type="sfxf">PL</mk>
      </mw>
      </mor>
      </w>
- <w>
      rues
- <mor type="mor">
- <mw>
- <pos>
 <c>n</c>
      </pos>
 <stem>ru</stem>
 <mk type="sfx">_FEM</mk>
 <mk type="sfx">_PL</mk>
      </mw>
      </mor>
      </w>
- <t type="p">
- <mor type="mor">
 <mt type="p" />
      </mor>
      </t>
 <media start="827.185" end="846.852" unit="s" />
      </u>
      </CHAT>
```

La lourdeur apparente du balisage n'en est pas une, la conversion en XML étant automatisée grâce à la plateforme chatter.jar, dont nous parlerons dans le chapitre suivant.

Nous allons quelque peu expliciter l'intérêt d'une telle représentation XML. Les transcripts de nos données sont au format .cha, qui est le format utilisé par le logiciel que nous avons utilisé, ne sont lisibles que pour le logiciel CLAN. Plusieurs problèmes se posent alors, en sachant que ceux-ci sont les mêmes quel que soit le logiciel :

1) seuls les utilisateurs de CLAN peuvent consulter ces transcripts. Le logiciel, comme l'ensemble des logiciels de transcription, est peu maniable et peu intuitif. Son utilisation nécessite l'étude des manuels ;
2) même pour une personne un tant soit peu familière avec le logiciel utilisé, il se peut qu'elle désire appliquer les possibilités d'analyses d'un autre logiciel, ce qui suppose la conversion des fichiers aux formats adéquats. L'interopérabilité entre logiciels n'est ni toujours possible, ni toujours optimale ;
3) les données ne sont pas pérennes : elles restent tributaires du logiciel et de ses mises à jour. Les évolutions technologiques nous ont appris que l'obsolescence des outils de lecture

survenait bien plus rapidement que ce que l'on pouvait prévoir. Il est par exemple aujourd'hui peu aisé, en ce qui concerne le matériel, de consulter des enregistrements magnétiques audio ou vidéo. Même en ce qui concerne les supports numériques, la disquette (floppy disk) encore très répandue il y a une dizaine d'années, a aujourd'hui quasiment disparu, et le DVD comme support de stockage semble destiné au même sort. Quant aux logiciels, nombre de programmes ne peuvent être lancés que sur les plateformes auxquelles ils ont été destinés ; s'il est toujours possible de les exploiter au moyen de bidouillages informatiques, rien n'assure que des transcripts tels les nôtres, au format .cha, soient lisibles d'ici quelques années sans avoir recours aux services de professionnels.

C'est pour ces raisons qu'une représentation largement conventionnée comme le XML intervient : son rôle est d'assurer l'utilisation des données indépendamment des disciplines, des plateformes, des logiciels et des systèmes d'exploitation utilisés pour la constitution des données d'une part, et d'assurer leur exploitation sur le long terme d'autre part.

Nous revenons maintenant sur ce que nous évoquions ci-dessus : le balisage XML offre la possibilité de baliser et de représenter les données selon des standards pérennes et reconnus, mais le choix des balises reste libre ; ceci aurait pu amener à des représentations disparates vu que tout chercheur aurait choisi des balises différentes. C'est ici que nous parlerons de la Text Encoding Initiative (TEI). La TEI rassemble depuis 1987 des experts et chercheurs, toutes nationalités et spécialisations confondues, pour proposer des conventions de balisage pour toute forme de document. Initialement, l'approche de la TEI était le SGML, mais a adopté en le XML à sa naissance. La TEI ne propose pas de schéma définitif, mais un ensemble de schémas pour la représentation des données linguistiques en langage XML, dont l'utilisateur en choisira un. Un document sera alors « TEI-conformant » s'il remplit au moins la condition suivante[239] : être représenté à un format XML valide selon l'un des schémas proposés par la TEI.

Les travaux détaillant le XML, la TEI et ses cadres sont nombreux (site de la TEI[240], les travaux de Lou Burnard (Burnard & Baumann, 2010) ou une présentation en français

[239] Pour les autres conditions, techniques et documentaires, voir sur le site de la TEI :
http://www.tei-c.org/release/doc/tei-p4-doc/html/CF.html

[240] Site de la TEI :
http://www.tei-c.org/index.xml

résumant le codage des annotations disponible au sein du travail de thèse de Bazillon[241] (2011, Chapitre 4 : 142-162)).

En ce qui concerne notre corpus, cette section justifie la version XML que nous proposons en parallèle à la version .cha de CLAN. Elle a été encodée en XML selon le schéma de la plateforme TALKBANK[242]. Nous passons maintenant à la présentation de corpus ou de bases de données ayant lien avec le XML. Ces corpus ou projets nécessitaient de toute façon une présentation, vu leur relative importance.

2.9.2 Le British National Corpus

Le BNC est un corpus de cent millions de mots d'anglais britannique moderne constitué entre 1990 et 1994, par un consortium réunissant des éditeurs de dictionnaires et des chercheurs universitaires. Il s'agissait des Presses Universitaires d'Oxford, Longman et Chambers, et des centres de recherche des universités de Lancaster, d'Oxford et de la British Library.

Les objectifs affichés du BNC étaient de créer « un corpus de langue au moins aussi grand que tout ce qui était librement utilisable jusque-là. Ce nouveau corpus se devait d'être synchronique et contemporain, et constituer un échantillonnage conséquent de l'ensemble de la langue anglaise, à la fois parlée et écrite ». (Burnard, 2007 : 18).

La proportion entre données écrites et données orales est de 10 contre un. Burnard (2007 : 22) rapporte que les raisons de cette disproportion sont matérielles, le coût de la retranscription d'un matériel oral étant dix fois plus élevé que le coût d'ajout d'un texte équivalent en nombre de mots. Le volume du corpus oral est donc d'environ 10 millions de mots même si, poursuit Burnard, « nombreux sont ceux qui suggéraient que, l'oral et l'écrit ayant selon eux la même importance en terme de signification dans la langue, ils devaient être présents en quantité égale dans le corpus ». Ceci relève de la question de la structuration des corpus en vue de leur représentativité ; nous avions également rapporté la structuration détaillée de la

[241] Thèse disponible en ligne :
http://www.theses.fr/2011LEMA3003/document

[242] Le schéma est consultable ici :
http://talkbank.org/software/talkbank.xsd
et sa documentation ici :
http://talkbank.org/software/xsddoc/

partie orale du BNC et des critiques qui furent émises. Le BNC ne fut pas conçu pour être un monitor corpus, tel la Bank of English ou le COCA.

Le corpus est annoté morphosyntaxiquement au moyen d'un système d'étiquetage atteignant un taux de précision de 95%.

La deuxième édition du BNC, aussi connue sous le nom de BNC World Edition a été publiée en décembre 1999, cinq années après la première apparition du BNC. Il s'agissait d'une édition où le taux d'erreurs dans l'étiquetage fut réduit. Mais c'est la troisième édition du BNC, parue en mars 2007 sous le nom de BNC XML Edition qui nous intéresse. Tel que l'indique son nom, le BNC fut édité au format XML, ce qui permit de réduire la complexité du balisage, d'améliorer l'utilisation du corpus et d'accroître la conformité de la structure d'annotation utilisée avec les normes internationales telle la TEI (Burnard, 2007 : 31).

Un corpus tel le BNC est un type de corpus qui manque à la France. Nous pouvons constater l'importance de son édition XML TEI en revenant sur la première édition de 1994 : si le codage alors adopté ne fut pas modifiée pour faire des données du BNC des données compatibles avec la TEI ou si les annotations n'avaient même pas été présentées en LDD, il aurait été très difficile de les exploiter aujourd'hui sans un apprentissage et une assimilation d'une grille de lecture dont l'usage se serait avéré unique. Pour illustrer l'intérêt du XML TEI, des données du BNC sont aujourd'hui consultables via Xiara[243], qui n'est ni un logiciel au format propriétaire, ni une plateforme, mais un moteur de recherche optimisé pour les documents TEI-conformant ; ceci signifie que Xiara peut effectuer des recherches sur tout ensemble de documents XML d'une part, mais l'intérêt réside dans le fait que d'autre part, tout moteur de recherche XML pourra être utilisé pour effectuer des recherches sur les données du BNC, les rendant ainsi indépendantes des évolutions logicielles.

Le BNC est un corpus d'importance de par son volume et sa disponibilité en XML, et les fonds qu'il nécessita en sont consécutifs (1,5 millions de livres sterling (Burnard, 2007 : 18)). Comme nous l'évoquions plus haut, les institutions françaises n'ont pas encore pris l'initiative

[243] Xaira : XML Aware Indexing and Retrieval Architecture. Disponible ici :
http://xaira.sourceforge.net/
Les données du BNC sont ensuite payantes ; une licence personnelle coûte 75£ et une licence institutionnelle 500£. Les licences sont disponible ici :
http://www.natcorp.ox.ac.uk/getting/index.xml?ID=intro

d'un corpus similaire, et même pour les institutions britanniques pour lesquelles le BNC constitue un précédent, de tels corpus sont très rares vu les sommes engagées, comme le note Burnard (2007 : 33) :

> Néanmoins, quand bien même cela serait évidemment intéressant, les chances restent très maigres d'obtenir les fonds nécessaires pour construire une série de corpus semblables au BNC, à intervalles réguliers, disons toutes les décennies.

2.9.3 American National Corpus

L'American National Corpus (ANC) est un projet en cours de développement depuis 1990, visant à la création d'un corpus d'anglais américain, comparable au BNC (Ide, Reppen, & Suderman, 2002, sect. 1). Le projet est financé par un consortium d'éditeurs et de sociétés intéressées par le traitement automatique des langues, notamment pour la constitution de dictionnaires d'anglais américain contemporain.

Le corpus s'est certes inspiré du BNC pour sa structuration mais contrairement à ce dernier, l'ajout de données et leur actualisation sont en cours ; ce qui fait de l'ANC un corpus de type « monitor corpus » (0), qui vise à devenir un corpus de référence d'au moins 100 millions de mots. Pour l'heure, il en compte 22 millions, et nous nous intéresserons tout particulièrement l'Open American National Corpus (OANC), qui constitue la partie totalement libre d'accès de l'ANC, sans achat de licence[244].

L'OANC contient aujourd'hui environ 18 millions de mots, dont 3 200 000 de langue parlée[245] ; plusieurs annotations ont été appliquées aux données, dont l'annotation morphosyntaxique et la lemmatisation.

L'OANC nous intéresse pour deux raisons : en premier lieu, le corpus est disponible en version XML : la pérennité et la facilité de manipulation des données XML que nous discutions plus haut peuvent être illustrées par le fait qu'elles peuvent être soumises aux requêtes du moteur de recherche XAIRA que nous avons présenté pour le BNC ; en second

[244] Les données sont téléchargeables ici :
http://www.anc.org/OANC/index.html#download

[245] D'après le site de l'OANC :
http://www.anc.org/OANC/index.html#download

lieu, il s'agit, à notre connaissance, de la plus importante masse de données librement et gratuitement téléchargeables[246].

Cette rapide présentation des BNC et ANC laissent donc paraître un point essentiel : les anglicistes ont accès à une quantité de données écrites et orales de très grande importance (qu'elles soient gratuites ou payantes), disponibles en XML et donc assurément exploitables tant du point de vue des outils d'accès, que du point de vue du long terme. Ces faits nous permettront de mieux comprendre l'initiative du projet COCOON, dont nous allons parler dans la section suivante.

2.9.4 Pérennité et diffusion des données en France

En France, la constitution de grands corpus à l'image de ceux que nous avons présentés, où à l'instar de corpus de références similaires à la BoE ou au COCA fait défaut, en raison de l'absence d'initiatives institutionnelles ou entrepreneuriales allant dans ce sens. Parallèlement, le nombre de « petits » corpus constitués ou en cours de constitutions fit naturellement éclore l'idée d'un regroupement des corpus français, à défaut de constituer un ou des corpus conséquents. C'est à cet objectif que s'attela par exemple le projet COllections de Corpus Oraux Numériques[247] (COCOON), dont les objectifs est d'accompagner « les producteurs de ressources orales, à créer, structurer et archiver leurs corpus »[248], en vue de constituer ce que COCOON appelle « une archive ouverte », définie ainsi :

> Une archive ouverte est un réservoir dans lequel le dépôt des données est effectué par l'auteur lui-même ou bien par une personne ou une institution (…). La description du document vient alors enrichir un catalogue qui permettra de faire connaître son existence à tous (ou juste à un groupe dans le cas d'un dépôt soumis a un contrôle d'accès) et surtout de pouvoir le retrouver parmi l'ensemble des autres documents. Le rôle d'une archive ouverte se cantonne principalement à la sauvegarde, au référencement et à l'accessibilité des documents.

[246] Les données du COCA sont beaucoup plus importantes et consultables gratuitement, mais nous précisons bien que dans le cas actuel, les données de l'OANC sont également téléchargeables, ce qui n'est pas possible pour les données du COCA.

[247] Anciennement Centre de Ressources sur la Description de l'Oral (CRDO) qui lui-même succéda au Centre de Ressources Numériques (CRN).

[248] D'après le site de COCOON :
http://crdo.risc.cnrs.fr/exist/crdo/

Pour assurer la diffusion et la pérennité des documents, COCOON ne garantit que la maintenance de fichiers « dont le codage et le format sont connus », dont le XML.

Nous citerons également l'Observatoire des pratiques linguistiques de la DGLFLF qui, par le biais du Corpus de la Parole présente ses axes de travail comme suit[249,250] :

- le soutien à des travaux d'étude et de recherche, la coordination et l'organisation en réseaux de ces travaux ;
- la diffusion des informations recueillies auprès des spécialistes, des responsables de politiques publiques et du large public ;
- la conservation, la constitution, la mise à disposition et la valorisation de corpus oraux enregistrés. Ces corpus constituent un outil de travail pour la recherche, mais acquièrent également, avec le temps, un caractère patrimonial.

L'initiative la plus récente est celle des Très Grandes Infrastructures (TGI) de recherche pour les SHS, au nombre de quatre, dont deux nous intéressent.

La première, le Très Grand Équipement Adonis[251] (TGE-Adonis), est une infrastructure de recherche nationale, dont la mission est « d'assurer l'accès et la préservation des données numériques produites par les sciences humaines et sociales », via une aide à l'archivage, la préservation et la diffusion des données SHS grâce à une plateforme d'utilisation.

La seconde est la **C**oopération des **O**pérateurs de **R**echerche **P**our un **U**sage des **S**ources numériques (CORPUS-IR), définie comme suit[252] :

> **Corpus** IR est *une plateforme de coopération* pour l'accès aux principaux ensembles documentaires (images, sons, textes) produits en priorité dans le cadre de recherches dans les SHS en général, et notamment en linguistique, psychologie, histoire, archéologie, philosophie, anthropologie, géographie, lettres et arts.

[249] Du site de l'Observatoire :

http://www.dglflf.culture.gouv.fr/observatoire/observatoire_accueil.htm

[250] Nous rappelons que l'inventaire des corpus français par Cappeau & Seijido (2005), fut établi à la demande de la DGLFLF.

[251] Site du TGE-Adonis :

http://www.tge-adonis.fr/

[252] D'après le site de CORPUS-IR :

http://www.corpus-ir.fr/

CORPUS-IR propose notamment la possibilité d'une aide (matérielle et technique), via une labellisation du corpus. Nous ne présenterons pas ici les conditions de la labellisation[253], ni les recommandations techniques concernant les corpus, disponibles sur le site de CORPUS-IR[254]. Il nous intéresse ici uniquement que les recommandations concernant les métadonnées encouragent le recours au schéma du Dublin Core[255] dans un fichier XML, ou le recours à la TEI.

Ainsi, en ce qui concerne les perspectives de pérennisation des données orales en France, la disponibilité des données au format XML est une obligation. Néanmoins, les initiatives du TGE-Adonis et de CORPUS-IR sont certes récentes, mais ne sont pas sans précédents. Si elles assureront sans doute la pérennisation des données et leur disponibilité grâce à des LDD universaux, il ne nous semble pas que la perspective d'un « réservoir de corpus » qui puisse offrir des résultats similaires à ceux obtenus lors du recours aux corpus anglo-saxons soit envisageable sur le court terme. Soit les linguistes qui constituent des corpus doivent se résigner à l'utilisation de certains standards inévitables quand il est question de regroupement des données, soit les institutions ou entreprises éditoriales françaises devront un jour prendre l'initiative de financement d'un corpus de référence homogène d'un point de vue constitutif et structurel. Que la France dispose, à terme, d'un réservoir de corpus ou de grands corpus de référence, leurs versions XML seront en tout cas indispensables.

2.10 Types de corpus

Nous avons évoqué des corpus dits « de référence », ou « monitor corpus », opposés aux corpus dits « fermés » lorsque nous discutions de représentativité. Nous avons également fait allusion aux « corpus de spécialisation », auxquels le corpus d'apprenants que nous avons constitué peut être apparenté. Nous éclaircirons ces notions dans cette section, avec des exemples de corpus qui méritent d'être présentés ; d'autant que la compréhension des structures et des méthodologies de constitution de ces corpus permettra de mieux comprendre les types de corpus qu'ils représentent.

[253] http://www.corpus-ir.fr/index.php?page=labelisation

[254] http://www.corpus-ir.fr/index.php?page=guide-technique

[255] Le Dublin Core est un schéma de métadonnées générique.

Nous enchaînerons avec une discussion de la langue des corpus, pour tenter de comprendre la complexité de termes en apparence anodins comme « corpus de français » ou « corpus d'interlangue », et ces notions de langues des corpus ont donnée naissance aux « corpus comparables » et « parallèles » d'un point de vue de langue du corpus, que nous discuterons également. Nous conclurons cette section par la problématique des données Internet ; nous tenterons de savoir si le WEB peut être utilisé comme ressource linguistique en tant que corpus.

2.10.1 Corpus statiques et corpus ouverts

Un corpus statique, ou fermé, est un corpus constitué pour représenter un nombre défini de domaines de la langue, à une époque donnée ; sa constitution n'envisage pas l'ajout de données au fil des années, mais une représentation dans le cadre de l'échantillonnage de la partie du langage qu'il s'est proposé d'étudier. À l'instar de Kennedy (1998 : 61), McEnery *et al.* (2011 : 9) comparent le corpus statique à un instantané photographique et le nomment « snapshot ».

Un corpus ouvert (corpus de suivi, corpus de référence, monitor corpus) est un corpus dont la constitution vise la représentation de l'ensemble du langage. Dès lors, ce type d'objectif suppose et implique les notions de masse de données importantes, représentation du langage probabiliste au sens entendu par Sinclair et constitution du corpus au fil des ans, à la fois pour des raisons pratiques mais également pour la représentation diachronique du langage. Kennedy (1998) et McEnery *et al.* (2011) comparent ce type de corpus à un film cinématographique pour poursuivre leur métaphore sur l'instantané photographique des corpus statiques.

Ces catégorisations sont à rattacher à des théorisations postérieures aux premiers corpus, notamment celles de Sinclair (1991) et de quelques grands colloques des années 1990 ayant été organisés pour que le point se fît sur les travaux sur corpus qui avaient été entamés trente ans plus tôt, comme l'état de l'art sur la linguistique de corpus édité par Leech (1991b) ou le symposium dont les actes ont été édités par Svartvik (1992). Ces catégorisations ne sont pas non plus immuables ; ainsi, en 1991, alors que les grands corpus de référence comme le BNC, l'ANC ou le COCA n'existaient pas encore, Leech considérait que le Brown était un corpus

de référence, alors qu'il fut par la suite régulièrement cité[256] mais comme corpus statique, ne représentant « que » la langue écrite des États-Unis des débuts des années 1960 :

> In terms of 'representativeness' we may distinguish relatively general-purpose corpora, such as the Brown and SEU corpora, from corpora designed for a more specialized function -- a domain-specific corpus representing the language of the oil industry, for example. (Leech, 1991a : 11)

Notre présentation commencera donc par les corpus statiques, en prenant pour exemples le Brown Corpus, et de son alter ego le LOB. Nous présenterons ensuite deux corpus de référence, qui sont la BoE et le COCA.

Avant d'entamer cette série d'exemples qui nous permettra de mieux comprendre la typologie des corpus, nous nous devons de dire quelques mots sur « les corpus de spécialisation » : un corpus de spécialisation est un corpus de type statique, mais qui se limite soit à la représentation de situations d'énonciations très particulières, soit à la représentation de domaines de la langue très spécifiques. Habert *et al.* (1997 : 38) évoquent les corpus de spécialisation qui se restreignent à « un langage spécifique, très contraint du point de vue lexical, syntaxique, voire textuel, que l'on trouve dans les domaines scientifiques et techniques ». Sinclair (1996) parle des corpus d'apprenants, d'enfants, de parole pathologique ou gériatrique comme des corpus de spécialisation.

Ainsi un corpus comme le CFPP2000 est un corpus statique, car il représente le français parlé parisien des années 2000, et ces trois spécificités majeures (français parlé, français parisien, français des années 2000) sont trop génériques pour que le CFPP2000 puisse être considéré comme un corpus de spécialisation. La subjectivité de ce jugement peut être atténuée par les possibilités du corpus : un corpus tel le CFPP2000 peut prétendre intégrer un large corpus de référence sur le français, tandis qu'un corpus oral d'apprenants en FLE, tel le nôtre, ne trouverait pas sa place dans une compilation des corpus censés représenter le français, et tend ainsi à être défini en tant que corpus de spécialisation. Les fonctions des corpus de spécialisation restent ponctuelles et destinées à des spécialistes.

Corpus fermés

[256] Notamment dans les ouvrages généralistes sur la linguistique de corpus (Biber, 1993a; Biber et al., 1998; Kennedy, 1998; McEnery & Hardie, 2011; McEnery & Wilson, 2001).

Le Brown Corpus, ou « Brown University Standard Corpus of Present-Day American English », considéré comme le premier corpus électronique[257], constitue un exemple de corpus fermé. Il a été compilé au début des années 1960 par Henry Kučera et Nelson Francis à la Brown University. Le manuel du corpus[258] indique que le corpus est constitué de 1 014 312 mots d'écrits anglais édités aux États-Unis dans le courant de l'année 1961, et dont les auteurs furent tous des locuteurs natifs de l'anglais étasunien. Le corpus comprend 500 échantillons d'environ 2000 mots chacun (Cf. Tableau 4 : Détail des échantillons du Brown Corpus) ; nous rappelons que nous avions discuté dans la section citée les critiques émises envers la structuration du Brown en termes de représentativité, et sur lesquelles nous ne reviendrons pas. Le corpus fut par la suite annoté morphosyntaxiquement, la liste des codes d'annotation est disponible sur le site. Le corpus est disponible en téléchargement libre.

Ce qui nous intéresse, 50 ans plus tard, est de constater qu'un corpus comme le Brown fut à sa constitution et ce pendant de nombreuses années considéré en tant que corpus de référence représentant la langue anglaise dans un sens quelque peu absolu. Deux critères de jugement ont considérablement évolué depuis, à savoir le regard qualitatif et quantitatif porté sur les représentations verticales et horizontales du corpus. En ce qui concerne la représentation verticale, Sinclair (1996) note que quand le Brown fut mis à disposition, le million de mots était considéré « comme un miracle ». Vers le milieu des années 1980, l'ordre de grandeur était de 20 millions de mots avec les données du projet COBUILD, de 200 millions de mots courant les années 1990 et nous atteignons aujourd'hui plusieurs centaines de millions de mots au sein du même corpus, en nous dirigeant sans doute vers le premier milliard[259].

[257] Cette affirmation se retrouve presque systématiquement dans la littérature. Toutefois, Léon (2005 : 41) nuance l'idée en rappelant qu'il y eut au moins un précédent, celui de la numérisation du Trésor de la Langue Française. L'initiative du Trésor fut prise en 1957 lors d'une conférence, « Lexicologie et lexicographie françaises et romanes », en vue de constituer un dictionnaire de français moderne et contemporain (1789-1960). Le titre de premier corpus électronique reviendrait donc au TLF selon Léon, qui concède néanmoins que le Brown, contrairement au TLF, fut mis à disposition des chercheurs qui souhaitaient le consulter dès sa numérisation.

[258] Le manuel est disponible ici :
http://khnt.aksis.uib.no/icame/manuals/brown/

[259] Nous ne parlons ici que de corpus structurés et comprenant plusieurs catégories dont au moins une de langue parlée. En ce qui concerne les corpus uniquement littéraires ou journalistiques, ils dépassent de loin les chiffres

Quant à la représentation horizontale, l'absence de données orales, d'écrits non édités ou de données diachroniques sont autant de lacunes qui eussent été rédhibitoires d'un point de vue contemporain. Le Brown, composé uniquement d'écrits étasuniens des années 1960, est aujourd'hui considéré comme trop spécifique pour représenter la langue anglaise, surtout si l'on compare avec des données telles celles des corpus de référence que nous présenterons ci-dessous. En revanche, les données du Brown ont un potentiel de représentativité horizontale pour que le corpus ne soit pas considéré comme trop spécialisé.

Ce type de jugement, dont nous concédons la part subjective comme pour celui que nous avons exprimé pour le CFPP2000, fait du Brown ce qu'on appelle un corpus fermé dont les deux principaux attributs sont : 1) une représentativité horizontale partielle sans qu'elle ne soit trop spécialisée ou unique et 2) une représentativité verticale figée, c'est-à-dire un volume de données en-deçà des standards contemporains des corpus de référence, et qui ne sont pas actualisées au fil du temps. Ces critères viennent illustrer la métaphore de « l'instantané photographique » du langage qu'offrent les corpus fermés.

Autre exemple de corpus fermé, le Lancaster-Oslo-Bergen Corpus ou LOB Corpus est le fruit d'une collaboration, entre 1970 et 1978, entre l'université de Lancaster, l'université d'Oslo et le « Norwegian Computing Centre for the Humanities » à Bergen. Nous le citons car le LOB a été créé en tant qu'équivalent britannique au Brown Corpus américain, et donc pour l'anglais tel qu'il est écrit en Angleterre. L'équivalence signifie ici que les procédés d'échantillonnage (choix du nombre et des types de textes) utilisés pour le Brown Corpus ont été repris pour le LOB[260]. Il en a résulté un corpus similaire dans la taille et la structuration, le LOB puisant évidemment ses données dans les écrits édités en Angleterre vers le début des années 1960. Les deux corpus Brown et LOB deviennent ainsi deux corpus comparables qui ne diffèrent que par la variation synchronique des données. Ces deux instantanés pris pour la même période illustrent, encore une fois, les caractéristiques des corpus fermés.

---

que nous évoquons ; le corpus littéraire de Google Books atteignait plus de 500 milliards de mots en 2011, d'après Michel *et al.* (2011).

[260] Avec, toutefois, de très légères variations dans le choix du nombre d'échantillons pour certaines catégories. Váradi (2001 : 590) détaille dans un tableau les échantillons des deux corpus ; McEnery *et al.* (2011 : 97) proposent une étude comparative poussée entre eux. Rapidement, le nombre d'échantillons des « écrits populaires » a été diminué de quatre, et les deux catégories « textes de loisirs » et « littérature » ont été augmenté de deux échantillons chacune.

Corpus de référence

Un corpus de référence, régi par la notion de « monitor corpus approach » qui a été notamment proposée par John Sinclair (1991 : 24), qui vise à pourvoir les linguistes d'un corpus dont les représentativités horizontales et verticales sont suffisamment importantes pour permettre tout type d'étude linguistique d'envergure de type calculs de fréquence, constitution de grammaires sur des bases empiriques ou constitution de dictionnaires. Ces impératifs représentationnels impliquent la notion de suivi et d'actualisation des données pour deux raisons. La première raison est d'ordre technique : un volume de données de plusieurs centaines de millions de mots ne peut être obtenu sur un court terme. La seconde raison, qui se confond avec un atout de ces corpus, est la volonté de pouvoir suivre l'évolution de la langue sur la durée de constitution du corpus. Ici se concrétise l'image que nous évoquions plus haut, celle qui compare le corpus de référence à un film cinématographique et non plus à un instantané photographique. Nous constatons donc la nécessité d'un financement continu pour la constitution d'un corpus de référence ou, dans le cas d'un financement ponctuel, la perspective d'un autofinancement dont l'efficience s'avérerait aléatoire, et qui soulève la question éthique de la tarification de données linguistiques. Pour ces raisons de financement, la France est très loin de disposer de tels corpus, et nous laissons la parole à Véronis qui résume la situation ainsi (2000 : 2) :

> Le retard du français dans ce domaine est considérable (…). Ainsi, au moment où le British National Corpus propose 100 millions de mots étiquetés du point de vue grammatical et 10 millions de mots de parole transcrite, la communauté francophone ne dispose même pas de l'équivalent du *Brown Corpus* (1 million de mots étiquetés), réalisé dans les années soixante.

Nos exemples se limiteront donc aux corpus anglo-saxons. Le corpus Bank of English (BoE)[261], constitué à l'université de Birmingham, est un exemple de corpus de référence. Nous rappelons que le BoE est le corpus de travail du projet COBUILD, et qu'il avait été financé par Harper Collins Publishers. La collecte des données commença dans les années 1980, mais les données ne furent regroupées sous le nom de BoE qu'à partir de 1991. Le corpus comptait 7,3 millions de mots en 1987 (Sinclair, 1990 : 185), 200 millions en 1994 (Järvinen, 1994 : 565), 400 millions en 2001 (McEnery & Wilson, 2001: 187), plus de 500

[261] Site de la BoE :
http://www.titania.bham.ac.uk/

millions en 2011 (McEnery & Hardie, 2011 : 7) et près de 650 millions de mots à l'heure où sont écrites ces lignes, d'après le site de la BoE qui présente le corpus ainsi :

> It contains 650 million words from a carefully chosen selection of sources, to give a balanced and accurate reflection of English as it is used everyday.

Les auteurs rapportent également que des données nouvelles sont ajoutées chaque mois. Järvinen (1994 : 565) précise par exemple qu'en 1994, la taille du corpus augmentait d'environ 10 millions de mots mensuellement. Les donnés sont principalement textuelles et, dans l'optique de représentativité probabiliste des auteurs (que nous avons discutée dans la partie consacrée à la représentativité), la stricte égalité entre les catégories n'est pas recherchée, mais un certain équilibre en termes d'ordre de grandeur (entre 10 et 15 millions de mots pour chaque catégorie). Nous retrouvons des catégories du genre « littérature anglaise », données de « BBC », du « Times » ou de « The Economist ». Dans la perspective de référence, les données ne sont pas non plus homogènes d'un point de vue synchronique : nous retrouvons des textes publiés en Angleterre, aux États-Unis ou en Australie.

D'après Sinclair (Wichmann, Fligelstone, McEnery, & Knowles, 1997 : 38-39), un équilibre est maintenu entre les données orales et les différents types de données écrites – journaux, magazines, ouvrages imprimés etc. – et les données orales représentaient en 1997 15 millions de mots collectés de manière hétérogène.

Une partie du corpus, appelée « Wordbank online », est disponible par abonnement payant sur le web. Elle contient 57 millions de mots, écrits et parlés.

Un autre exemple représentatif de corpus de référence est le Corpus of Contemporary American English[262] (COCA). Nous tenons à le citer car il représente le plus grand corpus consultable gratuitement[263]. Sa constitution a été entamée en 1990 sous la direction de Mark Davies, de la Brigham Young University. Le corpus contient à ce jour 425 millions de mots ; depuis 1990, le rythme est d'environ une augmentation de 20 millions de mots par an.

[262] Site du COCA :
http://corpus.byu.edu/coca/

[263] Mais non de manière illimitée : un étudiant ou chercheur en linguistique aura droit à 300 requêtes quotidiennes. Pour davantage de détails, voir sur le site du corpus :
http://corpus.byu.edu/userCategories_e.asp?u=499999.

Le COCA contient cinq types de données : la langue parlée, des écrits fictionnels, des écrits de magazines populaires, des écrits journalistiques et des écrits académiques. Chaque section contient un peu plus de 80 millions de mots et l'équilibre entre ces cinq types de données est respecté lors des nouveaux ajouts. En ce qui concerne les données orales, il s'agit de la transcription d'enregistrements provenant d'environ 150 émissions télévisuelles et radiophoniques.

Les recherches se font directement sur le site du corpus ; il n'est pas nécessaire d'installer de logiciel mais il est demandé à l'utilisateur de s'inscrire gratuitement. Il est possible de rechercher des occurrences particulières, des structures, des occurrences selon leurs catégories syntaxiques, les collocations, les lemmes et les synonymes. Toute combinaison de requête entre ces catégories est possible. En outre, l'interface permet de limiter les requêtes à un ou plusieurs types de données, et de comparer les résultats. Les limitations peuvent se faire également dans le temps, en choisissant l'année de collecte. Enfin, et toujours depuis l'interface, COCA offre la possibilité de comparer les résultats des requêtes avec les données de l'ANC, du BNC, de la BoE et de l'Oxford English Corpus. Ainsi, la recherche des 100 premiers collocats du terme « french », dans la section langue parlée, donne ceci :

Ex 15

> fries british english revolution spanish doors quarter english german bread toast german italian italian speak colonial restaurant minister france troops dutch cuisine jacques paris accent cooking fry translated subtitles laundry germans philosopher jean françois speaks louisiana chirac mister nicolas translation horn riviera chef sarkozy pierre wines bistro alps orleans languages le designer fluent nineteenth-century countryside creole colony painter belgian champagne mitterrand legion polynesia th-century algeria provincial ambassador loaf slices actress twist easel colonies canadians chateau swiss embassy th-century eighteenth-century lieutenant arabic intellectuals italians michel astronomer renaissance baguette cinema georges settlers henri culinary fleet impressionists chefs guiana revolutions cuffs antique indochina

Avec la possibilité d'accéder aux détails de chacun des collocats. Voici quelques autres exemples de requêtes possibles sur le site du COCA[264] :

Ex 16

[264] Ces exemples sont illustratifs des possibilités de requêtes sur le COCA, nous les avons empruntés aux concepteurs, sur le site du corpus.

Past tense verb + over in SPOKEN	ADJ + track in NEWSPAPERS
Synonyms of smart in FICTION	Verbs in MAGAZINES-Sports
Nouns in NEWSPAPERS-Money	Adjectives in ACADEMIC-Medicine
Adverbs in FICTION-Movies	Past tense verb + over in SPOK vs NEWS
ADJ + track in NEWS vs SPOK	Verbs in 2008 vs 2005-2008

Pour conclure sur la notion du volume des données, McEnery *et al.* (2011 : 7) notent que la base de données qui détient le record de taille et de croissance reste le web, qui pourrait être qualifié de « monitor corpus » immense. La question est de savoir si le web peut être considéré en tant que corpus, et cette problématique sera traitée en 2.10.4.

Conclusion sur les corpus statiques et les corpus de référence

La premier axe de typologie que nous venons de présenter n'a d'intérêt scientifique que du point de vue des méthodologies de constitution, qui elles-mêmes reflètent les objectifs du corpus. Cela signifie que nous ne devons pas voir dans un corpus de référence un corpus qui contiendrait, de par son statut, obligatoirement les informations d'un corpus statique constitué dans le même laps de temps. À l'inverse, il serait souhaitable que les corpus plus spécialisés s'affranchissent d'un certain statut quand ils ne sont constitués que pour une recherche ponctuelle. Habert (2000) parle dans ce cas de corpus éphémères, et poursuit avec raison en affirmant que leur réutilisation ne s'impose pas toujours. Or il ne s'agit pas d'imposer des thèmes de recherche, mais il nous semble que la situation du linguiste confronté à une masse de données de plusieurs centaines de millions de mots ne diffère pas de celle du chercheur face à un corpus trop spécifique s'il n'a pas en main la documentation et l'outillage nécessaire pour en faire des sources de données utiles[265].

Nous n'avons explicité les terminologies de « corpus de référence », « corpus statique » ou « corpus de spécialisation », par ailleurs assez versatiles, que pour indiquer trois grandes familles méthodologiques. Dans ce cadre des procédures, une documentation claire, exhaustive et standardisée reste le chemin le plus court pour une exploitation efficace des

[265] Nous citerons ici Jean Véronis qui, dans l'une de ses présentations, illustre par une comparaison judicieuse la position du chercheur face à une masse de données non outillée : « Les Big Data ne nous mettent-elles pas dans le même inconfort intellectuel que la physique quantique ? On observe, on prédit, mais comprend-on vraiment ? ».
La présentation est disponible sur le blog de Véronis :
http://blog.veronis.fr/2012/10/conf-big-data-et-technologie-du-langage.html

données, bien davantage que des représentations horizontales et verticales d'envergure. Il va sans dire qu'un corpus de référence documenté, outillé et facilement disponible dans une version standard comme celle proposée par l'ANC, reste un outil d'importance et dont la communauté linguistique de France devrait se doter.

2.10.2 Les langues du corpus

La documentation d'un corpus nécessite l'établissement d'un ensemble de critères dont celui de la langue des auteurs dans le cas de corpus écrits, et des locuteurs dans celui de corpus oraux. Or si l'on imagine par exemple la volonté de constitution d'un corpus oral de français parlé, l'identification du « français parlé » se verra dans l'obligation de prendre en compte les variations de la langue française. Les variations peuvent être diachroniques, diatopiques ou concerner les variations entre locuteurs natifs et locuteurs non natifs du français.

La question de la variation diachronique se pose principalement dans le cas de corpus écrits, les plus récentes données orales disponibles ne datant pas de plus d'un siècle. Néanmoins, les corpus oraux ouvrant la voie aux études diachroniques devraient devenir de plus en plus nombreux d'ici quelques années. Par exemple, le corpus du *Français fondamental* peut, d'un point de vue théorique, présenter un intérêt scientifique en ce qui concerne la variation diachronique entre le français de Paris des années 1950 et le français de Paris des années contemporaines ; ainsi Bazillon nous livre-t-il un comparatif entre les fréquences des dix mots les plus employés au sein du corpus du « français fondamental » et du corpus contemporain du projet EPAC[266] (Bazillon, 2011, p. 47).

En ce qui concerne la variation diatopique, il y a au monde 29 pays où le français est la langue officielle ou l'une des langues officielles, liste à laquelle il conviendra d'ajouter les pays comptant une communauté francophone importante (Maroc, Algérie, Tunisie, Liban etc.), ainsi que certaines régions ayant une tradition ou une minorité francophone (Vallée d'Aoste en Italie, Pondichéry en Inde etc.). Une fois le pays ou l'entité indiquée, le problème persistera en raison des variations régionales et territoriales de ce seul pays ; cette situation

[266] EPAC : Exploration de masse de documents audio pour l'extraction et le traitement de la PArole Conversationnelle. Voir le site du projet :
http://tln.li.univ-tours.fr/Tln_Epac.html

n'est évidemment pas spécifique du français, les variations diatopiques en anglais, espagnol, portugais, arabe ou autre sont très nombreuses également.

Quant aux variations entre locuteurs natifs et non natifs, il n'est pas aisé de définir ce qu'est exactement un locuteur natif. Meyer rapporte que l'International Corpus of English, dans sa volonté d'inclure des locuteurs natifs américains, les a définis ainsi :

> A native speaker of American English (is) someone who had lived in the United States and spoken American English since adolescence. (Meyer, 2002 : 46)

Nous distinguons au moins trois éléments difficilement évaluables dans cette définition : l'appréciation de la présence réelle ou non sur le sol américain, la possibilité d'évaluer l'emploi ou non de l'anglais sur le sol américain et le consensus sur l'âge de l'adolescence. D'autant plus que s'ajoute aux difficultés d'évaluation de ces deux facteurs la question de leur réelle pertinence : en 2000, 337 langues ont été recensées aux États-Unis : pour près de 47 millions d'américains, soit à peu près 18% de la population, l'anglais n'est pas la langue parlée à la maison (Shin & Bruno, 2003 : table 1). En ce qui concerne l'adolescence, Meyer parle de 10-12 ans. Là également, nous ne trouvons pas de consensus clair. Ces remarques ne constituent pas une critique mais le constat de la difficulté de définir ce qu'est un locuteur natif. Nous avons pris l'exemple flagrant des États-Unis, mais rares sont les pays dont la langue est homogène sur l'ensemble du territoire, même en ce qui concerne des pays dont les populations ne dépassent pas la dizaine de millions d'habitants comme la Belgique ou la Suisse.

Quant aux locuteurs non natifs (anglais), Schmied (1996 : 187) observe que « it can be difficult to determine where an interlanguage ends and educated English starts » (il peut être difficile de délimiter où l'interlangue se termine et où l'anglais natif commence), et les niveaux, la fluidité, la richesse lexicale et la maîtrise de la construction syntaxiques varient énormément entre apprenants. Par ailleurs et en-deçà du niveau atteint par l'apprenant, l'identification de sa langue maternelle, le nombre des autres langues apprises, l'âge de l'apprentissage de la langue cible, l'âge d'apprentissage de la première langue maternelle et un nombre de critères non linguistiques peuvent influer sur l'interlangue.

La richesse des variations d'une langue font en sorte que le corpus homogène pouvant être décrit avec exactitude n'existe pas, en raisons du nombre indéfini de données métalinguistiques pouvant être prises en compte. Néanmoins, le respect d'un protocole de

constitution et de fait, une documentation de corpus sont une condition pour que le corpus soit, non pas homogène, mais descriptible. La disponibilité de la documentation dans le corps du corpus ou en annexe permet au chercheur d'avoir à disposition certaines pistes de réflexion, voire d'explication de l'hétérogénéité de données pourtant constitués selon un protocole. Il ne s'agira donc pas pour les compilateurs d'un corpus de fournir des corpus purs, mais des corpus dont les métadonnées sur les profils linguistiques de ses auteurs ou locuteurs permettront au linguiste de formuler des hypothèses sur les variations qu'il y trouvera.

2.10.3 Corpus parallèles et corpus comparables

Nous avons discuté des questions des corpus structurellement proches ou identiques, ne différant que par leurs variations diachroniques ou géographiques ; ceci nous amène à consacrer la présente section à une présentation rapide des notions de « corpus parallèles » et de « corpus comparables ».

Deux corpus (ou plus) sont parallèles si leurs structures et leurs contenus respectifs sont rigoureusement identiques sauf en ce qui concerne un seul paramètre : la langue du corpus. La variation entre les deux langues peut être diachronique (français parlé parisien des années 1950 / français parlé des années 2000), régionale (français parlé parisien des années 2000 / français parlé de Bruxelles des années 2000) ou enfin, il peut s'agir de deux langues totalement différentes. Ces corpus ont naturellement émergé avec les avancées du TAL et de la TA. Néanmoins l'un des premiers exemples, les deux corpus Brown et LOB dont nous avons parlé plus haut, ne résulte pas de la traduction d'un corpus, mais de la constitution du LOB selon les mêmes procédures d'échantillonnage et de structuration que le Brown. La notion de comparabilité repose donc sur les deux principaux critères de la représentativité : l'échantillonnage et la taille de ceux-ci. Le parallélisme de deux corpus implique donc que ces deux corpus représentent deux langues différentes avec le même potentiel et dans les mêmes proportions.

Le seul parallélisme ne suffit pas à rendre exploitables deux corpus parallèles. Il est nécessaire que les traductions soient alignées ; cela signifie que chaque segment sera aligné avec la traduction lui correspondant. Le processus d'alignement se base sur l'élaboration d'algorithmes (Gale & Church, 1993), qui sont créés selon des critères spécifiques, telles les informations lexicales du texte (Chen, 1993). Les processus d'alignement sont complexes, ceci dû aux différences structurelles qui sont des résultantes usuelles de la traduction, le

parallélisme des corpus n'étant évidemment jamais parfait. Comme exemples de corpus parallèles, nous citerons le Hansard[267] canadien qui est le compte rendu des actes de la Chambre des communes canadiennes en anglais et en français (Somers & Jones, 1992) ; l'Europarl, qui est un corpus de textes parallèles en onze langues provenant des débats du parlement européen (Koehn, 2005), ou encore le « JCR-Acquis Multilingual Parallel Corpus », qui contient des documents juridiques de l'Union européennes en vingt langues officielles de l'Union ; il y a 8000 documents par langage, soit environ neuf millions de mots. Le corpus est aligné (Steinberger et al., 2006).

Contrairement aux corpus parallèles, les corpus comparables peuvent diverger sur plusieurs paramètres, mais restent proches de par leurs structures ou leurs contenus. Les critères de différenciation peuvent être qualitatifs pour des fins stylistiques, comme le genre, l'auteur ou la période (Déjean & Gaussier, 2002) ou quantitatifs se basant sur la fréquence des mots (Kilgarriff, 2001; Rayson & Garside, 2000). Déjean & Gaussier proposent la définition suivante pour deux corpus comparables bilingues :

> Deux corpus de deux langues l_1 et l_2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l_1, respectivement l_2, dont la traduction se trouve dans le corpus de langue l_2, respectivement l_1. (Déjean & Gaussier, 2002 : 2)

Plus les critères qualitatifs et quantitatifs sont communs, plus le degré de comparabilité tend vers le parallélisme. La notion de « sous-partie » recouvrira alors la totalité du corpus, si tant est que le parallélisme parfait puisse exister. L'utilisation de corpus comparables au lieu de corpus parallèles est d'ordre pratique : pour des raisons évidentes et tel que le rapportent Déjean & Gaussier (2002 : 3), les corpus comparables sont beaucoup plus accessibles que les corpus parallèles de bonne qualité.

[267] Un Hansard est la transcription officielle des débats parlementaires dans les gouvernements fondé sur celui du Royaume-Uni. Le Hansard est nommé ainsi d'après Thomas Hansard, qui commença à éditer les transcriptions du parlement britannique en 1803. Aujourd'hui nombre de parlements du Commonwealth publient leur Hansard, comme le Canada, la République d'Irlande, l'Afrique du Sud etc. Ces informations proviennent de « Story of Hansard », sur le site officiel du parlement du Royaume-Uni : http://www.hansard-westminster.co.uk/story.asp

Le « English-Norwegian Parallel Corpus[268] » (ENPC), l'un des premiers couples de corpus parallèles, offre un bon exemple des possibilités d'exploitation de tels outils. La constitution du ENPC fut entamée en 1994 et il contient des échantillons d'environ dix à quinze mille mots chacun, d'écrits fictionnels et non fictionnels en anglais et norvégien (Johansson & Hofland, 1994). Fin 1998, le corpus représentait environ 2 600 000 mots. Plusieurs sortes d'étude peuvent être menées sur un tel corpus : en premier lieu, ils constituent une base de données de référence nécessaire à la création et l'optimisation des logiciels de TA ; en second lieu et grâce aux échantillons alignés entre les deux langues, des études contrastives peuvent être opérée, que ce soit des études stylistiques ou syntaxiques ; enfin, l'alignement bidirectionnel entre les deux corpus permet à des apprenants en classe de langue la comparaison entre un texte en langue source et sa traduction en langue cible.

L'exploitation de corpus parallèles ou comparables en classes de langue montre des résultats encourageants (Bernardini, 2004). L'exploitation peut être aussi celle de corpus parallèles mais unilingues, contenant des échantillons de productions d'apprenants, comparables à des productions de natifs. L'objectif est de focaliser l'attention des apprenants sur leurs erreurs statistiquement récurrentes, et de mettre à leur disposition des productions de natifs aux structures relativement similaires (les échantillons alignés à leurs productions), soit une norme[269] à laquelle ils peuvent se référer. Granger & Tribble (1998) rapportent que cette approche aide les apprenants dans leurs processus d'apprentissage en leur faisant plus ou moins prendre conscience des phénomènes de fossilisation de leur interlangue, et de leur faire initier un processus de restructuration de leurs connaissances[270].

2.10.4 Le WEB en tant que corpus

Nous avions rapporté en 2.6.3.2 l'exemple d'études faites durant les années 1990 et qui ne se basèrent que sur une version numérique du Wall Street Journal ; cet exemple illustrait les possibles dérives inhérentes à la création du WEB. La tentation d'y voir une source de données linguistiques massive et facile d'accès perdure. Or se pose un ensemble de questions

[268] Site du corpus :
http://www.hit.uib.no/enpc/profil2.html

[269] Cette question de « norme », ainsi que toute la démarche en général, a ses critiques, voir Bernardini (2004).

[270] Cette approche est soumise à critique, notamment en termes de norme, d'authenticité et de résultats concrets, voir Bernardini (2004).

liées aux critères définitoires des corpus que nous avons étudiés : une fois les données à disponibilité, il s'agira d'apprécier leur représentativité, leur documentation, les annotations intrinsèques (autrement dit les informations non linguistiques) qu'elles comportent, leur provenance, leur mode de présentation ainsi que leur structuration. Dans cette section, nous tenterons donc de voir quelle est la nature des données que nous pouvons puiser à partir du WEB, et proposerons notre réponse à la problématique suivante : le WEB est-il un corpus ?

Le WEB contient indubitablement une masse de données linguistiques dont l'intérêt n'est pas à discuter. Un point important est à discuter ici, il s'agit du contenu de ces données. Le terme « données du WEB » signifie rien et tout à la fois ; il peut s'agit de données littéraires, journalistiques ou académiques, et la question de la représentativité horizontale des données WEB est alors posée. La démarche initiale du linguiste sera de définir clairement ses objectifs en formalisant le type de données auxquelles il désire avoir accès. Ce même linguiste devra ensuite se constituer une méthodologie d'accès à ces données. Nous allons discuter des méthodologies d'accès aux données du WEB avant de revenir sur les données elles-mêmes.

Plusieurs méthodes d'accès sont possibles. Considérons celle de Kilgarriff *et al.*, qui d'après leur définition d'un corpus, en concluent que le WEB en est un :

> We define a corpus simply as "a collection of texts" (...). The answer to the question "Is the web a corpus?" is yes. (Kilgarriff *et al.*, 2003 : 334)

Il ne s'agit pas d'une conclusion hâtive ou simpliste, car pour Kilgarriff *et al.*, « corpus » ne signifie pas « corpus exploitable » : pour faire d'un corpus un corpus susceptible de fournir une base de travail, ils suivent une méthodologie détaillée par Baroni & Kilgarriff (2006 : 90), qui consiste à filtrer et nettoyer les données, puis à les annoter morphosyntaxiquement et à les lemmatiser afin que les requêtes linguistiques puissent avoir lieu. Il ne s'agira pas ici de détailler leur méthodologie, mais de constater que même pour des auteurs reconnaissant pleinement le statut de corpus au WEB, sa consultation sans préparation préliminaire ne peut avoir lieu.

Les raisons de l'épuration tiennent de l'inaptitude des moteurs de recherche à répondre à des requêtes linguistiques. En effet, les résultats de moteurs tels Google ou de Yahoo contiennent des liens commerciaux, des doublons, des liens périmés et nombre de balises spécifiques aux pages WEB. Le recours aux données par le biais de moteurs de recherche est non seulement inefficace, mais les résultats peuvent s'avérer erronées même pour des recherches de

fréquence de mots, comme l'a montré Véronis[271]. Par ailleurs, et en comparaison avec des moteurs de recherche destinés aux corpus, les moteurs de recherche traditionnels ne permettent pas des recherches selon les registres de langue, les recherches selon la date sont très aléatoires (les moteurs traditionnels indexent la date de création de la page WEB et non la date de création de son contenu), les données du WEB ne sont pas annotées contrairement à celles d'un corpus et enfin, le concordancier d'un corpus peut combiner plusieurs recherches.

L'absence de toute annotation linguistique ou métalinguistique contraint donc l'utilisateur d'un moteur de recherche traditionnel à se limiter à la forme lexicale et orthographique traditionnelle, et à devoir fouiller et dépouiller les résultats. C'est pourquoi certains chercheurs ont tenté d'offrir à la communauté linguistique un outillage approprié ; nous citerons par exemple WebCorp[272], qui se présente comme un ensemble d'outils permettant l'accès au WEB en tant que corpus. L'accès est gratuit. WebCorp ne possède néanmoins pas son propre moteur de recherche et se base sur les moteurs de recherche traditionnels comme Google. Pour ce que nous avons pu en juger, les résultats sont peu satisfaisants. Nous avons par exemple tenté de rechercher l'emploi de la double pronominalisation ; dans un corpus traditionnel, nous aurions effectué la recherche sur les catégories grammaticales, pour WebCorp, nous ne pouvons que nous limiter à une seule combinaison par recherche, « te + le » pour ce qui suit (une partie des résultats) :

Ex 17

65: Lyrics-Copy en : Home > F > Frédéric François > Je **te le** jure www Search: Browse Artists: # A B C D E
66: T U V W X Y Z Ringtones Karaoke Je **te le** jure by Frédéric François Send "Je **te le** jure" Ringtone
67: Karaoke Je **te le** jure by Frédéric François Send "Je **te le** jure" Ringtone to your Cell Je **te le** jure Aucune
68: Send "Je **te le** jure" Ringtone to your Cell Je **te le** jure Aucune femme avant n´a su me plaire Tu as
69: Quand je t´ai vue mes yeux se sont ouverts Je **te le** jure Je **te le** jure J´étais barbare, tu m´as rendu Send "Je **te le** jure" Ringtone to your Cell MP3 Backing Tracks & Karaoke Videos

Les résultats sont redondants, très peu pertinents (paroles de chansons) et comportent des liens commerciaux. Le recours à d'autres outils similaires a donné des résultats encore plus décevants.

[271] L'étude est disponible sur son blog :
http://blog.veronis.fr/2005/02/web-le-mystre-des-pages-manquantes-de.html

[272] WebCorp a été créé et est maintenu au Research and Development Unit for English Studies (RDUES), à la Birmingham City University. Le site de WebCorp est disponible ici :
http://www.webcorp.org.uk/live/index.jsp

Le concordancier du WEB n'existe donc pas encore. Les études se basant sur des données du WEB passent pour la plupart par une étape de préparation similaire à celle que nous évoquions plus haut pour le corpus constitué par Kilgarriff *et al.* Nous citerons ici une expérience d'Ide *et al.* (2002) qui visait à répondre à deux problématiques : les données du WEB peuvent-elles fournir des données susceptibles d'intégrer le corpus ANC et, le cas échéant, l'obtention de ces données peut-elle être automatisée ? La conclusion des auteurs (2002, sect. 7) est que l'intégration de données du WEB à un corpus nécessite « un travail considérable, parfois manuel » afin que lesdites données soient linguistiquement exploitables.

L'étude date de 2002 mais la situation actuelle est toujours la même. Notre conclusion est la suivantes : le WEB n'est pas un corpus, ni un réservoir de corpus, mais un réservoir de données. La concrétisation des données en un corpus exige leur structuration et leur annotation en prenant en compte chacun des critères définitoires que nous avons étudiés tout au long de ce chapitre, les données doivent être :

1) collectées : cette étape est sans doute celle qui est la plus facilitée grâce au WEB, mais reste ardue. La collecte des données implique leur sélection, leur filtrage et leur épuration ;
2) numérisées : les données sont par défaut numériques puisque provenant du WEB, mais nécessitent leur conversion au format final décidé pour le corpus à constituer ;
3) représentatives : non pas dans le sens absolu, nous entendons ici que l'utilisateur du corpus à constituer doit pouvoir, grâce à la documentation du corpus, savoir ce que représente ces données ;
4) annotées : au minimum transcrites pour les données orales, et documentées quelles qu'elles soient ;
5) standardisées : aux formats que nous avons discutés dans la section que nous avons consacrée à la standardisation des corpus.

Ces étapes, qui peuvent paraître longues et fastidieuses, sont pourtant les mêmes qui régissent toute constitution de corpus. Le WEB est donc loin de représenter le corpus ultime qui rendrait inutile les travaux des linguistes de corpus des cinq dernières décennies, comme l'exprime Leech qui évoque les limites des moteurs actuels :

> Perhaps the future will bring 'intelligent search engines' (…). Meanwhile, while the internet is an added resource of immense potential, it does not remove the need to improve and update other textual resources, and does not render obsolete the corpus compiled according to design and systematic sampling. (Leech, 2006 : 146)

Nous nous devons maintenant de discuter d'un point que nous avions abordé plus haut, celui du contenu des données, problématique que nous allons lier à la méthodologie. Nous répétons que le contenu des données WEB est aussi varié que la langue elle-même, et que le choix des données à étudier doit se faire en amont de la constitution du corpus. Nous préconisons pour l'heure, puisque les moteurs dont parle Leech n'existent pas encore, le recours non pas aux moteurs de recherche, mais directement aux sites concernés selon le choix des données à étudier. Le choix des sites sera évident selon la volonté de constituer un corpus journalistique, littéraire, d'écrits académiques ou de données audiovisuelles à transcrire.

Nous constatons que quel que soit le domaine à étudier, le WEB n'est qu'une nouvelle ressource abondante et facile d'accès, qui n'offre pas vraiment de données dont la nature linguistique serait nouvelle, sauf pour ce qui concerne un point, celui de la communication virtuelle, et c'est sans doute un des points cruciaux des données WEB. L'étude de la syntaxe et du lexique sur les forums de discussion, les commentaires des articles, les messageries instantanées ou sur les sites personnels constituent de notre point de vue le seul pilier de ce qui pourrait s'appeler « les données WEB ». Un corpus littéraire, même s'il provient du WEB, reste un corpus littéraire, et non un corpus WEB[273], tandis que cette dénomination nous semble justifiée dans le cas d'un corpus de communications collectées à partir de la blogosphère francophone. À propos de « la langue Internet », nous laissons la parole à Véronis :

> Dans toutes les communautés et de tous les temps se sont développés des sociolectes et des argots, parfois totalement incompréhensibles par le commun des mortels, l'argot des gueux et de voleurs, etc. Peut-être que ce qui nous surprend autant, c'est que c'est la première fois où il naît par écrit, et de façon visible par toute la planète ? [274]

Les études concernant le sociolecte des communautés virtuelles sont nombreuses, qu'elles soient syntaxiques, morphologiques, lexicales ou traitant des contraintes de certaines situations de communications comme les messages limités à 140 caractères sur le site Twitter.

[273] Il est toutefois vrai que les journalistes n'écrivent pas leurs articles de la même façon s'ils sont destinés à être diffusés en ligne ou dans la version papier de leur journal, en ayant recours à un style plus concis et plus accrocheur (voir à ce propos Dagiral & Parasie (2010)). Ces particularités sociotechniques, selon le terme des auteurs, ont une importance qui justifierait de différencier « presse journalistique » de « journalisme en ligne ».

[274] La citation est un commentaire de l'auteur sur l'un de ses articles, disponible ici : http://blog.veronis.fr/2010/10/lexique-la-toomuchite.html

Nous ne prétendons évidemment pas à attirer l'attention sur un phénomène nouveau ; nos remarques ne visent qu'à situer ce que signifie le terme données du WEB, et nous y voyons donc essentiellement ces sociolectes. Là encore, la collecte d'un corpus sur ce type de données peut se passer des moteurs de recherche traditionnels, puisqu'elle pourra se faire directement sur les sites concernés. La constitution d'un corpus à partir des forums de discussion francophones peut avoir comme base de départ le téléchargement de sujets généralistes qui nécessiteront ensuite une épuration des données. La masse de ces échanges interactifs extrêmement variés dépasse de loin la quantité de données disponibles dans les corpus traditionnels.

Outres les études, les données WEB ont également donné naissance à plusieurs sites supervisés par des linguistes. Nous avons maintes fois cité le blog de Jean Véronis, et nous tenons à citer le Language Log[275], qui est un blog collaboratif dont la maintenance est assurée à l'University of Pennsylvania par le phonéticien Mark Liberman ; les principales thématiques traitent de la langue des médias et d'Internet ainsi que de la culture populaire en général. Le recours à Google en tant que moteur de recherche de corpus a souvent lieu pour la vérification ou l'assertion de théories sur le langage.

En conclusion, la richesse, la diversité et la facilité d'accès aux données WEB est à nos yeux une opportunité dont n'auraient pu rêver les linguistes il y a tout juste une quinzaine d'années, et tout semble indiquer que le potentiel de ces données va de pair avec la croissance du WEB. Si le WEB ne peut encore être listé en tant que « type de corpus » à part entière, nul doute que la maîtrise de ses outils sera cruciale pour les linguistes de corpus dans les années à venir, au risque de passer à côté de la chance de puiser dans un tel réservoir.

[275] Consultable à l'adresse suivante :
http://languagelog.ldc.upenn.edu/nll/

2.11 Conclusion du second chapitre

Ce second chapitre nous a permis de constater que la linguistique de corpus est une linguistique qui préconise l'approche empirique du langage en faisant appel aux corpus linguistiques. Essentiellement une méthodologie pour certains, théorie linguistique pour d'autres et ignorée par les tenants d'une approche rationnelle, la linguistique de corpus est depuis plusieurs années polyvalente : elle permet la construction de grammaires et de dictionnaires, elle sert à vérifier des hypothèses émises mais aussi à observer sans *a priori* des constantes linguistiques non intuitives.

Il est de nos jours raisonnable d'exiger que les données d'un corpus soient numériques. Le corpus est donc un ensemble de données numériques collectées selon un protocole de collecte à des finalités exprimées. Le corpus doit contenir autant d'informations que possible sur les méthodes de collecte, les situations d'énonciation et les locuteurs, car c'est uniquement grâce à cette documentation que la représentativité d'un corpus peut être jugée. En ce qui concerne les annotations, seule la transcription des corpus oraux nous semble exigible ; tous les autres types d'annotation doivent cependant suivre les conditions du schéma d'annotation. Elles doivent donc être suppressibles, documentées et ne pas altérer les données brutes. Aucune taille minimale ne peut être délimitée, mais des données équilibrées et documentées sont à privilégier par rapport à un plus grand volume de données moins susceptibles d'être correctement exploitées. Les protocoles de constitution, de collecte, d'annotation doivent privilégier la clarté à une complexification qui diminuerait le potentiel d'exploitation du corpus. Particulièrement en ce qui concerne le protocole de transcription mais également pour l'ensemble des protocoles, deux impératifs nous paraissent indispensables : le premier est de les fournir (ce qui signifie qu'ils ont été constitués), le second est de les élaborer dans une perspective de diffusion et de mise à disposition de la communauté scientifique.

La constitution d'un corpus est donc un processus relativement libre dans ses finalités et les faits de langue qu'il compte représenter, mais rigoureux au sens où les corpus constitués ne doivent plus uniquement servir les équipes qui les ont élaborés. Quand Robert Lees déclarait à Nelson Francis que la constitution du Brown corpus était une « perte de temps et de l'argent du gouvernement », il était sans doute dans le tort. Mais constituer d'autres « corpus fantômes », de nos jours où la numérisation et Internet permettent l'accès aux données à un

niveau international, serait effectivement une perte de temps et d'argent. Nous conclurons par les termes d'Habert *et al.* (1997 : 9) qui, tenus en 1997, sont plus que jamais d'actualité :

> Cette utilisation de corpus annotés, de grande taille, variés et assortis d'outils d'exploration puissants, permet d'observer plus finement les phénomènes et remet en question une partie des postulats de la linguistique.

3.Chapitre 3 : Corpus CIL-FLE

3.1 Raisons d'être du corpus CIL-FLE et description générale du corpus

Nous rappelons[276] que les corpus, dans le domaine de l'enseignement des langues, peuvent soit servir à la création d'outils didactiques comme les dictionnaires pour apprenants basés sur corpus (Longman's Learner Corpus[277]), soit être eux-mêmes utilisés en tant qu'outils d'exploration de la langue par les apprenants[278], soit constituer des bases de données de productions orales et écrites d'apprenants, destinées à renseigner les chercheurs et les enseignants sur l'acquisition de la langue tel que le formule Meyer :

> The idea behind using learner corpora to develop teaching strategies is that they give teachers an accurate depiction of how their students are actually using the language, information that can then be incorporated into textbooks and lesson plans. (Meyer, 2002 : 27)

C'est au troisième type d'usage que nous nous intéressons et en l'occurrence, les corpus d'apprenants peuvent dans ce cas aussi bien concerner des apprenants en cours d'acquisition (Langue étrangère) que des immigrants en contact quotidien avec une langue qui n'est pas leur langue maternelle (Langue seconde). Tel que le rapporte Granger, l'intérêt pour ce type de corpus ne remonte pas à plus de 25 ans en ce qui concerne l'anglais :

> However, investigations of non-native varieties have been a relatively recent departure: it was not until the late 1980s and early 1990s that academics and publishers started collecting corpora of non-native English, which have come to be referred to as learner corpora. (Granger, 2002 : 5)

Quant au français, soit les corpus d'apprenants en FLE, ils sont encore plus récents et très peu répandus. Citons par exemple le projet Frida[279], lancé en 1998 à l'Université catholique de Louvain en Belgique, constitué de 200 000 mots de productions écrites d'apprenants en FLE. En productions orales, les données les plus importantes dont nous ayons connaissance sont les corpus regroupés par le projet FLLOC[280] de l'Université d'Essex en Grande-Bretagne, où

[276] Cf. 1.7.

[277] Cf. 1.4.

[278] Cf. 1.7.3.

[279] French Interlanguage Database : http://sites.uclouvain.be/cecl/projects/Frida/frida.htm

[280] French Learner Language Oral Corpora :

plusieurs corpus différents de productions orales en FLE peuvent être téléchargés[281] ; le projet LANCOM[282], à l'Université catholique de Louvain également, indique mettre à disposition 18 heures d'enregistrements et leurs transcriptions, mais le projet semble être abandonné depuis 2001. Nous remarquons ainsi que les corpus les plus significatifs en FLE n'ont pas été constitués en France. La consultation du recensement de l'IRCOM[283] amène au constat suivant : il n'y a en France aucun corpus d'apprenants en FLE en consultation libre. Les principaux projets terminés ou en cours concernant l'acquisition d'une L2 sont le projet COREIL[284] et l'IPFC[285] ; non seulement les données de ces deux projets ne sont pas mises à disposition, mais la finalité de ces deux corpus est une étude phonétique, phonologique et prosodique de l'interlangue.

Plusieurs raisons justifient ainsi la création d'un corpus oral d'apprenants en FLE ; elles sont intrinsèques au processus lui-même mais aussi contextuelles. Elles sont en premier lieu intrinsèques au processus car il va de soi qu'une analyse de l'interlangue ne peut reposer sur des intuitions personnelles, et que le chercheur se doit de travailler sur des données authentiques. En second lieu, elles sont contextuelles car les corpus d'apprenants font défaut en France, et les rares projets ayant construit leur corpus ont privilégié les études phoniques. Ainsi, au contraire des corpus COREIL ou IPFC, les données du corpus que nous allons présenter ont été constituées dans l'objectif d'analyses syntaxiques, lexicales ou morphologiques, et non strictement phoniques. C'est dans ce contexte que le projet CIL (Corpus Inter Langue), en cours de finalisation à l'Université Rennes 2, a été lancé en 2008. Dans sa globalité, le projet CIL rassemble les productions orales et écrites d'apprenants en FLE et ALE. Nous avons participé à l'ensemble des axes du projet mais avons pris la

http://www.flloc.soton.ac.uk/

[281] Les différents corpus et leurs descriptions sont disponibles à l'adresse suivante :
http://www.flloc.soton.ac.uk/list.html

[282] LANgue et COMmunication :
http://bach.arts.kuleuven.be/elicop/ProjetLANCOM.htm

[283] Nous rappelons que l'IRCOM est le Consortium Corpus Oraux et Multimodaux, l'inventaire des corpus est disponible à l'adresse suivante :
http://ircom.huma-num.fr/site/corpus.php

[284] Projet visant à la constitution d'un corpus oral d'apprenants de français L2, débuté en 2010, non achevé, ne mettant pas à disposition les données déjà constituées pour le moment, cf. fiche de l'IRCOM :
http://ircom.huma-num.fr/site/description_projet.php?projet=coreil

[285] Cf. note n° 102.

responsabilité de la partie orale FLE[286] (dorénavant CIL-FLE), et c'est spécifiquement de ce corpus que nous traiterons dans ce chapitre, en présentant en détail son contenu, sa documentation et ses protocoles de constitution et de transcription. La nécessité d'une présentation exhaustive vise à faire d'un ensemble de données un corpus scientifique au sens que nous avons défini tout au long du second chapitre. En effet, tel que le formulent Habert *et al.*, un corpus dont les protocoles de collecte et de discussion ne sont pas disponibles n'est pas exploitable :

> Sans une documentation jointe, un corpus est mort-né. L'un des dangers de la facilité actuelle à rassembler des textes électroniques est précisément que les objectifs du regroupement ainsi que ceux des annotations effectuées ne soient pas enregistrés : le corpus cesse d'être utilisable dès que se perd la mémoire de ces choix. La documentation doit couvrir deux volets distincts : les sources utilisées et la responsabilité éditoriale de constitution du corpus d'une part, les conventions d'annotation d'autre part. (Habert *et al.*, 1997 : 156)

3.1.1 Locuteurs, enregistrements et archivage

Les locuteurs enregistrés sont des adultes, hommes et femmes, ayant pour langue maternelle l'arabe, l'anglais, l'espagnol ou le chinois mandarin. Le choix de ces quatre langues permet d'une part de trouver facilement des apprenants à enregistrer, et la limitation des langues maternelles des apprenants à quatre langues permet d'autre part d'envisager des études comparatives entre apprenants d'une même langue maternelle et apprenants de différentes langues. Les locuteurs sont en cours d'apprentissage du français ou ont appris le français. Ils sont tous d'un niveau allant de B1 à C1 selon le CERCL (Cadre Européen Commun de Référence pour les Langues). À ce jour, CIL-FLE regroupe les données de 49 apprenants. 11 d'entre eux ont pour langue maternelle l'arabe levantin septentrional, 8 l'anglais (Angleterre, États-Unis et Nouvelle-Zélande), 12 le chinois mandarin et 18 l'espagnol (Colombie, Venezuela, Pérou, Cuba, Mexique, Espagne). Les données orales obtenues constituent 49 entretiens d'une durée moyenne de 23 minutes par entretien, ce qui représente un total de près de 19 heures d'enregistrement, soit 130 000 mots dont 105 000 produits par les apprenants.

Les données ont été recueillies par nous-même pour une partie des locuteurs arabes, en Syrie. Les autres enregistrements ont été effectués par les étudiants de Master 1 en Linguistique et

[286] Comme nous allons le voir, la partie FLE en général comporte un entretien, une lecture imposée et une production écrite. C'est uniquement à l'entretien que nous nous intéresserons en détail.

Didactique des langues à l'Université Rennes 2. Depuis l'année universitaire 2009-2010 et dans le cadre du module « C3B 8M Enseignements méthodologiques », les étudiants du Master 1 LDL participent à la constitution du corpus CIL. Cette démarche a permis le recueil de 8 à 10 enregistrements annuels. Les enregistrements ont tous été effectués avec un Edirol R09. Les fichiers audio sont au format WAV 16 Bit, et la taille de CIL-FLE est d'environ 13 giga-octets. La formation des étudiants du Master 1 concerne plusieurs niveaux. Après leur avoir expliqué la nature du corpus et ses objectifs, nous les formons à l'entretien, à l'utilisation des appareils d'enregistrement, à l'utilisation du logiciel de transcription CLAN et à la transcription des données orales. Cette dernière étape est soumise à des contraintes à la fois inhérentes au processus de transcription dans sa globalité mais également aux difficultés techniques logicielles. En raison de l'impossibilité de parfaitement former un groupe de 50 personnes à la transcription en quelques séances, et de certains obstacles techniques parfois rébarbatifs, les protocoles distribués et exploités par les étudiants ont été simplifiés. Après avoir récupéré les premières versions des étudiants, nous en avons effectué une vérification à la fois technique et linguistique, et nous les avons corrigées afin qu'elles répondent aux conditions des protocoles.

Les séances d'enregistrement comportent trois tâches : un entretien, une lecture imposée et une tâche de production écrite. En ce qui concerne les entretiens, ils ne sont pas préparés et il n'y a aucune question prédéfinie. Les entretiens ne sont pas libres pour autant car il a été mis en place des directives que les enquêteurs ont tenté de suivre : une présentation de l'apprenant par lui-même, l'évocation de faits du passé, l'évocation de descriptions et de généralités, les projets d'avenir du locuteur et un argumentaire. Ces directives ont pour ambition l'obtention d'un entretien représentatif des faits de langue, afin de permettre la diversification des analyses. Ensuite se déroule une lecture imposée d'un seul et même texte pour tous les apprenants[287]. La longueur moyenne de cette lecture est d'environ deux minutes. Ces lectures permettront des analyses phonologiques, phonétiques ou prosodiques. Enfin, la séance se termine par une production écrite[288] en présence de l'enquêteur, immédiatement après l'entretien. Les dictionnaires ne sont pas permis et le laps de temps imparti est de 30 minutes pour une production d'une à deux pages.

[287] Cf. annexe 2.

[288] Cf. annexe 2.

Afin de faciliter l'archivage, les recherches et les commandes d'analyse de CLAN par la suite, les fichiers ont été nommés sur le modèle suivant pour un seul locuteur :

- eng_ca_re_83_f_11.wav
- eng_ca_re_83_f_11_lecture.wav
- eng_ca_re_83_f_11.cha
- eng_ca_re_83_f_11.docx
- eng_ca_re_83_f_11.jpg

Le premier fichier est le fichier sonore de l'entretien, le second le fichier sonore de la tâche de lecture, le troisième le fichier de transcription, le quatrième la saisie sous WORD de la tâche de production écrite et le dernier un scan de la production écrite de l'apprenant. Les noms des fichiers sont à lire comme suit :

- eng : anglais, langue maternelle de l'apprenant. Cela sera *ara* pour l'arabe, *zho* pour le chinois et *spa* pour l'espagnol. Ces codes sont ceux de la norme internationale ISO-639 pour la représentation des langages qu'utilise le logiciel CLAN ;
- ca_re : sont les deux premières lettres du prénom et du nom de l'apprenant. Conformément aux recommandations du *Guide des bonnes pratiques* pour la constitution des corpus oraux (Baude *et al.*, 2006 : 67), toute possibilité d'identification des locuteurs a été supprimée (nous avons remplacé les passages sonores concernés par des bips). Toujours conformément au *Guide* (Baude *et al.*, 2006 : 68), une version brute non anonymisée a été sauvegardée ;
- « 83 » indique l'année de naissance du locuteur, « f » son sexe et « 11 » indique 2011, l'année de l'enregistrement.

Ces métadonnées ont été associées au fichier de transcription au format .cha. Toutes les possibilités de recherche (outil de recherche par nom ou par *tag*, outil de recherche de CLAN) sont donc offertes. Les raccords et les coupures nécessaires sur les fichiers audio ont été effectués avec le logiciel Audacity, qui a également servi à supprimer toute information d'ordre privée des enregistrements. Conformément au *Guide* (Baude *et al.*, 2006 : 60), les locuteurs ont tous été dûment informés des finalités des séances d'enregistrement, et ont signé

une autorisation d'exploitation scientifique des données obtenues[289]. Ils ont en outre rempli une fiche signalétique détaillée ; ces fiches ont été numérisées et jointes au corpus[290]

3.1.2 CHILDES et logiciel de transcription CLAN

CHILDES[291] est un projet créé en 1984 par Brian MacWhinney qui regroupe plusieurs corpus oraux d'acquisition du langage en tant que langue première, seconde ou étrangère. Il vise notamment à constituer une base conséquente de corpus multilingues, consultable et gratuite, à faciliter la recherche en mettant à disposition des chercheurs les outils logistiques et les méthodologies nécessaires à la création de corpus, à favoriser l'entraide entre les chercheurs s'intéressant aux sciences du langage et à mettre à disposition un espace commun pour les corpus en acquisition du langage, où au moins les protocoles techniques sont homogènes grâce au logiciel CLAN.

Le logiciel CLAN est un logiciel de transcription et d'analyses des données orales, écrites ou filmées. Il est gratuit et téléchargeable sur le site de CHILDES[292]. Nous n'envisageons pas ici d'offrir un manuel d'utilisation ou d'exploitation du logiciel, car le site de CHILDES met à disposition un manuel de transcription[293], ainsi qu'un manuel d'exploitation[294]. Ces manuels sont en anglais et régulièrement mis à jour en fonction de l'évolution du logiciel. Chacun d'entre eux compte près de 220 pages, ce qui démontre le potentiel de CLAN et nous ne présenterons ici les fonctionnalités du logiciel que de manière succincte dans le cadre d'expliciter certains choix de transcription et quelques-unes des possibilités d'exploitation du corpus. Nous signalons que les membres du projet Colaje ont rédigé en français un guide simplifié du logiciel (Bourdoux *et al.*, 2011)[295].

Cependant, nous consacrerons quelques lignes à justifier le choix de CLAN pour la transcription des corpus du projet CIL. En premier lieu, CLAN offre une interopérabilité satisfaisante avec les autres logiciels de transcription ; les fichiers .cha de CLAN sont

[289] Cf. annexe 3 pour la consultation formelle des autorisations.

[290] Cf. annexe 4 pour la consultation de la fiche signalétique.

[291] http://childes.psy.cmu.edu/

[292] http://childes.psy.cmu.edu/clan/

[293] http://childes.psy.cmu.edu/manuals/chat.pdf

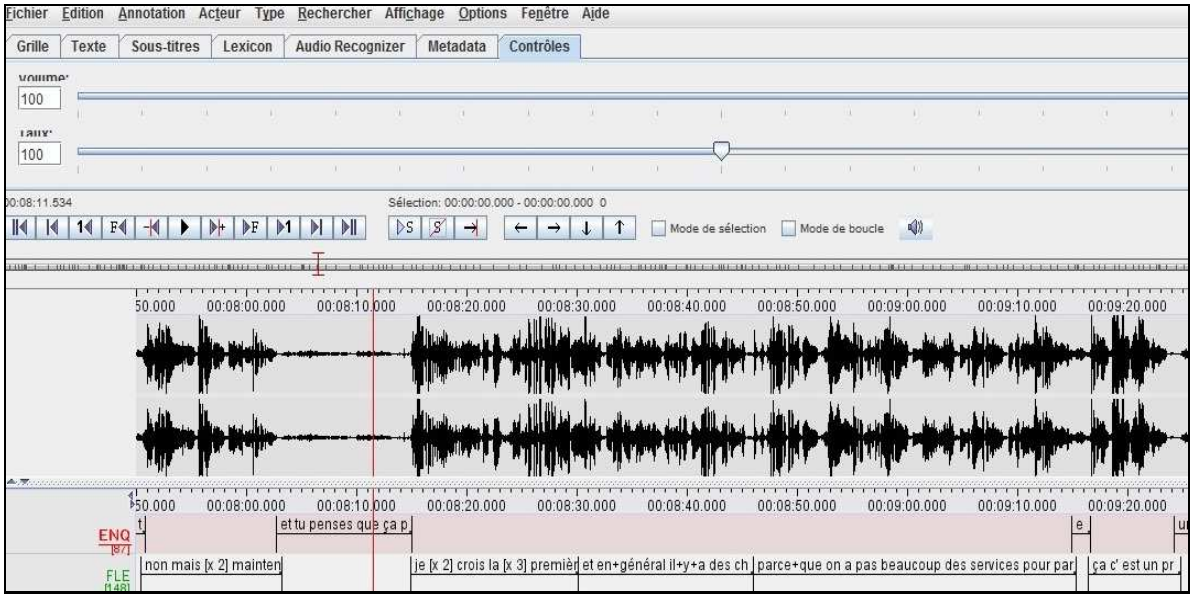
[294] http://childes.psy.cmu.edu/manuals/clan.pdf

[295] http://colaje.scicog.fr/images/stories/PDF/Guide-CLAN_colaje-juin-11.pdf

convertibles et peuvent être lus avec les autres principaux logiciels de transcription tels PRAAT, PHON ou ELAN. Ces conversions ne sont pas systématiquement satisfaisantes, mais économisent néanmoins un temps de travail non négligeable au cas où il serait nécessaire de traiter les données de CIL au moyen d'un autre logiciel. D'autre part, CHILDES met à disposition un outil de conversion des données au format XML[296], ce qui assure d'une part la pérennité des données et donc la possibilité de les exploiter avec des outils à venir.

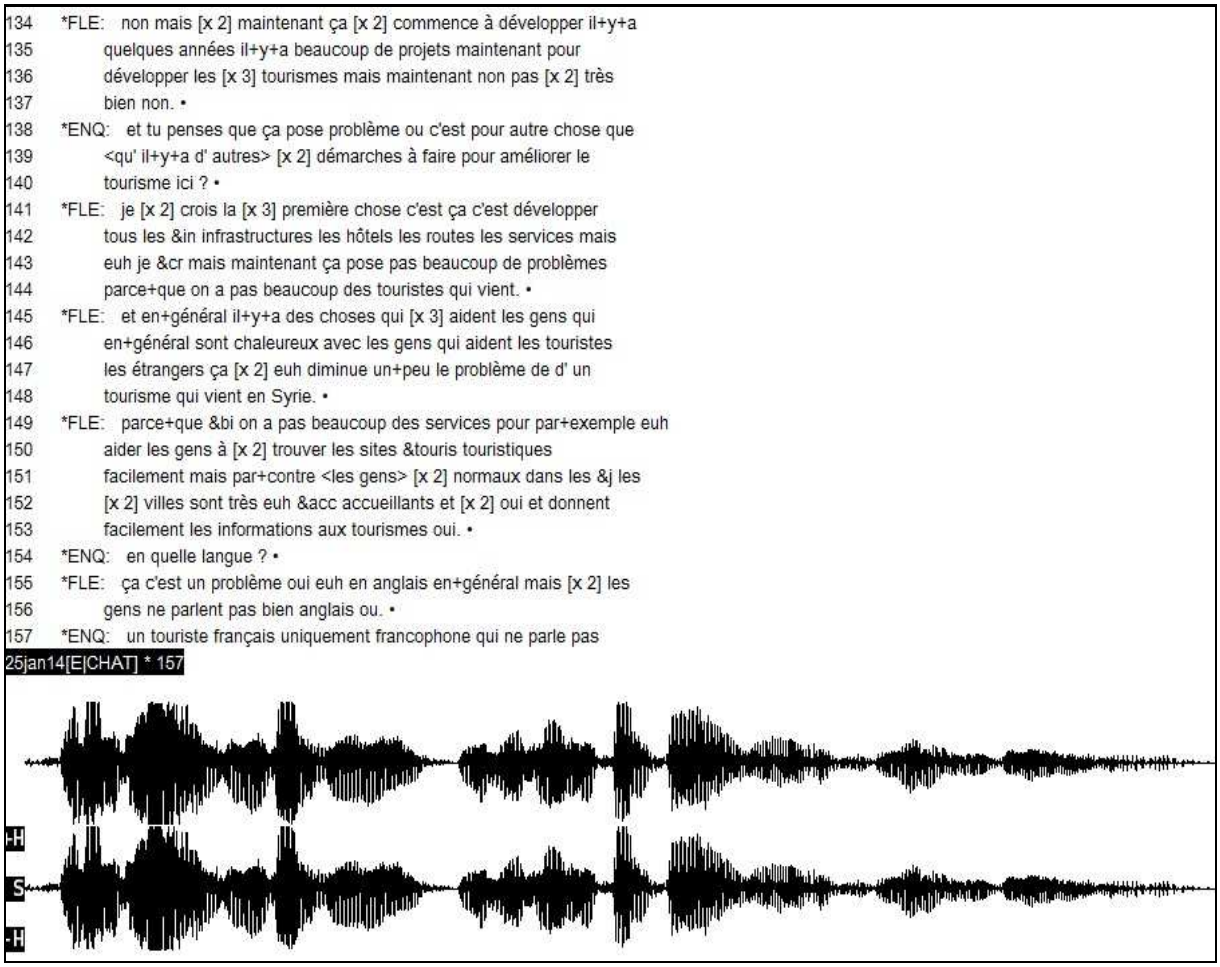
Néanmoins, le choix de CLAN et non d'un autre outil ne se résume pas à ces aspects techniques, mais réside essentiellement dans son efficience dans le cadre d'une analyse discursive. En effet, les logiciels de transcription offrent une représentation spatiale des transcripts qui peut être soit horizontale sous forme de « partition musicale », soit verticale. Voici deux représentations du même segment sonore de 1 minute 30 secondes, le premier est la représentation horizontale sous ELAN, le second est la représentation verticale sous CLAN :

Figure 3: Représentation spatiale des transcripts sous forme de partition horizontale



[296] Les fichiers de transcription ont tous été convertis au format XML, et la version XML du corpus est disponible sur la clé USB jointe. Cf. annexe 1.

Figure 4 : Représentation spatiale des transcripts sous forme verticale



Le premier type de représentation est à privilégier dans trois situations essentielles :

1) dans le cas de corpus destinés à des analyses phonétique ou phonologique (corpus IPFC) :
2) dans le cas de corpus ou plus de deux locuteurs sont en présence, et où des phénomènes comme les chevauchements ou la prise de parole ont leur importance (corpus d'analyse conversationnelle) ;
3) dans le cas de corpus s'intéressant à des phénomènes paralinguistiques comme le langage paraverbal ou la gestuelle (corpus multimodaux).

En ce qui nous concerne, notre corpus n'est pas un corpus d'analyse conversationnel car le rôle de l'enquêteur est de susciter la parole chez l'apprenant ; la finalité n'est pas les mécanismes de l'interaction mais bien la production orale spontanée. En outre, nous

rappelons que nos objectifs sont essentiellement discursifs. Dans le cadre d'analyses syntaxiques, morphologiques ou lexicales, la lisibilité qu'offre une représentation verticale est importante ; en outre, CLAN étant destiné à ce type d'analyse, les outils à cet effet comme les annotateurs morphosyntaxiques sont plus développés que ceux possibles sur des logiciels comme PRAAT.

3.2 Protocole de transcription

La transcription des données orales d'apprenants en FLE constitue le cœur de cette étude en raison de ce qu'elle comporte de difficultés inhérentes à la transcription de l'oral de manière générale, difficultés auxquelles viennent s'ajouter les complications liées au fait qu'il s'agit ici d'interlangue. Si certaines constantes peuvent se retrouver dans l'interlangue et que donc des solutions peuvent être proposées pour certaines difficultés de transcription récurrentes, certains phénomènes ont été traités au cas par cas et les solutions proposées ne constituent en aucun cas la seule alternative possible. L'interlangue est par définition un système en construction où les phénomènes peuvent être ponctuels ou systématiques, où les barrières entre les fautes et les erreurs sont poreuses et où certains faits oraux sont tels que leur représentation graphique ne peut correspondre aux normes du français. Comme nous allons le voir, les déviations qui posent problème au niveau de la transcription peuvent être de nature syntaxique, morphologique ou lexicale. Nous rappelons qu'il est admis depuis longtemps que la transcription est en soi une analyse des données, « a selective proces reflecting theoretical goals and definitions » (Ochs, 1979 : 44) et qu'un corpus est par conséquent une construction. Cette construction doit néanmoins éviter d'être définitive ; puisque nous considérons la transcription comme une analyse des données, elle se doit d'être réfutable.

Nous nous sommes imposé un premier principe : nous n'envisageons pas la représentation exhaustive du signal sonore, tel que le fait par exemple le protocole du corpus Léonard :

> Ce qui importe avant tout dans le cadre du projet Colaje, c'est d'avoir une transcription aussi complète que possible. Cela implique que chacun essaie de transcrire *toutes* les informations qu'il observe et entend. (Bourdoux *et al.*, 2011 : 4)

Tel que le précisent les auteurs, il ne s'agit pas d'un principe immuable quel que soit le projet, mais d'un choix théorique qui implique des analyses et des approches différentes. Nous

rappelons que dans notre présentation théorique sur le processus de transcription[297], nous avions défendu l'idée d'une transcription incitant à l'écoute[298]. Nos transcripts visent donc essentiellement à offrir au chercheur la possibilité de retrouver, dans un corpus de plusieurs dizaines de milliers de mots, les phénomènes linguistiques auxquels il serait intéressé, et cela grâce aux manipulations logicielles disponibles ; pour le chercheur qui approcherait le corpus sans objectifs d'analyse *a priori*, les transcripts lui permettent d'aborder les données de manière lisible. Dans les deux cas, il est nécessaire de réécouter les segments à analyser, et l'alignement du son sur les transcripts vient compenser le « minimalisme » que nous défendons.

Dans la partie théorique sur la transcription, nous avions en premier lieu évoqué la question de la perception des données, puis celle des conventions de transcription :

1) le type de transcription ;
2) la segmentation des données ;
3) le choix des données à représenter.

Dans cette section, nous aborderons ces problématiques d'un point de vue pratique en présentant les choix opérés, qu'ils soient d'ordre technique et liés à CLAN ou purement linguistiques : ces deux types sont étroitement imbriqués et comme le remarquent Bourdoux *et al.* (2011 : 3), « les choix techniques sont des mises en application des choix scientifiques et c'est le va et vient entre technique et recherche scientifique qui doit être constamment en cours ». Tout au long de notre démarche, nous en profiterons pour illustrer les principales difficultés suscitées par la transcription de l'interlangue.

3.2.1 Perception des données

Comme le notent Bilger *et al.* (1997 : 57), « la compréhension de certaines séquences sonores n'est pas aussi univoque qu'on pourrait le penser, et peut amener le transcripteur à hésiter entre plusieurs interprétations concurrentes ». La fiabilité des perceptions est encore moindre lorsqu'il s'agit d'interlangue : Zechner (2009, section 4.4) rapporte en effet que le taux de désaccord entre transcripteurs de données natives ne dépasse jamais les 5%, tandis que le taux

[297] Cf. 2.7.2.

[298] Cf. 2.7.2.3.

de désaccord entre transcripteurs de données d'apprenants ne descend jamais en dessous des 10%, atteignant dans certains cas 34%, pour une moyenne globale de 15 à 20%. Il ne s'agit pas ici de difficultés de perception liées à la mauvaise qualité sonore des enregistrements, mais bien de divergences de perception donnant lieu à plusieurs transcriptions possibles. Pour résoudre ce type de problème, l'équipe du GARS avait proposé la multi-transcription (Bilger *et al.*, 1997), soit la possibilité de proposer plusieurs versions graphiques d'un même passage sonore. Dans notre démarche, nous n'avons pas adopté cette solution pour plusieurs raisons. Premièrement, quel que soit le nombre de transcripts proposé pour un seul et même passage sonore, il ne pourra prétendre recouvrir l'ensemble des perceptions possibles, et la multi-transcription s'apparente alors à une multi-analyse. Bien que nous ne réfutions aucunement le caractère analytique d'une transcription unique et que nous ne prétendions pas proposer une quelconque version graphique neutre, nous n'avons pas retenu la possibilité d'une multi-transcription car nous rappelons que notre objectif n'est pas de fournir les différentes interprétations possibles, mais la possibilité de recherche et de fouille de corpus qui permettra au chercheur, via la réécoute, de construire sa propre interprétation. Il nous faut ici d'ailleurs faire remarquer que lors de la vérification de la première transcription des étudiants, notre propre perception a parfois été influencée par le transcript du segment sonore que nous avions écouté au préalable. Cela confirme d'une part la divergence des perceptions, mais également le caractère analytique de toute version graphique d'un signal sonore.

En ce qui concerne les fragments inintelligibles que nous évoquions, ils sont parfois dus à une piètre qualité sonore, mais la plupart d'entre eux sont des phonèmes compréhensibles mais intranscriptibles en morphèmes. Pallaud (2002 : 289), montre dans sa vérification des transcripts de douze corpus que certains éléments restent non élucidés. Nous discutons de ces fragments ici car ce qui n'a pu être élucidé pourrait l'être par un autre transcripteur. Lors de la vérification des transcripts, nous avons pu proposer une version pour des fragments qui avaient été marqués comme non intelligibles par les étudiants, et pu percevoir (ou croire percevoir ?) d'autres fragments après la lecture des transcripts. Conformément au protocole technique de CLAN, ces fragments ont été signalés par xxx, qu'il s'agisse de mots isolés ou de passages plus longs.

3.2.2 Type de transcription

La première question qui se pose est le type de transcription à adopter dans le cas de l'interlangue. Comme nous l'avons vu[299], EAGLES recommande la transcription orthographique pour les études discursives et suggère de réserver la transcription phonétique pour les études phoniques. Les recommandations ne concernent cependant pas l'interlangue et ses spécificités qui peuvent suggérer la nécessité d'une transcription phonétique quelle que soit la finalité du corpus. Les auteurs qui ont traité de la question sont rares, et la question de la transcription de l'interlangue n'a jamais constitué une problématique à part entière. Ainsi Blanche-Benveniste évoque-t-elle rapidement le sujet en ces termes :

> Le parler des étrangers maîtrisant mal le français impose d'autres contraintes. M.A. Mota, étudiant le parler d'émigrés portugais en France, avait montré que la transcription phonétique était la seule solution correcte :
>
> > [mwa vulɛ pa partir – bõ avã bulɛ partir – wi ɔ kumãse - kw~ẽ kumãnsej isi plœre - vulɛ pa rɛʃte frãs - mẽtnã - sẽ o kõntrɛr – mi plɛ plus mwa du kɔ mõ mari]
>
> Une transcription orthographique aurait risqué de donner une version exagérément « optimiste », en proposant par exemple des imparfaits graphiquement très au point, comme *voulais, pleurais* :
>
> > Moi voulais pas partir. Bon, avant, voulais partir, oui, au commencer (commencé). Quand commençais (commencer, commencé) ici, pleurais. Voulais pas rester en France. Maintenant, c'est au contraire. Me plaît plus moi de que au mon mari.
>
> ou, au contraire, une version exagérément « pessimiste », dans laquelle les verbes terminés par [ɛ] ou [e] auraient tous des finales d'infinitifs :
>
> > Moi vouler pas partir. Bon, avant bouler partir, oui, au commencer. Quand commencer ici, pleurer. Vouler pas rester. (Blanche-Benveniste, 2010 : 38)

Il ne nous semble pas que la transcription phonétique soit « la seule correcte »[300] car ces raisons avancées en défaveur d'une transcription orthographique peuvent être discutées. Dans le cas de la transcription de [vulɛ] ou de [plœre] par exemple, Blanche-Benveniste souligne le risque de transcrire *voulais* ou *pleurais*, car ces transcriptions seraient « graphiquement très au

[299] Cf. 2.7.2.2.

[300] L'étude sur laquelle se base Blanche-Benveniste, celle de M.A. Mota, est une thèse soutenue en 1978 à l'Université de Lisbonne. Blanche-Benveniste n'indique pas les références ou le titre de cette thèse et nous n'avons pas pu la consulter.

point ». Il s'agit d'un faux problème car selon les termes de l'auteure elle-même, la question n'est que graphique dans ce cas-là, alors que l'objet de la recherche est l'oral. Or dans le cas d'un apprenant s'exprimant à l'oral, il suffit que l'apprenant marque oralement le temps passé, et peu importe sa graphie. Certes, le fait que la désinence verbale de l'imparfait soit homophonique au morphème infinitif peut évoquer un doute sur la compétence réelle de l'apprenant sur la formation de l'imparfait, mais il s'agit alors d'une question qui ne relève pas de la transcription, dont le rôle se limite à la représentation du signal sonore. Dans l'exemple, l'apprenante a recours à l'indicateur temporel *avant*, ce qui indique qu'elle s'exprime au passé. Le fait que *voulais*, *voulé* ou même *vouler* soient homophones n'est pas pertinent dans le choix qui devrait être fait dans ce cas-là, qui serait de notre point de vue *voulais*. Le problème sera ensuite de traiter l'absence du pronom sujet et la lacune de l'apprenante est donc d'ordre syntaxique. Nous avons relevé ce type d'occurrence dans notre corpus :

> Ex 18 : ça était très petit mais la chauffage ne **marchait** pas très bien.[301]

Nous pourrions nous interroger si l'apprenante a ici véritablement opté pour l'imparfait en raison de la valeur descriptive de son aspect et si elle n'aurait pas écrit *marcher* ou *marché*. Comme nous le disions, cette question ne nous semble pas légitime en termes d'évaluation de la langue orale qui ne permet pas de trancher dans le cas des homophones[302]. Il en va de même pour les exemples suivants :

> Ex 19 : j' ai habité à [x 2] Brighton hum dans le sud de l' Angleterre hum pendant huit ans et après ma famille et moi hum &co **commençé** de habite à Guernesey.[303]

> Ex 20 : je **né** au Mexique.[304]

> Ex 21 : et [x 2] en+fait je **resté** en France xxx un an oui depuis octobre deux mille neuf.[305]

[301] eng_li_wh_89_f_11.

[302] D'autant plus que nous pouvons supposer qu'un nombre important de natifs commettraient eux-mêmes des fautes d'orthographe similaires.

[303] eng_be_la_87_h_11.

[304] spa_al_ga_79_h_12.

[305] zho_zh_bo_84_f_11.

Dans (19), l'apprenant montre qu'il maîtrise l'utilisation du passé composé avec *j'ai habité* et que les faits évoqués le sont au passé en raison de l'indicateur temporel *pendant huit ans* ; le problème n'est donc pas la graphie de *commencé* mais la syntaxe en raison de l'omission de l'auxiliaire, comme c'est le cas pour (20) et (21). Dans les exemples suivants :

> Ex 22 : euh on a passé trois jours et on est **amusés** beaucoup[306].

L'erreur est ici l'omission du pronom complément.

> Ex 23 : hum même ma sœur Nesrine ma sœur Nesrine on a elle a fait le marié et **elle est voyagé** elle est partie chez son mari à Émirats aux Émirats.[307]

> Ex 24 : et donc pendant mon &enf mon enfance **j' ai resté** un+peu distant avec la musique.[308]

L'erreur est ici le choix erroné de l'auxiliaire.

> Ex 25 : et à l' école **commencé** la musique après le bac mais finalement **réussi** la [x 2] lycée de musique et la lycée d' informatique aussi <tous les deux> [x 2].[309]

L'erreur est ici l'omission du pronom personnel sujet ainsi que de l'auxiliaire. Dans cette série d'exemples, la transcription phonétique préconisée par Blanche-Benveniste mènerait au même constat, et la transcription orthographique que nous proposons, bien que « graphiquement très au point », n'empêche pas l'identification des lacunes des apprenants, et ne constitue donc pas ce que semble redouter – à raison – Blanche-Benveniste, à savoir une correction des productions orales des apprenants.

Toutefois, nous rappelons que tout choix de représentation ne peut être systématique à partir d'un ou de plusieurs exemples en raison de la diversité des éléments contextuels. Si nous avons choisi de graphiquement marquer le passé dans les exemples cités, nous avons dans d'autres cas opté pour le morphème infinitif –er, soit en raison de l'absence d'indicateurs temporels, soit parce que nous avons jugé qu'il pouvait s'agir d'un phénomène rare chez des apprenants avancés mais bien présent, à savoir celui du recours à la forme infinitivale à la place d'un verbe fini (Bartning, 1997), comme dans les exemples suivants :

[306] ara_ka_be_81_h_09.

[307] ara_ka_be_81_h_09.

[308] spa_en_ro_92_h_13.

[309] spa_al_ga_79_h_12.

> Ex 26 : et elle m' a dit oh tu peux reste ici et moi j' ai une copine euh qui **louer** une chambre parfois je peux appeler Agnès euh et demander euh si elle **louer** une chambre maintenant euh et maintenant.[310]

> Ex 27 : euh je j' **aider** avec les devoirs je [x 2] fais le goûter après école.[311]

Les exemples présentés par Blanche-Benveniste et ceux que nous venons d'examiner concernent essentiellement le marquage de la morphologie verbale du passé, d'autres concernent le marquage discontinu de manière générale tel que nous le détaillerons plus tard. Il ne s'agit pour l'instant que de constater que l'orthographe standard est apte à représenter les erreurs des apprenants sans dénaturer, ou « améliorer » leur interlangue, au moins en ce qui concerne les objections de Blanche-Benveniste concernant les morphèmes du passé.

Mais le choix d'une transcription en orthographe standard soulève immédiatement la question des trucages orthographiques et celle des disfluences.

3.2.2.1 Trucages orthographiques

Nous rappelons avoir discuté de différentes positions théoriques quant aux trucages orthographiques ou aménagements graphiques[312]. En ce qui nous concerne, nous n'y avons pas eu recours et toutes les transcriptions suivent l'orthographe standard, sauf en ce qui concerne les formes néologiques qui sont l'une des spécificités de l'interlangue tel que nous le verrons. Mais les phénomènes de la langue parlée habituels n'ont pas fait l'objet d'une transcription spécifique en raison des problèmes que nous avions abordés durant le second chapitre : les trucages orthographiques alourdissent la tâche du transcripteur sans qu'il n'y ait d'impact qualitatif positif sur les données. Nous considérons même que l'impact est négatif car les trucages diminuent la lisibilité du corpus et compliquent les recherches et l'analyse du corpus. Voici une liste d'exemples où il aurait pu y avoir des trucages orthographiques :

> Ex 28 : ah oui bien+sûr oui j' aimerais travailler ici me développer professionnellement et **je sais pas** réaliser quelques activités.[313]

[310] eng_la_sk_86_f_10.

[311] eng_ca_re_83_f_11.

[312] Cf. 2.7.2.2.

[313] spa_ar_ca_73_f_10.

> Ex 29 : et **je suis** le dernier maintenant bah ça veut dire **je suis** le dernier au+niveau+de la famille pour se marier et je suis le dernier parmi mes amis.[314]

Il s'agit ici d'un assourdissement du phonème [ʒ] en [ʃ] et d'une chute du schwa. Les termes en gras ont été prononcés [ʃepa] et [ʃɥi], mais ont été transcrits selon l'orthographe habituelle.

> Ex 30 : euh un **petit** jour en septembre j' ai rencontré mes amis.[315]

> Ex 31 : je crois que la société française est plus compliquée que la Chine peut+être c'est à+cause+de l' immigration je crois puisque ici **il+y+a** beaucoup d' étrangers.[316]

Il s'agit ici d'amuïssements : *petit* a été prononcé [ti], et *il y a* a été prononcé [ja]. Le signe + sera explicité plus bas.

> Ex 32 : bon ça dépend aussi parce+que il+y+a bon **enfin** on mange aussi des choses bien lourds.[317]

Il s'agit ici de l'aphérèse du terme *enfin*, qui a été prononcé [fɛ̃].

Le traitement diffère en ce qui concerne les phénomènes bénéficiant d'un assez large consensus typographique, ou qui sont présents dans certains dictionnaires, notamment en ce qui concerne les apocopes :

> Ex 33 : parce+que au **labo** à la **fac** comme on est toujours avec les mêmes personnes on parle toujours bah souvent euh on parle sur les mêmes sujets on discute sur les mêmes sujets on utilise toujours les mêmes vocabulaires et+cetera.[318]

> Ex 34 : bah rien bah je joue du **foot** des fois je [x 2] vais &à au **gym** des fois.[319]

Enfin, pour conclure sur les aménagements graphiques, nous précisons que le *ne* de négation n'a pas été rajouté lors de son absence à l'oral.

3.2.2.2 Disfluences

[314] ara_ab_ma_80_h_10.

[315] ara_ka_be_81_f_09.

[316] zho_we_ch_88_f_10.

[317] spa_na_kl_87_f_12.

[318] ara_ab_ma_80_h_10.

[319] spa_en_ro_92_h_13.

Lorsque l'orthographe standard est le choix opéré pour un protocole de transcription, la seconde question qui se pose est celle des disfluences. Précisons d'emblée que d'un point de vue terminologique, il n'y a pas consensus : Candea (2000 : 9-12) recense les termes de « pauses non-silencieuses », « pauses pleines », « pauses sonores », « auto-réparations », « disfluences », « bribes, bafouillage, bégaiements d'hésitation », « phénomènes d'hésitation », « hésitation vocale », « marques du travail de formulation », « marques de recherche de formulation » et « gestion de la formulation » dans la littérature française ; « hesitation phenomena », « hesitation pauses », « repair phenomena », « speech disfluencies », « filled pauses » et « fillers » dans la littérature d'expression anglaise. Cette richesse terminologique démontre la polysémie du terme et le fait qu'il regroupe plusieurs sous-catégories. Nous adopterons pour notre part le terme de disfluences et considérons qu'il regroupe dans le cadre de notre protocole les phénomènes suivants : 1) les répétitions 2) les hésitations et les marqueurs de structuration de la conversation (d'après la terminologie d'Auchlin (1981)) 3) les amorces et faux départs.

1) Les répétitions

Si la répétition porte sur un mot, il est suivi de la séquence [x a], *a* étant le nombre de répétitions. Si la répétition porte sur une séquence, celle-ci est mise entre chevrons et est suivie de la séquence [x a]. Les hésitations entre deux occurrences répétées sont ignorées. Ainsi :

> Ex 35 : euh les [x 3] mondial euh des <ça célébrait à> [x 2] à &Alle Allemagne ouais c'est ça.[320]

Sans les balises de répétitions, l'énoncé serait : *euh les les les mondial euh des ça célébrait à ça célébrait à ça célébrait à à &Alle Allemagne ouais c'est ça.* Il va de soi qu'il ne s'agit pas d'un balisage automatique de tous les homographes, qu'ils soient homophones ou non :

> Ex 36 : c'est **ça ça** dépend les &reli régions mais en+général oui c'est plutôt désertique.[321]

> Ex 37 : parce+que c'est une différence en+effet c'est pas comme **nous nous** sommes habitués à avoir l' argent même si c'est pas beaucoup on a.[322]

[320] spa_en_ro_92_h_13.

[321] ara_ra_ki_79_h_09.

Dans l'exemple (36), nous avons jugé d'après les marqueurs prosodiques que le premier *ça* était rattaché à *c'est*, et le second à *dépend*. C'est également les marqueurs prosodiques qui nous ont permis de distinguer, dans l'exemple (37), le premier *nous* en tant que pronom sous forme disjointe, du second *nous* en tant que pronom personnel sujet ; concernant ce même exemple et à un autre niveau, ce sont toujours les marqueurs prosodiques qui nous font avancer qu'il ne s'agit pas ici du verbe *s'habituer* au passé composé, ce qui démontre l'importance de la réécoute puisque nous avons fait le choix de ne pas marquer la prosodie, qui est un processus d'annotation différent de la transcription. Toujours en ce qui concerne les répétitions, considérons l'exemple suivant :

> Ex 38 : ah oui oui <avec un> [x 2] un Français xxx je crois que hum des hommes ici sont **très très** beaux très beaux.[323]

Il ne s'agit pas ici de syntaxe, puisque les deux occurrences de *très* appartiennent à la même catégorie fonctionnelle (adverbe) et portent toutes les deux sur *beaux* ; la répétition est ici une épanalepse, soit une figure de style volontaire dont le balisage eût été erronée car elle ne rentre pas dans la catégorie des disfluences.

2) Les hésitations et les marqueurs de structuration de la conversation

Conformément aux recommandations EAGLES qui préconisent l'utilisation d'au moins deux types de syllabes « fillers », une voyelle et une nasale[324], toutes les hésitations du corpus ont été transcrites soit *euh*, soit *hum*, quelles que soient leurs durées, leurs significations ou les pauses silencieuses qui peuvent les séparer ; nous considérons que la transcription se doit uniquement de permettre la localisation de ces phénomènes pour une potentielle analyse et non pas les analyser en proposant des transcripts reflétant leur durée ou une quelconque interprétation sémantique.

En ce qui concerne les marqueurs de structuration de la conversation, nous en avons distingué quatre types :

[322] eng_jo_ga_91_h_12.

[323] spa_ar_ca_73_f_10.

[324] Ces recommandations ont été traduites et citées par Delais-Roussarie (2002 : 30).

- Les marqueurs de structuration comme *voilà*, *quoi*, *en fait*, *bon* etc. ; qu'ils soient simples ou composés, ce premier type de marqueur ne pose pas de problème particulier de transcription ;

- Les marqueurs de structuration exprimés dans la langue maternelle de l'apprenant ont été balisés comme l'ensemble des termes non français, balises sur lesquelles nous reviendrons plus en détail :

> Ex 39 : oui c'est pour ça parce+que **ya'ni@s:ara** tu sais la matière euh le travail au banque euh.[325]

- Certaines onomatopées et interjections : en ce qui concerne les premières, elles ont été retranscrites selon le dictionnaire. Quant aux secondes, les « interjections émotives » (Riegel *et al.*, 2009 : 773), nous les avons transcrites selon l'orthographe de la *Grammaire méthodique du français*, mais avons tenu à unifier les orthographes. Ainsi l'ensemble des [o], [a] et [e] ont été transcrits *oh*, *ah* et *eh*. Seuls [ba] et [bɛ̃] ont bénéficié des deux orthographes *bah* et *ben*, car nous avons considéré ces deux interjections différentes l'une de l'autre :

> Ex 40 : oui j' ai travaillé deux fois **ben** une fois c'était en à en Belgique et le deuxième fois c'était **bah** à Rennes.[326]

3) Amorces et faux départs

Les amorces et les faux départs ont été marqués par la balise &, conformément au protocole technique de CLAN :

> Ex 41 : mais j' étais en Bretagne oui parce+que mon père il est **&fr français** il est breton donc j' ai j' étais près de Rennes mais pas Rennes exactement donc non.[327] (amorce)

> Ex 42 : euh chambre il+y+a euh cinq chambres et deux toilettes et une cuisine et **un &sal un balcon** et un petit [x 3] salle pour habiller.[328] (faux départ)

> Ex 43 : et **&j** quand je le finis je voudrais bien.[329] (amorce incomplétée)

[325] ara_ge_fr_80_f_08.

[326] spa_pe_gu_93_h_13.

[327] eng_ol_la_91_f_12

[328] zho_zh_bo_84_f_11

Le balisage va contre notre volonté de faciliter, autant que faire se peut, la lecture des transcripts et de ne pas les alourdir, et nous aurions pu soit ignorer le phénomène, soit le transcrire sans balises. Nous avons adopté la première option dans un premier temps, mais avons considéré qu'il n'était pas possible d'ignorer les amorces en raison de leur récurrence : CLAN comptabilise en effet 588 amorces et faux départs différents pour une fréquence totale de 1437 occurrences, soit une moyenne de 30 amorces par apprenant, les dix amorces les plus fréquentes ainsi que leurs fréquences étant : 94 &j, 83 &s, 54 &c, 37 &l, 36 &f, 30 &a, 24 &d, 24 &p, 23 &m, 17 &i. La moyenne de 13 amorces et faux départs tous les 1000 mots chez les apprenants de CIL-FLE est trois fois supérieure à la moyenne des natifs si l'on considère les chiffres de Pallaud & Henry (2003) qui ont calculé une fréquence de 4 amorces tous les 1000 mots sur un corpus d'environ 45000 mots.

La seconde option n'a pas été adoptée pour des raisons techniques. Transcrire *sal* au lieu de *&sal* pose deux problèmes à CLAN : d'une part il considérera *sal* comme un mot, et cela entraînera des difficultés pour le calcul de la fréquence, pour la reconnaissance lexicale ainsi que pour l'annotation morphosyntaxique ; d'autre part, l'absence de balise n'aurait pas permis la localisation du phénomène.

Il n'est pas toujours aisé d'identifier les amorces et les faux départs. Considérons les exemples suivants :

> Ex 44 : hum et <j' ai> [x 2] pris **l' &é** euh l' éponge.[330]
>
> Ex 45 : et je me demandais s' ils **est étaient** mexicains ou pas mais bon heureusement ils ont commencé à parler.[331]
>
> Ex 46 : et ces problèmes c'est [x 2] sont [x 2] surtout **qui qu' ils** [x 2] n' accompagnent pas l' enfant enfin et dans un cours comme si c'était des [x 2] personnes adultes quoi.[332]

Dans l'exemple (44), nous avons considéré que [le] était l'article défini féminin *la* suivi de l'amorce de *éponge*, avec l'élision de la voyelle finale de l'article. Mais [le] aurait pu être considéré comme l'article pluriel *les* et constituer alors un faux départ ; la transcription eût

[329] eng_al_jo_86_f_10.

[330] eng_la_sk_86_f_10.

[331] spa_ar_ca_73_f_10.

[332] spa_pe_gu_93_h_13.

alors été : *j'ai pris les l'éponge*. Dans les exemples (45) et (46), nous avons à l'inverse considéré qu'il n'y avait pas d'amorce, alors que *est étaient* aurait pu être transcrit *&é était*, et *qui qu'ils* transcrit *qu' &i qu'ils* ; l'homophonie entre ici en jeu et nous avons conscience que les choix finalement adoptés sont interprétatifs et réfutables.

Les exemples suivants suscitent également plusieurs interprétations qui ne concernent pas les homonymes :

> Ex 47 : j' arrive à savoir les **signifis** voilà donc voilà c'est ça je dois améliorer mon [x 2] français à l' écrit surtout.[333]

> Ex 48 : ouais c'est obligatoire parce+que pour &l pour &m pour mes parents euh ça fait **&long** euh quand on quand [x 2] j' ai quand je suis entrée le quand je suis déjà entrée le [x 3] collège.[334]

Dans l'exemple (47), il convient de s'interroger sur *signifis* : s'agit-il de l'amorce de *significations* ? La transcription aurait alors été *&signifi*, mais les marqueurs prosodiques nous ont fait juger qu'il ne s'agissait pas d'une amorce mais d'un néologisme pour *signification*. Dans l'exemple (48), nous aurions pu considérer qu'il s'agit ici d'une utilisation adverbiale de *long*, et la balise de l'amorce aurait alors été erronée ; là encore, ce sont les marqueurs prosodiques qui indiquent une amorce de *longtemps*. Enfin, dans les exemples suivants :

> Ex 49 : oui j' aime beaucoup et euh le **&ciné** &ci le film de français aussi ouais.[335]

> Ex 50 : et comme+ça si c'est accepté c'est bien sinon il faut faire des **&manip** des **&manipula** manipulations **&supplément** &su supplémentaires pour compléter le travail.[336]

Dans l'exemple (49), *ciné* aurait pu être considéré comme une apocope, ainsi que *manip* dans l'exemple (50). Toutefois les marqueurs prosodiques suggèrent que ce n'est pas le cas. Ce jugement est conforté dans l'exemple (50) puisque *manip* est suivi de *manipula*, qui n'est pas une apocope mais assurément une amorce. Ces observations ne concernent pas uniquement les apocopes : toujours dans l'exemple (50), *supplément* n'est pas le nom commun mais

[333] spa_jo_go_62_h_12.

[334] zho_fa_li_86_f_13.

[335] zho_zh_bo_84_f_11.

[336] ara_ab_ma_80_h_10.

l'amorce de l'adjectif *supplémentaire*, comme en témoigne plusieurs indices qui sont la prosodie et le fait que l'amorce soit suivie d'une autre amorce puis de l'adjectif.

3.2.3 Segmentation des données

Le premier principe de segmentation des données a évidemment été la prise en compte des tours de parole. Nous avons attribué à tous les enquêteurs l'étiquette ENQ et le rôle « Investigator », parmi ceux proposés par CLAN ; les apprenants ont tous bénéficié de l'étiquette FLE et du rôle « Student ». Ainsi, chaque fichier de transcription est un dialogue entre ENQ et FLE. Chaque ligne de transcription, ou tire[337], est alignée au segment sonore qui lui correspond par une « bullet » à la fin de la tire, qui délimite en millisecondes le début et la fin du segment sonore afin de permettre la réécoute d'une tire. Nous avons aligné le son aux transcripts manuellement, en tentant de suivre les points suivants :

- Le tour de parole de FLE peut être découpé en plusieurs tires, car FLE peut prendre la parole pour plusieurs minutes en continu ;

- aucune marque de prosodie n'est utilisée pour la segmentation, hormis le point d'interrogation quand l'interrogation est explicite. Le point final de chaque tire indique à CLAN la fin de la tire et ne comporte aucune valeur prosodique ;

- quand il y a chevauchement de la parole, deux tires de transcription sont affectées au même passage sonore, l'une avec l'étiquette ENQ, l'autre avec l'étiquette FLE ;

- en ce qui concerne la segmentation du flux sonore lors du passage d'une tire à l'autre, nous avons respecté, dans la mesure du possible, les unités syntaxique et sémantique. (Cf. Figure 4 : Représentation spatiale des transcripts sous forme verticale pour un exemple de la segmentation de la prise de parole de FLE). Nous n'avons pas tenu à effectuer une segmentation en unités syntaxiques minimales, ni en énoncés uniques ; la délimitation de ces phénomènes en interlangue orale étant un sujet d'analyse à part entière.

[337] Une « tire » est une ligne de transcription ou d'annotation. Dans les corpus oraux en général, un segment peut avoir plusieurs tires : une tire de transcription et une ou plusieurs autres d'annotation ou de commentaires non soumis à protocole.

3.2.4 Choix des données à représenter et manière de les représenter

Selon les recommandations de Delais-Roussarie (2002 : 9-10), nous indiquerons ici les phénomènes que nous avons ignorés, ceux que nous avons choisi de représenter ainsi que la façon de les représenter typographiquement. En premier lieu, nous avons ignoré l'ensemble des événements non linguistiques « qu'ils soient sonores ou non (aboiements d'un chien, toux, rires, hochements de tête, etc.), communicatifs ou non (rires, pleurs, éternuements, etc.) » (Delais-Roussarie, 2009 : 9). En ce qui concerne les événements linguistiques, nous avons ignoré la grande majorité des « backchannels » ou, selon la terminologie de Mondada (2004 : 4), les continuateurs, qui sont définis par Bertrand *et al.* comme suit :

> Le terme de backchannel (…) est employé de manière générique pour référer à l'ensemble des signaux verbaux, vocaux et gestuels, émis par l'interlocuteur d'un dialogue pour montrer son écoute, sa compréhension, son accord, etc. au discours produit. (Bertrand *et al.*, 2009 : 3)

Ainsi les *oui*, *d'accord*, *bien sûr* etc., ainsi que les acquiescements sonores dont le rôle est celui défini par Bertrand *et al.* n'ont pas été segmentés, localisés ou transcrits, principalement parce que la majorité de ces phénomènes ont été produits par l'enquêteur et que nous ne sommes pas concernés par l'interaction. Enfin, nous avons également décidé de ne pas marquer les pauses : la signalisation des pauses, avec indication ou non de leurs durées, alourdit le texte alors que des outils comme ELAN permettent de les repérer et de les quantifier automatiquement. Enfin, les liaisons n'ont pas été marquées, sauf les liaisons erronées : il ne s'agit pas ici d'un repérage d'erreur mais la simple manifestation graphique de la réalisation d'un phonème, comme dans l'exemple suivant :

> Ex 51 : un [x 2] vieux &n homme rester ensemble et.[338]

Cette manière de faire reste lacunaire en ce qui concerne la liaison puisqu'elle ne marque pas l'absence de liaison. En outre, dans l'exemple suivant :

> Ex 52 : ça c'est important parce+que à l' étranger personne peut nous aider peut nous aider on peut juste vivre par nous-mêmes.[339]

[338] zho_wa_ha_88_f_10.

[339] zho_yi_wa_93_f_13.

Il pourrait sembler nécessaire de marquer la répétition <*peut nous aider*> *[x 2]*, mais nous ne l'avons pas fait car le segment la liaison entre *nous* et *aider* a été réalisée dans un cas mais pas dans l'autre.

Considérons maintenant certains phénomènes et la manière dont ils ont été marqués ou balisés. Nos choix concernent les faits à noter ; les balises sont celles imposées par CLAN.

1) Les lettres dénommées en tant que telles

Chaque lettre est suivie de la balise @l :

> Ex 53 : c'est pourquoi je parle je prononce <un+peu> [x 2] plus bien que l' autre syrienne parce+que il+y+a beaucoup de lettres français trop difficiles à [x 2] parler à [x 2] prononcer comme **p@l v@l r@l e@l u@l**.[340]

2) Les majuscules

Aucune majuscule n'est utilisée en début d'énoncé, seules sont utilisées les majuscules pour les noms propres.

3) Les acronymes et les signes épelés

Chaque lettre est écrite en majuscule puis suivie du symbole _, excepté la dernière. Si le sigle est lu, comme par exemple UNESCO, le mot est écrit en toutes lettres et avec une majuscule :

> Ex 54 : **Dutfa** diplôme universitaire de traduction français arabe et arabe français et euh c'était uniquement de la traduction ou vous aviez aussi des cours (…) à Byblos à la banque et au **C_L_F_C**.[341]

4) Les titres et les nombres

Les titres et les nombres sont transcrits en toutes lettres.

> Ex 55 : *ENQ: tu as terminé quand ?
> *FLE: en **deux mille deux**.
> *ENQ: qu'est+ce+que tu as fait après ?
> *FLE: après je suis <j' ai> [x 2] travaillé euh **six** mois ici en Syrie dans une journal de publicité qui s' appelle Alwasit Alwassila pardon après j' ai je suis allé à Beyrouth j' ai

[340] ara_ma_sa_84_f_10.

[341] ara_di_ma_83_f_09.

travaillé **deux** ans et demi comme comptable dans une journal quotidien libanais qui s' appelle Albalad.[342]

5) Apostrophes, mots composés, traits d'union et locutions

Afin de procéder aux calculs de fréquence, aux analyses lexicales ou à l'annotation morphosyntaxique, tout logiciel doit se baser sur une base de données de référence où toute occurrence est répertoriée et catégorisée ; cette base est appelée « lexique ». Le premier lexique de CHILDES fut anglais, mais depuis que le projet connaît une envergure internationale, des équipes de différents pays ont constitué d'autres lexiques dans d'autres langues. C'est ainsi que Christophe Parisse a constitué un lexique français pour CLAN, composé d'entrées appelés « items »[343]. Le logiciel CLAN considère que toute suite de lettres précédée et suivie d'un espace constitue un « single lexical item », ou item. Morphosyntaxiquement, un item peut contenir plusieurs morphèmes, car la segmentation en morphèmes aurait conduit à des problèmes de lisibilité et s'apparenterait davantage à une analyse morphosyntaxique qu'à la transcription. De même, il n'est pas possible d'agglutiner les morphèmes grammaticaux en un seul item en raison du marquage discontinu. La transcription que nous proposons n'offre donc ni segmentation totale en morphèmes en items, ni agglutination systématique des morphèmes en items. En revanche, quand deux morphèmes sont liés par une apostrophe dans la norme, il est nécessaire pour le bon fonctionnement de CLAN de les séparer par un espace : ainsi *j'ai appris* est transcrit *j' ai appris*, avec un espace entre *j'* et *ai*, car il aurait sinon fallu que l'entrée *j'ai* soit insérée dans le lexique telle quelle, et qu'elle ne soit pas considérée comme l'item pronom personnel sujet *j'* suivi de l'item verbe auxiliaire *ai*, mais comme un seul et unique item qui aurait nécessité une étiquette morphosyntaxique. Quand l'apostrophe de la norme ne lie pas deux morphèmes, comme c'est le cas pour *aujourd'hui*, l'espace n'est pas inclus.

À l'inverse de la séparation des items par un espace, d'autres ont été liés avec le symbole +, et cela concerne les constructions considérées comme fonctionnant en « blocs », car leurs segments ne peuvent commuter indépendamment les uns des autres. Si les éléments de ces constructions sont déjà liés par une apostrophe dans la norme, l'apostrophe a été conservée

[342] ara_ra_ki_79_h_09.

[343] Le lexique est disponible en annexe électronique. Le tableau des différentes catégories morphosyntaxiques adoptées figure en annexe 6.

sans être remplacée par un + et l'espace n'a pas été inclus. Il s'agit de *c'est* et *il y a* en tant que présentatifs, transcrits *c'est* (pas d'espace inclus) et *il+y+a* (liaison par +), ainsi que de leurs variations, transcrite *c'était / il+y+avait / il+n'y+avait+pas* etc. Cela concerne également la tournure interrogative *est-ce que*, transcrite *est+ce+que*, et liée au pronom interrogatif quand celui-ci est présent : *qu'est+ce+que / qu'est+ce+qu' / qu'est+ce+qui*. De même, nous avons lié la plupart des locutions non verbales, qu'il s'agisse de locutions prépositives comme *autour+de*, conjonctives comme *parce+que* ou adverbiales comme *en+fait* ou *un+peu* :

> Ex 56 : mais moi je j' aime &en **en+fait** quand j' étais petit j' avais des problèmes avec ça **parce+que autour+de** moi tout+le+monde aimait la musique et ça me fait **un+peu** mal.[344]

L'exemple (56) montre également que nous avons lié les pronoms indéfinis comme *tout+le+monde* ou *quelque+chose* ; nous rappelons qu'un pronom indéfini comme *quelqu'un* est transcrit comme tel puisque l'apostrophe fait office de liaison pour CLAN.

L'utilisation du trait d'union n'est pas possible pour CLAN, les mots composés ont donc été transcrits avec le symbole + :

> Ex 57 : hum campagne pour moi c'est campagne ouais parce+que enfin c'est comment on dit avant [x 2] collège c'est &cam c'est la campagne avec mon &gran **grand+mère** et **grand+père** c'est comme+ça.[345]

En ce qui concerne les toponymes composés, ils ont été liés par le symbole _ , car il n'est pas possible de faire suivre le symbole + d'une majuscule :

> Ex 58 : à Rennes **Saint_Malo Saint_Michel** Dinan Dinard et Brest la **Côtes_d_Armor**.[346]

Si la plupart des items composés de CLAN, c'est-à-dire les segments liés par le symbole +, se trouvent dans le lexique, ce dernier est une base de données dynamique. Les choix des compilateurs du lexique ne sont donc pas imposés, et il est possible d'en inclure de nouveaux. Ainsi des items comme *ici+là* ou *il+est+là*, retenus car pertinents dans le domaine de

[344] spa_en_ro_92_h_13.

[345] zho_fa_li_86_f_13.

[346] spa_al_al_72_f_11.

l'acquisition du langage chez les enfants, n'ont pas été retenus pour CIL-FLE ; d'un autre côté, nous avons rajouté au lexique certaines constructions que nous avons rencontré lors de la transcription mais qui n'étaient pas présentes dans le lexique[347] :

a+t+il {[scat v:poss]}	au+début {[scat adv]}	au+fur+et+à+mesure {[scat adv]}
au+milieu {[scat adv:place]}	au+milieu+d' {[scat prep:art]} "au+milieu+de"	au+milieu+de {[scat prep:art]} "au+milieu+de"
au+milieu+des {[scat prep:art]} "au+milieu+de"	au+milieu+du {[scat prep:art]} "au+milieu+de"	au+niveau {[scat adv]}
au+niveau+d' {[scat prep:art]} "au+niveau+de"	au+niveau+de {[scat prep:art]} "au+niveau+de"	au+niveau+des {[scat prep:art]} "au+niveau+de"
au+niveau+du {[scat prep:art]} "au+niveau+de"	au+pair {[scat adv]}	autour+de {[scat prep]}
avant+qu' {[scat conj]} "avant+que"	avant+que {[scat conj]}	comme+ci+comme+ça {[scat adv]}
de+temps+en+temps {[scat adv]}	de+toute+façon {[scat adv]}	du+coup {[scat adv]}
en+ce+moment {[scat adv]}	en+face {[scat adv:place]}	en+face+d' {[scat prep:art]} "en+face+de"
en+face+de {[scat prep:art]} "en+face+de"	en+face+des {[scat prep:art]} "en+face+de"	en+face+du {[scat prep:art]} "en+face+de"
en+gros {[scat adv]}	en+tant+qu' {[scat pro:rel]}	il+n'y+a {[scat v:exist]}
il+n'y+a+pas {[scat v:exist]}	il+n'y+a+plus {[scat v:exist]}	il+n'y+a+rien {[scat v:exist]}
il+n'y+avait+pas {[scat v:exist]}	il+n'y+en+a+pas {[scat v:exist]}	il+y+aura {[scat v:exist]}
il+y+avait {[scat v:exist]} "il+y+a&IMPF"	moitié+moitié {[scat adv]}	mot+à+mot {[scat adv]}
n'est-ce+pas {[scat co]}	n'importe+quel {[scat adj]}	n'importe+quelle {[scat adj]}
n'importe+quelles {[scat adj]}	n'importe+quels {[scat adj]}	n'importe+où {[scat adv]}
par+ci+par+là {[scat adv]}	par+coeur {[scat adv]}	à+part {[scat adv]}
plus+ou+moins {[scat adv]}	près+de+la {[scat prep:art]}	par+hasard {[scat adv]}
sauf+qu' {[scat conj]} "sauf+que"	sauf+que {[scat conj]}	s'il+vous+plaît {[scat co]}
y+a+t+il {[scat v:exist]}	à+cause+de+la {[scat prep:art]} "à+cause+de"	soi+même {[scat pro]} "soi+même"
à+cause+du {[scat prep:art]} "à+cause+de"	à+ce+moment+là {[scat adv]}	à+cause+des {[scat prep:art]} "à+cause+de"

[347] L'ensemble des ajouts au lexique de CLAN se fait dans un fichier séparé ; ce fichier doit être envoyé à CHILDES afin qu'il soit étudié et que ses entrées soient ajoutées au lexique principal. Pour la liste complète des ajouts comprenant également les items simples et les items composés nominaux, cf. annexe 7.

à+la+fin {[scat adv]}	à+l'aise {[scat adv]}	à+l'inverse {[scat adv]}

Nous tenons à préciser que si cette démarche facilite l'annotation morphosyntaxique, ou permet le repérage des locutions dans l'interlangue, certaines occurrences posent difficulté et ne seront pas repérables à moins d'un autre niveau d'annotation. C'est le cas dans l'exemple suivant où la locution a été déconstruite :

> Ex 59 : mais <on n' est> [x 2] pas obligés de se rencontrer tous ensemble bien+sûr j' ai des amis **par+là+par+ci**.[348]

L'apprenant a ici inversé les segments de la locution, qui ont certes été liés mais nous avons catégorisé l'item par+là+par+ci comme néologisme, et non comme la locution adverbiale *par ci, par là*.

6) Mots en langue étrangère

Les apprenants ont eu recours à des termes qui ne sont pas français. En règle générale, ces termes sont suivis de la balise @s:ara si le terme est en arabe, @s:eng en anglais, @s:spa en espagnol et @s:zho en chinois. Plusieurs cas se présentent. Il peut s'agir de marqueurs de structuration du langage comme nous l'avons vu précédemment. Dans un autre cas, l'apprenant ignore le mot en français :

> Ex 60 : des [x 3] vêtements un+peu <comment on dit ça> [x 2] **me'techma@s:ara**.[349]

D'autre part, un terme peut ne pas avoir d'équivalent en français. Ceci concerne notamment les termes culturels tels les termes culinaires ou vestimentaires :

> Ex 61 : oui les plats euh xxx je sais pas les plats le peut+être euh ya'ni@s:ara les [x 2] plats anciens (…) **sej'at@s:ara** peut+être oui **sej'at@s:ara** il+y+a j' ai un ami français qui aime beaucoup les **sej'at@s:ara**.[350]

Certaines difficultés se sont posées. Nous avons considéré que la balise @s indique un terme étranger mais que cette balise ne devait être appliquée qu'au cas où l'apprenant l'utilise en tant que terme étranger, mais pas dans les cas où le terme est effectivement étranger, mais considéré comme français par l'apprenant. Afin de pouvoir opérer cette distinction entre

[348] ara_ab_ma_80_h_10.

[349] ara_ra_ki_79_h_09.

[350] ara_ma_ze_84_f_09.

terme étranger et ce qu'il convient de nommer « néologismes », des indicateurs sont venus nous aider lors de la transcription, comme l'autocorrection :

> Ex 62 : oh surtout euh la **gramática@s:spa français la grammaire &gr grammaire** française c'est [x 2] compliqué la façon d' écrire le français c'est compliqué.[351]

> Ex 63 : ah je fais du rollerblade@s:eng au parc et xxx toujours **stupid@s:eng stupide** et bête euh donc avec mes élèves j' essaye de faire autre chose.[352]

Dans le premier, la correction est « visible », alors que dans l'exemple (63), l'apprenante a prononcé la première occurrence [stjupid], et la seconde [stypid], et le balisage ne pose donc pas problème puisque l'autocorrection semble indiquer que l'apprenant a conscience que ces termes ne sont pas français. En revanche, dans les exemples suivants :

> Ex 64 : parfois euh mais pas récemment mais &j en+même+temps j' ai jamais été c'est quoi le mot euh capturé je n' étais pas euh oui &j oui waouh le **criminal**.[353]

> Ex 65 : avec ma **familia** peut+être je serai si ils sont là dans le même ville ça serait plus facile.[354]

Si *criminal* et *familia* sont respectivement des termes anglais et espagnol, nous avons jugé qu'ils ont été utilisés dans ces deux exemples comme termes français. Certes, il s'agit ici de deux formes phonétiquement proches de *criminel* et *famille*, et ne constituent pas des néologismes lexicaux, mais des déformations phonétiques. Cependant, l'utilisation de termes phonétiquement déformés par la norme de la langue maternelle est une question d'importance dans l'acquisition des langues et il ne nous a pas semblé judicieux de baliser ces termes comme nous l'avons fait pour les termes étrangers. Le point que nous venons de présenter soulève deux questions importantes, celle des relations entre phonétique et transcription de l'interlangue d'une part, et celle de la barrière entre simples emprunts à la langue maternelle et néologismes d'autre part. La section suivante traite de ces deux questions, également concernées par les difficultés de la transcription de l'interlangue.

[351] spa_jo_go_62_h_12.

[352] eng_al_jo_86_f_10.

[353] eng_jo_ga_91_h_12.

[354] spa_al_al_72_f_11.

3.2.5 Difficultés de transcription

Dans l'article de Bilger *et al.* (1997), intitulé « Transcription de l'oral et interprétation. Illustration de quelques difficultés », les auteurs proposent une typologie des problèmes rencontrés lors de la transcription de la langue de natifs. Le premier type de problèmes recensé est lié à la « qualité sonore » où, outre les facteurs physiques comme les bruits perturbateurs et les chevauchements, les auteurs parlent de « caractéristiques articulatoires des locuteurs » en citant pour exemple les parlers régionaux et les énoncés d'enfants. L'interlangue aurait parfaitement pu figurer aux côtés de ces exemples de « parlers », « correctement perçus » mais « incompréhensibles » (Bilger *et al.*, 1997 : 59), indépendamment d'une mauvaise qualité sonore. À ce stade, le transcripteur est donc en difficulté dans l'interprétation de l'énoncé, dans la construction de son sens afin de le restituer sous forme graphique. Nous rappelons que nous n'avons pas retenu l'option de la multi-transcription et que nous avons décidé de ne proposer qu'une seule version, quitte à ce que celle-ci soit réinterprétée lors de la réécoute. Pour conclure sur le protocole de transcription, nous consacrerons donc les sections suivantes à une catégorisation des phénomènes qui nous ont amené à hésiter sur la transcription de l'interlangue, qui possède ses propres spécificités, aux côtés des difficultés inhérentes au processus dans sa globalité.

3.2.5.1 Difficultés liées à la prononciation des apprenants

Dans leur relevé, Bilger *et al.* donnent pour exemple la problématique de l'interprétation de la signification, en citant le cas des noms propres ainsi que celui des termes techniques ignorés du transcripteur. En effet, considérons l'exemple suivant de notre corpus :

> Ex 66 : la dernière fois euh j' ai &re euh j' ai allé voir euh **L'amour_dure_trois_ans** c'est un c'est un &veu c'est un comédie c'est a été bien le film..[355]

Les étudiants de M1, dans la première version de transcription, avaient transcrit *la mort dure trois ans*. Ce n'est toutefois pas au niveau des références extralinguistiques que nous avons rencontré les problèmes les plus difficiles, mais bien dans les « caractéristiques articulatoires » des apprenants. Par exemple, en ce qui concerne les apprenants hispanophones, le pronom personnel *je* est très souvent prononcé [ʒe], voire [je]. Faut-il dans ce cas le transcrire *jé* et lui créer une étiquette morphosyntaxique propre ? Ou bien annoter

[355] spa_al_ga_79_h_12.

cette prononciation ? Nous avons décidé de ne faire ni l'un ni l'autre pour ce cas, conformément au double principe de lisibilité et de minimisation des annotations au sein des transcripts, et ce d'autant plus que les barrières entre [ʒe] et [ʒə] sont dans de nombreuses occurrences très peu nettes.

Ce choix de ne pas transcrire certaines particularités phonétiques de l'interlangue porte néanmoins en lui le danger d'une transcription « correctrice » qui normaliserait, en quelque sorte, des spécificités qui, justement, constituent l'un des intérêts d'un corpus d'interlangue. Nous avons donc hésité, dans de nombreux cas, entre la restitution phonétique de ce nous avons perçu et la représentation graphique normée de ce que nous avons compris. Dans les exemples suivants :

> Ex 67 : <j' ai> [x 2] eu un **accident**.[356]
>
> Ex 68 : de nouveau euh avec des avec un livre pour apprendre nouveau **plat** chaque jour.[357]
>
> Ex 69 : euh ma chambre c'est très petit c'est juste comme un petit **placard**.[358]

Les termes en gras ont été prononcés [aksidɑ̃t], [plat] et [plakard], mais ont été retranscrits selon l'orthographe standard, car nous avons jugé que les écarts n'étaient pas suffisants pour justifier l'ajout de termes néologiques au lexique. Nous avons conscience qu'ici, nous « gommons » un aspect important de l'interlangue, mais nous nous justifions toujours le même principe : nos transcripts ne prétendent pas figurer tous les dysfonctionnements de l'interlangue, mais offrir la possibilité de les repérer, et une analyse phonétique ou phonologique nécessitera de toute manière une écoute du corpus préliminaire. De plus, il est tout à fait possible d'annoter ce type de phénomène en particulier dans un deuxième temps, mais le faire au premier niveau de la transcription légitimerait toutes les autres annotations, car il faudrait alors annoter toutes les erreurs lexicales, morphosyntaxiques, syntaxiques etc. De manière générale, les exemples (67), (68) et (69) ne sont d'ailleurs pas représentatifs du traitement que nous avons la plupart du temps suivi ; dans les exemples suivants, les termes en gras ont été transcrits tels que nous les avons perçus, certains d'entre eux nécessitant « l'invention » de nouvelles graphies :

[356] eng_la_sk_86_f_10.

[357] eng_li_wh_89_f_11.

[358] eng_li_wh_89_f_11.

Ex 70 : elle a habité dans le **norte** par conséquence elle était hum elle avait froid beaucoup de temps.[359]

Ex 71 : oui c'est un atelier du Cirefe et c'est l'atelier de **chante** c'est [x 2] intéressant pour [x 2] faire quelque+chose avec à la &musi avec la musique.[360]

Ex 72 : euh <j' étudiais droit> [x 2] je suis avocate maintenant et aussi je fais une **espécialisation** je sais pas si ici c'est la même chose c'est droit pénal je suis pénaliste.[361]

Ex 73 : même si je suis capable de presque tout lire quand je lis un **test** en [x 3] français je [x 3] comprends il+y+a des mots que parfois ça m' arrive que je ne comprends pas mais dans le **contest**.

Ex 74 : une **vielle** dame.[362]

Ces deux traitements différents des deux dernières d'exemples démontrent qu'il est difficile de mettre en place des instructions de transcription systématiquement applicables et qui ne prendraient pas en compte les cas particuliers, ce qui confirme encore une fois la nature analytique du processus de transcription. Toujours en ce qui concerne la prononciation des apprenants, ces deux cas sont à mettre en parallèle :

Ex 75 : euh oui et je ne sais pas beaucoup de canard et j' aime beaucoup les **désertes**.[363]

Ex 76 ; il+y+a des plages superbes superbes vraiment il+y+a des **desserts**.[364]

Dans l'exemple (75), l'apprenante a prononcé [dezɛrt] pour *desserts* alors que l'apprenant de l'exemple (76) a prononcé [desɛr] pour *déserts*. La transcription a ici posé problème au niveau de l'orthographe : comment transcrire le second autrement que *desserts*, afin de le différencier au niveau graphique du mets alimentaire ? Il en va de même pour les exemples (71) et (73) où les transcripts de *chante* et *test* sont les homographes de formes existantes en français ; si cette question ne s'est pas posée pour des occurrences comme *norte* (70) où le transcript n'est l'homographe d'aucune forme en français, ni comme nous allons le voir, pour des formes

[359] eng_li_wh_89_f_11.

[360] spa_al_ga_79_h_12.

[361] spa_ga_ll_87_f_11.

[362] zho_wa_ha_88_f_10.

[363] eng_li_wh_89_f_11.

[364] spa_al_al_72_f_11.

comme *buvé*[365] ou *nerviocisme*[366] où nous n'avons pas trouvé difficulté à « inventer » une orthographe, nous n'avons pas pu nous résoudre à faire de même pour le cas des homographes que nous avons cités (par exemple, dans (76), transcrire *desser*, *descert*, *deser* ?).

Dans la série d'exemples qui suit, les choix ont été plus faciles, car il ne s'agit plus de l'ajout d'un phone final, mais de réelles déformations dont la représentation graphique ne pose pas problème au sens où ces occurrences ne sont pas confondues avec des termes existants en français :

> Ex 77 : dans ça et ça aussi il+y+a un [x 2] très grand problème parce+que après l' enfant il [x 2] se développe sans [x 2] aucune aide **solitare** xxx voilà pour dire comme+ça.[367]
>
> Ex 78 : et maintenant elle est **commerciant** <elle est> [x 2] elle reste aussi à la maison mais pour aider mon père.[368]
>
> Ex 79 : alors+que en Colombie l' histoire bah c'est [x 2] pas c'est pas très long comme en France parce+que [x 2] il+y+a eu les la &con les conquêtes la conquête et voilà et donc <ça a> [x 2] ça pas aidé à [x 4] une bonne **progress** de la culture.[369]
>
> Ex 80 : hum oui je [x 3] **nettoye** [netwaj] un+peu la &cui la cuisine et les chambres.[370]

Un autre problème concerne la réalisation des liaisons, l'absence de réalisation des liaisons et la réalisation de liaisons non canoniques. En ce qui concerne toutes les liaisons réalisées, qu'elles soient obligatoires ou facultatives, elles n'ont pas été marquées car cela relèverait d'un autre niveau d'annotation. Nous n'avons pas non plus marqué l'omission des liaisons obligatoires pour les mêmes raisons. De même, un phénomène que nous avons remarqué très fréquent est la substitution de [s] à [z] n'a pas été marqué :

> Ex 81 : euh j' enseigne le l' anglais pour les petits enfants euh deux trois quatre cinq **six ans**. [sisɑ̃][371]

[365] ara_ca_ma_80_f_09.

[366] spa_en_ro_92_h_13.

[367] spa_pe_gu_93_h_13.

[368] spa_ci_pe_86_f_13.

[369] spa_pe_gu_93_h_13.

[370] eng_be_la_87_h_11.

[371] eng_ca_re_83_f_11.

Ex 82 : très chaleureux la familia@s:spa l' accueillement **les amis**. [lesami][372]

Ce choix pose problème en raison de l'homophonie entre *ils ont* où la liaison a été réalisée [s] et *ils sont*, comme dans le cas suivant :

Ex 83 : **ils ont** beaucoup plus de caractère.[373]

où l'on peut s'interroger s'il s'agit d'une faute dans la réalisation de la liaison, ou d'une confusion entre *être* et *avoir* ; nous avons jugé qu'il s'agissait du premier cas. Nous n'avons pas non plus noté la réalisation non canonique des liaisons :

Ex 84 : bah oui **et aussi** [etosi] ils sont presque tous euh de mon euh nous avons presque la même âge.[374]

En revanche, nous avons traité l'exemple suivant différemment :

Ex 85 : et ça arrive mais ici c'est <si tu> [x 2] &f &fai fais &t une erreur.[375]

Dans (85), l'apprenante a réalisé la liaison [t] entre *fais* et *une* ; sans doute l'apprenante s'est-elle représenté graphiquement *fait* et non *fais*, ce qui aurait pu l'induire en erreur. Mais ce qui nous préoccupe au niveau de la transcription est de représenter ce type de phénomène car il ne relève plus des réalisations ou non de liaisons obligatoires ou facultatives, mais d'un autre type d'erreur. Dans ce cas, nous avons transcrit les phonèmes réalisés et balisés comme des amorces.

3.2.5.2 Difficultés de transcription liées aux formes homonymes

En ce qui concerne les hésitations entre les formes homonymes lors du processus de transcription, Bilger *et al.* (1997 : 66) écrivent que « le contexte sémantique et/ou grammatical, de même que l'intonation, permet d'éliminer ce genre d'ambiguïté », mais que « certaines hésitations subsistent ». Dans le cas de l'interlangue, un paramètre autre que ceux cités entre en jeu, qui est la prise en compte de la probabilité de l'emploi de telle ou telle

[372] spa_al_al_72_f_11.

[373] eng_li_wh_89_f_11.

[374] eng_jo_ga_91_h_12.

[375] eng_al_jo_86_f_10.

structure par des apprenants, et un jugement subjectif ne peut alors être évité. Nous avions discuté des formes homonymiques des flexions verbales mais celles-ci ne sont pas les seules concernées. Considérons l'exemple suivant :

> Ex 86 : mais si tu as votre maison à Madrid c'est normal de rester là et je ne dois je ne veux aller à en location parce+que c'est très cher à Madrid le [x 2] &appart les appartements **ils** sont très [x 2] chers et je suis bien chez moi.[376]

Dans cet exemple, *ils* a été prononcé [i], et les étudiants de M1 avaient transcrit *les appartements y sont très chers*. Le transcript des étudiants correspond à ce qui est effectivement perçu, et le résultat graphique est irréprochable d'un point de vue syntaxique. Nous avons pourtant choisi de transcrire *ils* : la prononciation [i] pour *il(s)* est en effet « attestée en français depuis des siècles » (Cappeau & Gadet, 2007 : note n° 6), et nous avons jugé que le recours au pronom *y* était moins probable qu'une redondance anaphorique chez cette apprenante. L'homonymie entre un pronom et une autre forme peut également faire hésiter dans l'exemple suivant :

> Ex 87 : en+fait je travaille avec un une famille française euh et [x 3] j' ai appris **de** [x 3] que [x 2] les repas c'est très important.[377]

Où nous aurions pu transcrire *d'eux*. Nous avons émis le même type de jugement pour ce qui suit :

> Ex 88 : j' ai trois frères un frère au Golfe j' ai voyagé **quelques temps** chez lui euh une sœur qui est mariée et une sœur à la maison je vis avec mon père ma mère et ma petite fille petite sœur pardon.[378]

Si la locutrice avait été une native du français, nous aurions transcrit *quelque temps*, qui aurait signifié *un certain temps*. Mais il nous semble que l'apprenante exprime ici la répétition au sens de *plusieurs fois*, d'où la transcription *quelques temps*, tournure non usitée en français. Enfin, dans l'exemple suivant :

> Ex 89 : et [x 2] donc je joue au foot **tous le dimanche** et le samedi si je suis libre.[379]

[376] spa_ma_du_92_f_12.

[377] spa_ga_ll_87_f_11.

[378] ara_mo_sa_79_f_09.

[379] eng_be_la_87_h_11.

La partie en gras a été prononcée [tulədimɑ̃ʃ] ; *tous* et *tout* étant homonymes, la transcription *tout* pourrait être justifée par le fait que l'occurrence ait été suivie par [lə] et non par [le]. Nous pensons néanmoins que l'apprenant a bien voulu dire *chaque dimanche*.

Comme nous pouvons le constater, les formes homonymes posent un problème d'interprétation du sens selon la forme choisie, comme c'est le cas dans toute transcription. Mais dans le cas de l'interlangue, la démarche est plus délicate car selon la forme choisie, il peut y avoir ajout, modulation ou suppression d'erreur.

3.2.5.3 Difficultés de transcription liées au marquage du genre et du nombre

Dans les exemples recensés par Bilger *et al.* (1997 : 69-71) concernant la difficulté de marquer le genre et le nombre lors d'une transcription, la difficulté majeure réside dans l'interprétation du sens dans les cas d'homonymies. La transcription de l'interlangue soulève un autre type de problème que nous ne rencontrons pas dans la transcription de la langue native, relatif au fait que les marques du genre et du nombre peuvent n'être que graphiques. Commençons par considérer les exemples suivants :

> Ex 90 : et ma mère peut+être elle m' a **influencée** beaucoup.[380]
>
> Ex 91 : donc ça veut &d euh **mes parents ils souhaitent** que je peux avoir euh des je sais pas je peux avoir des talents sur de l' art ou et+caetera et en+effet **ils aiment** beaucoup la les chansons des films donc peut+être **ils nous donnent et ils me donnent** ce prénom.[381]

Dans l'exemple (90), il est possible que l'apprenante écrive *influencé* sans faire l'accord et sans donc marquer le genre. De même, dans l'exemple (90), il est possible que l'apprenante écrive *il souhaite*, *il aiment* ou *ils donne*, soit ne pas marquer le genre sur le pronom, le verbe ou les deux. Mais dans ces cas où les marques du genre et du nombre ne sont pas orales, ces supputations ne sont pas légitimes car c'est uniquement l'oral qui nous intéresse. Un principe de transcription étant tout de même à appliquer, nous avons décidé de systématiquement marquer le genre et le nombre quand leurs désinences ne se manifestent pas à l'oral, quand

[380] zho_yi_wa_93_f_13.

[381] zho_yi_wa_93_f_13.

bien même l'apprenant aurait pu commettre une erreur à l'écrit. Là encore, le fait d'obtenir des transcripts « graphiquement très au point » ne constitue pas un problème.

Mais en interlangue, quand les marques du genre et du nombre sont orales et que le marquage discontinu n'est pas respecté par l'apprenant, cela donne des énoncés où la flexion verbale attendue n'est pas réalisée :

> Ex 92 : les Iraniennes ils sont des [x 2] personnes magnifiques **ils boit** beaucoup mais ils rigolent toujours ils [x 2] on a dansé tout la soirée.[382]
>
> Ex 93 : non un+peu [x 2] **ils comprend** mais ils parlent pas ils ne parlent pas.[383]
>
> Ex 94 : bon je pense que **tous les mamans tous les grands+mères veut** aider **veut** conseiller.[384]

ou des déclinaisons erronées :

> Ex 95 : mon père est mort il+y+a dix ans **ma mère est mort** il+y+a &s peut+être cinq ans à+peu+près oui.[385]
>
> Ex 96 : ils sont d' **origine français**.[386]

ou l'absence de marquage discontinu :

> Ex 97 : euh en France oui hum **tout seule** parce+que les [x 2] amis il il+n'y+a+pas de temps et donc je vais **tout seule**.[387]

Jusqu'ici, les exemples que nous avons rapportés n'ont pas soulevé de difficultés de transcription car les morphèmes étaient oralement marqués. Le problème concerne les énoncés où le marquage discontinu n'est pas respecté et que dans le même temps, les morphèmes ne sont pas oralement marqués, comme dans l'exemple suivant :

> Ex 98 : mais il+y+a aussi **des choses** qui **est liées** au pétrole aussi.[388]

[382] spa_na_kl_87_f_12.

[383] ara_di_ma_83_f_08.

[384] spa_al_al_72_f_11.

[385] ara_ab_so_83_f_09.

[386] ara_ab_ma_80_h_10.

[387] zho_zh_bo_84_f_11.

[388] ara_ra_ki_79_f_09.

Le morphème pluriel est manifeste oralement dans le déterminant indéfini *des* mais le marquage discontinu n'a pas été respecté dans le choix de *est* au lieu de *sont* : mais comment transcrire [lije] ? Si la marque féminin ne pose pas problème car aucun indice oral ne vient indiquer l'interruption du marquage discontinu du féminin, il n'en va pas de même pour le morphème pluriel et l'on pourrait proposer le transcript *des choses qui est liée*. Cependant, nous avons considéré que le principe de marquer le genre et le nombre quand leurs désinences ne sont pas orales pouvait s'appliquer dans ce type d'occurrence car nous considérons que l'interruption du marquage discontinu n'est pas systématique, et que nous ne pouvons « supposer » une erreur qui n'a pas été commise. L'application de ce principe donne les transcripts suivants :

> Ex 99 : euh je crois je vais euh moitié+moitié &moi euh la moitié pour créer une entreprise et moitié pour [x 2] les &éd éducations [x 2] **pour les enfants qui est pauvres.**[389]

> Ex 100 : j' ai fait le connaissance de beaucoup de **personnes** qui **est** plus **âgées** que moi.[390]

Toutefois, pour les termes qui ne portent pas la marque morphologique du genre, et avec lesquels l'apprenant a utilisé un déterminant marqué différemment du genre du terme, nous avons choisi d'appliquer un marquage discontinu :

> Ex 101 : j' ai [x 2] un ami qui euh le maison **le [x 2] cheminée cassé** les fenêtres tous les fenêtres cassées.[391]

> Ex 102 : je crois le fondamental euh **le connaissance fondamental**.[392]

Dans ces deux exemples, nous n'avons ainsi pas transcrit *cassée* ou *fondamentale* car nous avons jugé que dans ces cas uniquement, le recours au déterminant masculin constituait une erreur dans l'attribution du genre au terme et non une erreur de marquage discontinu. Enfin, en ce qui concerne le nombre, signalons les erreurs des apprenants liées à l'invariabilité de certains lexèmes, où nous avons choisi de ne pas transcrire le morphème pluriel :

[389] zho_ju_zh_87_h_12.

[390] eng_la_sk_86_f_10.

[391] eng_ca_re_83_f_11.

[392] zho_ju_zh_87_h_12.

Ex 103 : on a amusé beaucoup et on a fait la fianç la **fiançaille** de mon frère Achraf um c'est tout en juin.[393]

Ex 104 : tout+le+monde apprend **la math** et l' espagnol ou le français pour &ché ça dépend de chacun et chaque pays et donc.[394]

Ex 105 : euh c'est comme euh d' abord j' aime bien le calcul je suis trop bien à la **mathématique**.[395]

Ex 106 : vacances d' été euh la **vacance** d' été dernier ?[396]

En revanche, nous avons marqué le genre dans le cas des erreurs de flexions lexicales, ce qui a donné des formes inexistantes en français :

Ex 107 : tu sais les liens **familials** sont très forts et par+exemple mon [x 2] père ma mère ils me manquent beaucoup.[397]

Ex 108 : mais en+général oui bien+sûr mais tous [x 2] les milieux sort et mais les [x 4] sorties dépend les [x 3] milieux **socials** les milieux économiques et tout ça oui.[398]

3.2.5.4 Difficultés de transcription liées au lexique de l'interlangue

L'une des spécificités de l'interlangue est la présence de nombreux « néologismes », et il nous faut ici définir ce terme en commençant par un retour sur la question des erreurs lexicales. Granger & Monfort (1994) décrivent dans leur article les différentes conceptions du lexique en général, ainsi que les différentes conceptions du lexique du point de vue de l'enseignement des langues. Dans leur volonté de procéder à une analyse des erreurs lexicales en interlangue, ils adoptent le schéma suivant en tant que catégorisation de ces erreurs (Granger & Monfort, 1994, section 4.4.2.) :

[393] ara_ka_be_81_f_09.

[394] spa_al_ga_79_h_12.

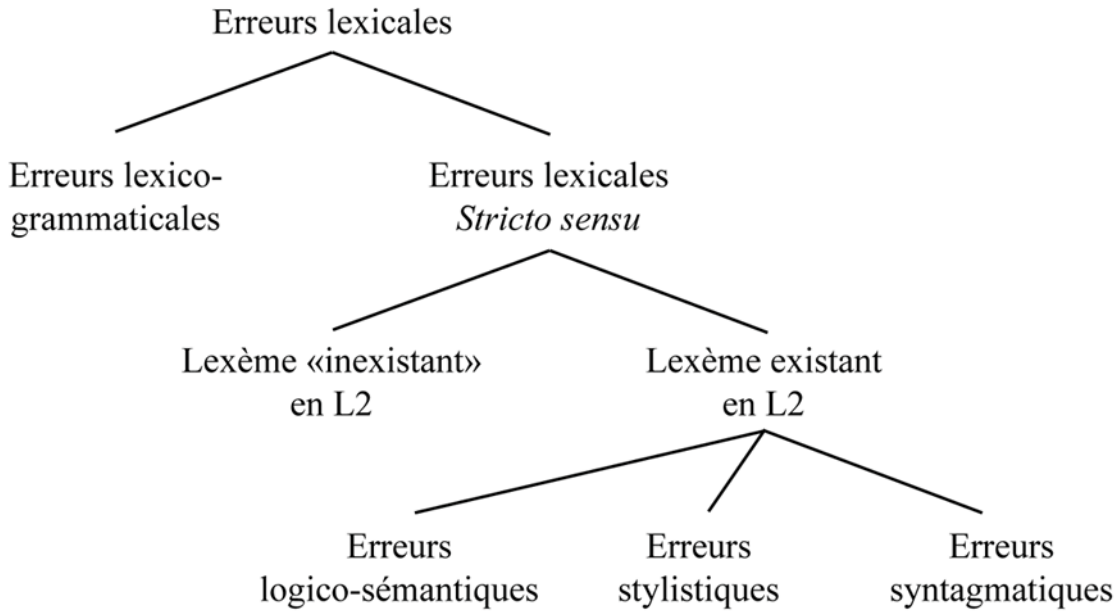
[395] zho_ju_zh_87_h_12.

[396] zho_wa_zh_90_f_11.

[397] spa_ar_ca_73_f_10.

[398] ara_ra_ki_79_f_09.

Figure 5 : Catégorisation des erreurs lexicales selon Granger & Montfort (1994)



Dans la démarche qui est la nôtre, il ne s'agit pas, au stade de la transcription, de relever l'ensemble des erreurs lexicales du corpus ou de les annoter, mais de considérer les erreurs lexicales qui posent problème lors du processus de transcription. Considérons le premier ensemble qui regroupe les erreurs lexico-grammaticales du type[399] :

> Ex 109 : il me semble que **le vie**.[400]
>
> Ex 110 : mais [x 2] ici c'est je peux **améliorer** plus. (m'améliorer)[401]
>
> Ex 111 : <et ma> [x 2] soeur et mon [x 2] beau+frère [x 2] **se** vivent dans presque à+côté [x 2] à Shangaï donc c'est un+peu très loin. (vivent)[402]
>
> Ex 112 : je trouve le système éducative en France est très [x 2] différent **que** le système aux États_Unis. (de / du système)[403]

Nous ne retiendrons pas ce premier type d'erreurs pour deux raisons. En premier lieu, notre conception du lexique est effectivement plus restreinte et nous considérons que les erreurs des

[399] Les exemples suivants, relevés de notre corpus, sont similaires à ceux que proposent Granger & Monfort pour illustrer les erreurs lexico-grammaticales.

[400] eng_al_jo_86_f_10.

[401] spa_ga_ll_87_f_11.

[402] zho_fa_li_86_f_13.

[403] eng_al_jo_86_f_10.

exemples cités sont morphosyntaxiques ou syntaxiques en raison du morphème masculin qui remplace le morphème féminin en (109), de l'omission de pronom personnel réfléchi en (110), de l'ajout de ce pronom en (111) et de la substitution de *que* à la préposition *de* en (112). En second lieu, ce type d'erreur ne pose pas problème au premier niveau de la transcription. Nous ne sommes donc concernés ici que par les erreurs lexicales *stricto sensu*, et nous commencerons par les lexèmes existants en français, sous-partie où nous retrouvons trois autres catégories, dont nous écartons également les erreurs stylistiques ainsi que les erreurs syntagmatiques. Les premières regroupent les « confusions entre les différents registres : familier, soutenu, littéraire, technique, archaïque, etc. » (Granger & Monfort, section 4.2.6.) et les secondes les erreurs de l'axe syntagmatique qui vont à l'encontre de l'usage collocatif du français du type :

> Ex 113 : on diminue le texte en par+exemple cent [x 2] mots. (réduit)[404]
>
> Ex 114 : seulement il+y+a un petit **morceau** de le du pays que c'est à+côté+de le mer. (partie)[405]

Si ces deux types d'erreurs sont effectivement lexicaux, leur prise en compte lors de la transcription par une quelconque annotation relèverait d'une analyse et, à l'instar des erreurs lexico-grammaticales, elles ne posent pas de problème particulier de transcription. Les catégories que nous discuterons ici sont donc les erreurs lexicales concernant les lexèmes inexistants en français ainsi que les erreurs logico-sémantiques, deux catégories qui définissent ce que nous appelons néologismes en interlangue. Les différentes catégorisations que nous empruntons ou que nous proposons ne constituent pas une typologie exhaustive des erreurs lexicales[406].

1) Les lexèmes inexistants en français

Nous avons distingué quatre types de formation de lexèmes inexistants en français. La première, qui est la plus récurrente, concerne les déformations phonétiques de lexèmes existants en français. Nous avions évoqué ce type d'occurrence et constaté que les limites entre les déformations phonétiques nécessitant une graphie non canonique et celles où nous

[404] ara_di_ma_83_f_09.

[405] spa_ga_ll_87_f_11.

[406] Concernant les erreurs lexicales, il conviendra de consulter la thèse d'Anctil (2010) qui offre un relevé critique des travaux antérieurs sur la question ainsi qu'une typologie détaillée.

avons jugé raisonnable d'adopter la graphie conventionnelle sont floues[407]. La seconde catégorie est celle où il ne s'agit plus de simple prononciation, mais de réelles déformations morphologiques de lexèmes existants en français. Il peut s'agir de troncation (*ambian – ambiance / asia – asiatiques*) ou de constructions (*concourances – concours / spécifiquer – spécifier*), mais nous n'avons pourtant pas considéré ces occurrences comme de réelles dérivations, mais comme des déformations. En raison de ces déformations, nous n'avons pas hésité pour ces cas-là à « inventer » de nouvelles graphies. Les troisième et quatrième catégories concernent les dérivations morphologiques basées sur des lexèmes français pour la troisième catégorie et étrangers pour la quatrième. Le tableau suivant montre un relevé exhaustif des lexèmes inexistants en français qui rentrent dans l'une des quatre catégories que nous venons de présenter :

[407] Cf. la série d'exemples (67) à (80).

Tableau 8 : Lexèmes inexistants en français dans le corpus CIL-FLE[408]

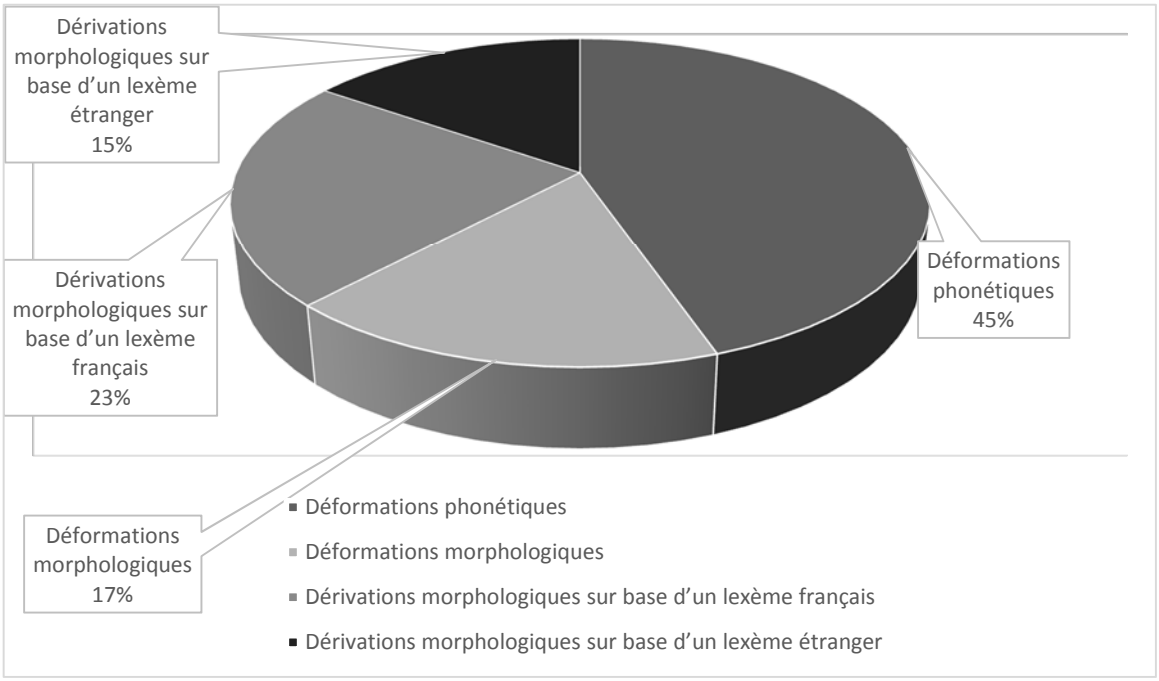
Déformations phonétiques	Déformations morphologiques	Dérivations morphologiques sur base d'un lexème français	Dérivations morphologiques sur base d'un lexème étranger
abstrate "abstraite"	aéres "aériennes"	accueillement "accueil"	beer "bière"
anilation "animation"	ambian "ambiance"	accueilleux "accueillant"	conserwas "conserves"
aréoport "aéroport"	asia "asiatique"	aleppaine "aleppine"	corte "corps"
beaulieues "banlieues"	chousé "choisi"	damassienne "damascène"	désiner "concevoir"
bizoutérie "bijouterie"	colombie "colombien"	discapable "incapable"	entertainer "amuser"
calendéris "calendriers"	concourances "concours"	disségnateur "dessinateur"	étudiantile "étudiante"
commerciant "commerçant"	concréter "concrétiser"	disségner "dessiner"	expectations "suppositions"
communité "communauté"	cubana "cubain"	enriche "enrichir"	fascionés "à la mode"
comparation "comparaison"	étudiz "études"	enseigneur "enseignant"	fornisme
constitucional "constitutionnel"	flu "flûte"	footsalle "foot en salle"	forniste
contest "contexte"	ingénétor "ingénieur"	francheuses "franches"	incièrte
coussons "coussins"	ingéniéro "ingénieur"	gourmandies "gourmandises"	march "mars"
criminal "criminel"	intreprétrice "interprète"	inscriée "inscrite"	migrantes "émigrés"
culturals "culturels"	léglises "églises"	inscrirée "inscrire"	neuratique "nerveux"
cussine "cuisine"	méditerranée "méditerranéen"	interpersonnelles "personelles"	norte "nord"
democracia "démocratie"	nute "nuit"	interprétrice "interprète"	practiquer "pratiquer"
dictatura "dictature"	politica "politique"	marrains "parrains"	protecter "protéger"
didactica "didactique"	politico "politique"	nerviocisme "nervosité"	random "aléatoire"
diploma "diplôme"	presseux "paresseux"	offri "offre"	relax "détendu"
directa "directe"	processes "progrès"	opératives "opérations"	ségures "sûrs"
droge "drogue"	racinée "enracinée"	pensateurs "penseurs"	space "espace"
écomique "économique"	radiographe "radiographie"	physiothologie "physiopathologie"	summer "été"
éconique "économique"	signifis "significations"	pizzeri "pizzeria"	tatement "contact"
émeubles "meubles"	solidare "solitaire"	populé "peuplé"	
espécialisation "spécialisation"	spécifiquer "spécifier"	porciculture "élevage de porcs"	
espécialité "spécialité"	yog "yoga"	receptibles "réceptifs"	
espir "espoir"		relationné "lié"	
esporte "sport"		relationnent	
esportes "sports"		relationner "lier"	
éstablissement "établissement"		relationnés "liés"	
estage "stage"		ressemblement "ressemblances"	
estrangères "étrangères"		restringieux "limités"	
estressé "stressé"		revivrer "revivre"	
estudier "étudier"		vitement "rapidement"	
exprimir "exprimer"			
faculta "facultés"			
familia "famille"			
familias "familles"			

[408] Le genre et le nombre sont marqués selon l'utilisation du terme dans le corpus.

formals "formels" galatte "galette" gouvernementelle "gouvernementale" laiter "allaiter" mama "maman" mélage "mélange" méterranéenne "méditerranéenne" niège "neige" obliguée "obligée" pauverté "pauvreté" pauvles "pauvres" plouse "pelouse" prense "presse" professionnal "professionnel" professionnalle "professionnelle" progrèsse "progrès" provence "province" rituals "rituels" royame "royaume" rute "route" souvants "suivants" terribe "terrible" tranqui "tranquille" una "une" universé "université" veur "veux"			

Comme nous l'avons mentionné, sur les 150 occurrences, les déformations phonétiques sont les plus nombreuses :

Figure 6 : Répartition des lexèmes inexistants en français dans le corpus CIL-FLE



Nous avons conscience que les déformations phonétiques peuvent ne pas être considérées au regard du lexique. Mais nous en parlons ici pour trois raisons. La première est que la manifestation d'une prononciation erronée est ici lexico-morphologique. La seconde est que nous attribuons les néologismes de cette catégorie à une prononciation erronée, mais ceci reste une hypothèse et non une certitude ; rien ne nous permet de certifier que *ritual*, *commerciant* ou *exprimir* sont des déformations phonétiques et non de réelles constructions morphologiques erronées, bien que nous le supposions mais les limites entre déformations et constructions, et donc entre erreurs de prononciation et erreurs lexicales, ne sont pas tranchées. Enfin, la dernière raison est d'ordre technique : quelle que soit la catégorie d'erreur à laquelle appartiennent ces occurrences, elles doivent être étiquetées en tant que néologismes pour CLAN puisqu'elles sont représentées graphiquement par des suites qui ne sont pas présentes dans le lexique de référence. Nous rappelons que ces quatre catégories doivent être détaillées dans le cadre d'une analyse approfondie des erreurs lexicales, nous ne les proposons que dans le cadre d'une réflexion sur le processus de transcription face à ces phénomènes.

Dans certains cas, les erreurs de morphologie verbale ont également nécessité la création de nouvelles graphies. Nous ne discutons pas ici de l'absence de conjugaison d'un verbe ou la confusion entre deux désinences qui ne posent relativement pas de problème, mais du recours

de l'apprenant à des formes morphologiques inventées dont résultent des graphies non conventionnelles :

> Ex 115 : et après il+y+a une autre personne qui [x 2] le garde mais l' handicapé ne se **sense** bien ne se **sense** pas bien avec lui.[409]
>
> Ex 116 : à un moment tu dis bah ça fait un an là maintenant et je me sens je &p probablement &j je [x 2] **regrè** c'est ça attends c'est le contraire de progrès c'est ?[410]
>
> Ex 117 : c'est vrai qu' ici <c'est plus> [x 2] amical hein les personnes se **réunent** pour [x 2] partager avec ses amis les parents les enfants toutes mais en Colombie par+contre c'est différent.[411]

La liste qui suit recense les « inventions graphiques » correspondant à ce phénomène dans le corpus :

apprende "apprenne"	aye "aie"	ayes "aies"
buvé "bu"	croyent "croient"	faisent "font"
faisez "faites"	parcouri "parcouru"	étudi "étudié"
li "lu"	lirer "lire"	mettrerait "mettrait"
nettoye "nettoie"	peinsent "peignent"	peuve "peut"
participassais "participais"	doivait "devait"	découvri "découvert"
regrè "régresse"	réapprende "réapprends"	réunent "réunissent"
saient "savent"	saivent "savent"	tudie "étudie"
sense "sent"	étudent "étudient"	participassais "participais"

2) Les erreurs logico-sémantiques

CLAN ne peut repérer que les néologismes dont la forme est inexistante en français. Quand la forme du néologisme correspond à une forme existante en français, il ne peut être repéré qu'au cours d'une analyse dédiée. Il s'agit donc ici « des erreurs qui sont dues à une méconnaissance du sens dénotatif du mot » (Granger & Monfort : section 4.2.5.). Les sources de ces erreurs sont diverses :

Les interférences entre la langue maternelle de l'apprenant et le français :

[409] spa_se_ma_93_h_12.

[410] ara_ab_ma_80_h_10.

[411] spa_ga_ll_87_f_11.

Ex 118 : <le guide> [x 2] euh de septième de huitième et de neuvième euh met des [x 4] **marches** euh très précises. (Le terme *marches* est une traduction littérale de l'arabe du même terme, mais qui signifie en arabe *étapes*)[412]

Ex 119 : il+y+avait des miettes sur le **compteur**. (*counter* en anglais)[413]

La confusion entre deux lexèmes français :

Ex 120 : mais maintenant le première décembre que c'est le première jour <qu' il a pris> [x 2] **procession** du pouvoir.[414]

Ex 121 : je suis **attaché** au labo. (Au sens de *rattaché*)[415]

Ex 122 : mais j' aimerais bien avoir un grand jardin avec <des arbres> [x 2] arbres et des fleurs avec une piscine parce+que si j' ai des enfants je crois que ça sera bien pour [x 2] eux et dans la **champagne** j' aimerais bien que soit à la **champagne**.[416]

La confusion entre les différentes catégories grammaticales du lemme :

Ex 123 : euh ils ont gardé cette politique mais maintenant il+y+a une [x 4] **ouvert** pour les langues étrangers et mais surtout pour l' anglais. (ouverture)[417]

Ex 124 : oui dans des [x 2] quartiers populaires qui est un+peu **conservés** il ne faut pas habiller très euh. (conservateurs)[418]

Ex 125 : mais [x 2] justement de cette semaine dernier dernière je j' étais en+train+de d' aller la chercher à l' école donc on a eu l' opportunité de faire des petites activités comment aller [x 2] au parc ou manger dans un **boulangère** oui et voilà et pas beaucoup enfin aller au parc ah ouais j' ai déjà dit excusez moi. (boulangerie)[419]

Ex 126 : tous les &indus tous les industries qui [x 3] est basée sur les [x 2] produits **agriculture** par+exemple. (agricoles)[420]

Nous n'avons pu justifier certaines occurrences car elles demeurent incompréhensibles :

[412] ara_ab_so_83_f_09.

[413] eng_la_sk_86_f_10.

[414] spa_ci_pe_86_f_13.

[415] ara_ab_ma_80_h_10.

[416] spa_ma_mo_89_f_12.

[417] ara_ra_ki_79_f_09.

[418] ara_ra_ki_79_f_09.

[419] spa_pe_gu_93_h_13.

[420] ara_ra_ki_79_f_09.

> Ex 127 : <elle avait> [x 2] mis la langue française <dans le> [x 2] **plateau** et Manu Chao je l' aime beaucoup.[421]

Pour l'ensemble de ces exemples, ni l'annotation, ni la création de nouvelles graphies n'ont été retenues.

3.3 Perspectives et possibilités logicielles de CLAN

Le protocole de transcription appliqué a la volonté de limiter les phénomènes oraux représentés. Nous ne considérons pas ceci comme un appauvrissement puisque le son, aligné aux transcripts, permet la réécoute de ces phénomènes. Une transcription minimale élargit au contraire le champ des possibilités d'exploitation du corpus. Nous illustrerons ici quelques-unes de ces possibilités en indiquant, quand cela sera nécessaire, les indications logicielles correspondantes en précisant à nouveau que nous ne désirons pas fournir un manuel d'exploitation technique[422].

En premier lieu, CLAN n'est pas un logiciel de *speech language* mais de *spoken language*, et il ne se prête pas à des analyses phonétiques, phonologiques ou prosodiques. À cet effet, PRAAT ou ELANB sont plus indiqués. CLAN possède toutefois la commande PHONFREQ permettant le calcul de la fréquence des phonèmes transcrits en fonction de leur position, à la condition d'une annotation phonétique préliminaire. Il est possible de conduire ce type d'analyse sur les entretiens, autrement dit sur l'interlangue spontanée, ainsi que sur les lectures imposées afin de procéder à des analyses comparatives.

Les annotations métalinguistiques présentes dans chaque fichier de transcription peuvent laisser envisager des études sur la maîtrise des apprenants des variations sociolinguistiques. Bien que Mougeon *et al.* (2002 : 21) constatent que le répertoire sociolinguistique des apprenants en FLE « n'inclut (presque) pas de variantes non standard marquées », les auteurs remarquent que leurs observations reposent sur un nombre trop restreint d'études pour pouvoir éliminer l'impact des critères d'âge ou de sexe sur l'interlangue.

---

[421] spa_ar_ca_73_f_10.

[422] Cf. notes n°293 et n°294 pour les manuels complets en anglais. Un manuel simplifié en français est disponible à l'adresse suivante :
http://www.ddl.ish-lyon.cnrs.fr/fulltext/Chenu/Fchenu_Formation_Childes2012.pdf
Cf. annexe 1 pour l'installation et l'utilisation des outils primaires de CLAN.

Comme la plupart des logiciels, il est possible d'opérer un calcul de fréquence des items du corpus via la commande FREQ. Il est évidemment possible de spécifier le locuteur à analyser : en conduisant la commande sur les énoncés d'apprenants uniquement, nous constatons que les apprenants ont produit 105495 occurrences en utilisant 5451 items différents, dont les quinze plus fréquents sont les suivants, avec leur nombre d'occurrences :

Syntaxe de la commande à écrire : FREQ +tFLE +o +u *.cha >freq.cex

3881 euh	3395 je	3328 et
3234 c'est	2577 de	2316 le
2192 j'	2005 la	1659 les
1634 oui	1611 pas	1511 mais
1365 à	1347 ai	1294 pour

2465 items, soit 45,22% des items différents, n'apparaissent qu'une seule fois dans le corpus[423]. Il est également possible de calculer la fréquence d'un seul item ou d'un ensemble d'items. La commande FREQ n'indique toutefois que les fréquences des items ; afin de localiser des occurrences, nous utiliserons le concordancier.

Le concordancier de CLAN permet via la commande COMBO la localisation en contexte d'un item, de plusieurs items qui se suivent, d'un item qui ne sera pas suivi d'un item spécifié etc. Afin de faciliter les recherches, il est également possible de créer des listes d'items sur lesquelles les commandes opéreront. Ces listes sont plus ou moins déjà constituées dans le lexique de CLAN mais il faut les ajuster pour qu'elles puissent être utilisées pour les commandes, et il est alors possible de localiser toutes les occurrences de telle catégorie grammaticale (adjectifs, adverbes, verbes irréguliers) ou les occurrences d'items regroupés sous des listes qui ne sont pas grammaticalement prédéfinies, et qu'il faudra constituer : liste des marqueurs discursifs, liste d'items en vue de l'analyse d'un phonème en particulier, liste des néologismes etc.

Nous avons par exemple constitué une liste regroupant l'ensemble des possibilités de la double pronominalisation, intitulée doublepron.cut, et nous l'avons placée dans le dossier lib

[423] Les résultats de la commande FREQ sur l'ensemble du corpus sont trop volumineux pour figurer ici ou même en annexe.

de CLAN, elle est constituée sur le modèle : me^le / me^la etc[424]. Le symbole ^ indiquant à CLAN que le second item devra immédiatement suivre le premier. En lançant la commande :

COMBO +s@doublepron.cut +tFLE *.cha

Puis en supprimant manuellement les occurrences du type :

> Ex 128 : mais pour **(1)nous (1)les** euh mes frères et mes soeurs et moi nous avons étudié la langue française.[425]

Nous obtenons les seules six occurrences de double pronominalisation du corpus :

> Ex 129 : je **(1)vous (1)en** prie.[426]
>
> Ex 130 : je vous **je (1)t' (1)**en prie.[427]
>
> Ex 131 : et ma mère m' a proposé tu pouvais aller à la plage avec tes sœurs une semaine et tu vas **(1)me (1)le** laisser sans aucun problème mais je lui dis non écoute il viendra avec nous.[428]
>
> Ex 132 : de **(1)le (1)lui** de lui laisser avec avec eux.[429]
>
> Ex 133 : ah oui Paris ah oui j' ai oublié je viens d' oublier ouais voilà ouais Paris ouais bien+sûr et j' ai été aussi ah je **(1)m' (1)en** rappelle plus c'était à Brest voilà c'est à Brest.[430]
>
> Ex 134 : c'est mal en tête parce+que je **(1)te (1)l'** ai dit je n' ai pas la talent.[431]

Il est ainsi possible de localiser toute suite d'occurrences, d'éditer l'ensemble du corpus en une seule commande, de choisir le format de sortie et de convertir les fichiers vers d'autres formats que le format .cha.

[424] En ce qui concerne le protocole technique à suivre pour la construction de telles listes et les commandes idoines, voir MacWhinney (2014b : 73).

[425] ara_ge_fr_80_f_08.

[426] ara_di_ma_83_f_09.

[427] ara_ma_ze_84_f_09.

[428] spa_al_al_72_f_11

[429] spa_al_al_72_f_11.

[430] spa_pe_gu_93_h_13.

[431] zho_fa_li_86_f_13.

Intéressons-nous maintenant à un autre type d'analyse, qui concerne les annotations, et considérons en premier lieu celle qui peut être automatisée sous CLAN : l'annotation morphosyntaxique. CLAN permet de lancer la commande MOR qui générera une ligne %mor après chaque tour de parole. Cette ligne contiendra un classement morphosyntaxique des items du tour de parole. Nous donnons l'exemple suivant pour un seul tour de parole :

> Ex 135 : il+y+a aussi le domaine euh le les [x 2] transit qui passent de la Syrie c'est très important maintenant euh le la place de la Syrie.

La commande MOR *.cha insérera une ligne d'annotation %MOR qui suivra chaque tour de parole, comme suit :

```
*FLE:   il+y+a aussi le domaine euh le les [x 2] transit qui passent de la
        Syrie c'est très important maintenant euh le la place de la Syrie. •
%mor:   v:exist|il+y+a^v:exist|il+y+a&CPL conj|aussi^adv|aussi
        pro:obj|le&MASC&SING^det|le&MASC&SING n|domaine&_MASC co|euh
        pro:obj|le&MASC&SING^det|le&MASC&SING pro:obj|les&PL^det|les&PL
        n|transit&_MASC^v|transir-PASS-_3SV pro:rel|qui^pro:int|qui
        v|passer-SUBJV:PRES-_3PV^v|passer-PRES-_3PV
        prep|de^prep:art|de^det|de^adv|de
        pro:obj|la&FEM&SING^det|la&FEM&SING n:prop|Syrie ?|c ?|'
        v:exist|être&PRES&3SV^v:aux|être&PRES&3SV^adj|est adv|très
        n|important^adj|important^v|importer-PPRE
        adv|maintenant^v|maintenir-PPRE co|euh
        pro:obj|le&MASC&SING^det|le&MASC&SING
        pro:obj|la&FEM&SING^det|la&FEM&SING
        n|place&_FEM^v|placer-SUBJV:PRES-_3SV^v|placer-SUBJV:PRES-_1SV^v|placer-IMP-_2SV^v|placer-PRES-_3SV^v|placer-PRES-_1SV
        prep|de^prep:art|de^det|de^adv|de
        pro:obj|la&FEM&SING^det|la&FEM&SING n:prop|Syrie .
```

Cette première démarche attribue à un item toutes les catégories grammaticales possibles qui peuvent être attribuées à tous ses homographes. L'item *place* peut ainsi être :

le nom féminin *place* :	n|place&_FEM
le verbe *placer* à la 3ème personne du singulier au subjonctif présent :	v|placer-SUBJV:PRES-_3SV
le verbe *placer* à la 1ère personne du singulier au subjonctif présent :	v|placer-SUBJV:PRES-_1SV
le verbe *placer* à la 2ème personne du singulier à l'impératif :	v|placer-IMP-_2SV
le verbe *placer* à la 3ème personne du singulier au présent indicatif :	v|placer-PRES-_3SV
le verbe *placer* à la 1ère personne du singulier au présent indicatif :	v|placer-PRES-_1SV

Afin de désambiguïser les annotations morphosyntaxiques, il faudra lancer le programme POST, inclus dans CLAN, qui propose la solution la plus fréquente en cas d'ambiguïté, en se basant sur une base de données de référence désambiguïsée. Ainsi, pour morphosyntaxiquement annoter le fichier ara_sa_ka_80_f_09, nous lançons la commande suivante :

MOR +tFLE ara_sa_ka_80_f_09.cha

La seconde étape est la désambiguïsation grâce à la commande suivante :

POST ara_sa_ka_80_f_09.cex

Enfin, la troisième et dernière étape est la désambiguïsation manuelle des items que le programme POST n'a pu désambiguïsés. Voici une partie de l'annotation obtenue :

*FLE: donc euh d' habitude quand on arrive à la saison du printemps
c'est+à+dire vers la fin de l' année donc on commence à espérer à
[x 2] attendre l' été pour faire beaucoup de choses beaucoup d'
activités faire des pique+niques des voyages même et+cetera .
%mor: conj|donc co|euh prep|de n|habitude&_FEM conj|quand pro:subj|on&3S
v|arriver-PRES-_3SV prep|à det|la&FEM&SING n|saison&_FEM prep:art|du
n|printemps&_MASC&_SINGPL conj|c'est+à+dire prep|vers
det|la&FEM&SING n|fin prep:art|de pro:subj|le/la&MASC n|année&_FEM
conj|donc pro:subj|on&3S v|commencer-PRES-_3SV prep|à v|espérer-INF
prep|à v|attendre-INF det|le/la&SING n|été&_MASC prep|pour
v:mdl|faire-INF adv|beaucoup prep|de n|chose-_PL adv|beaucoup
prep|de n|activité&_FEM-_PL v:mdl|faire-INF det|des&PL
n|pique+nique&_MASC-_PL det|des&PL n|voyage&_MASC-_PL adv|même
adv|et+cetera .

Il est alors possible de lancer les recherches sur les catégories morphosyntaxiques du lexique de CLAN, qui sont les suivantes :

adj	adj	adjectif
adv	adv	adverbe
adv:int	adv:int	adverbe interrogatif
adv:neg	adv:neg	adverbe de négation
adv:place	adv:place	adverbe de lieu
adv:yn	adv:yn	adverbe oui/non
co	co	communicateur ou interjection
conj	conj	conjonction
det	det	determinant (articles définis et indéfinis)
det:dem	det:dem	déterminant démonstratif
det:gen	det:gen	déterminant général (autres déterminants que les
det:poss	det:poss	déterminant possessif
n	n	nom
n:let	n:let	lettre
n:prop	n:prop	nom propre

num	num	numéro
on	on	onomatopée
pct	pct	ponctuation
prep	prep	préposition
prep:art	prep:art	préposition-article
pro	pro	pronom (général)
pro:dat	pro:dat	pronom personnel datif
pro:dem	pro:dem	pronom démonstratif
pro:int	pro:int	pronom interrogatif
pro:obj	pro:obj	pronom personnel objet direct
pro:refl	pro:refl	pronom réfléchi
pro:rel	pro:rel	pronom relatif
pro:subj	pro:subj	pronom personnel sujet
pro:y	pro:y	pronoms y, en
unk	unk	catégorie indéfinie (xxx, yyy)
v	v	verbe (conjugué)
v&INF	v:aux	verbe auxiliaire
v&PP	v:exist	verbe d'existence (être et il+y+a)
v&PPRE	v:inf	verbe infinitif
v:aux	v:mdl	verbe modal (je fais cuire un gâteau)
v:aux&INF	v:mdllex	verbe modal lexical (je fais un gâteau)
v:aux&PP	v:poss	verbe d'appartenance (verbe avoir lexical)
v:exist	v:pp	verbe participe passé
v:exist&INF	v:ppre	verbe participe présent
v:exist&PP		
v:inf		
v:mdl		
v:mdl&INF		
v:mdl&PP		
v:mdl&PPRE		
v:mdllex v:mdllex&INF		
v:mdllex&PP		
v:mdllex&PPRE		
v:poss		
v:poss&INF		
v:poss&PP		

Par exemple, nous pouvons lancer sur le fichier annoté la commande suivante :

combo +t%MOR +s"v*|*" ara_sa_ka_80_f_09.cha >verbes.cex

qui nous donnera la totalité des verbes utilisés par l'apprenante, au nombre de 239. La commande :

combo +t%MOR +s"v|*PP*" ara_sa_ka_80_f_09.cha >passecompose.cex

affinera les recherches sur les verbes conjugués au passé composé, au nombre de 32 chez cette apprenante.

Le contenu de ces listes morphosyntaxiques est modifiable, et il est également possible d'ajouter des listes supplémentaires (nous avons ajouté par exemple la catégorie « neo »). Il n'est cependant pas possible de modifier l'intitulé des étiquettes à moins de créer un nouveau lexique pour CLAN.

Selon MacWhinney (2014b : 157), l'annotation morphosyntaxique d'un corpus d'anglais peut voir sa marge d'erreur descendre en dessous de 6%. Il précise que ce taux est plus faible que le taux d'erreur de la plupart des annotateurs humains. En ce qui nous concerne, il est nécessaire de préciser qu'il ne peut être envisagé, dans un premier temps, une annotation automatisée de CIL-FLE pour des raisons liées au programme POST. La base de données sur laquelle se base le programme a été constituée grâce à des productions d'enfants, et pour l'annotation de productions d'enfants. Les spécificités de l'interlangue étant autres, nous avons constaté qu'il était très souvent nécessaire de désambiguïser manuellement, mais aussi de corriger certains choix de POST. Si une annotation morphosyntaxique du corpus est envisagée, la constitution d'une nouvelle base de données pour POST est à prévoir, et l'efficacité du programme sera proportionnelle au volume de la base de données. Enfin, il est préférable qu'une telle annotation soit effectuée par une équipe, en raison du temps que prendrait une telle annotation pour une seule personne, mais surtout parce qu'une confrontation des jugements et des vérifications mutuelles sont nécessaires à ce type de processus.

Pour conclure sur possibilités d'annotation de CLAN, nous dirons qu'un corpus d'apprenants appelle naturellement à une annotation évaluative. Comme le note Granger (2007 : 2), « une description détaillée d'erreurs d'apprenants ne peut que contribuer à un but essentiel de l'apprentissage, à savoir aider les apprenants à atteindre un haut niveau de précision dans la langue cible ». Le logiciel propose un ensemble très fourni de balises pour annoter les différents types d'erreurs (MacWhinney, 2012 : 99-100), mais la typologie des erreurs en interlangue est une question très ouverte qui ne fait pas consensus[433] et il conviendra dans le cadre d'un projet d'annotation évaluative de CIL-FLE d'ajuster les balises proposées par CLAN à l'interlangue ou de proposer une autre typologie originale ou inspirée par des travaux antérieurs.

[433] Voir à ce sujet James (2013).

Granger (1999 : 201) précise que les corpus d'apprenants, en tant que corpus spécialisés, nécessitent naturellement leurs propres techniques d'analyse. Nous avons constaté que CLAN nécessitait effectivement que son utilisateur s'approprie un bagage technique relativement important, dès le processus de transcription. À ce propos, Parisse & Morgenstern écrivent :

> Il est de nos jours absolument nécessaire de transcrire ses enregistrements audio ou vidéo avec des outils informatisés. Ces outils peuvent sembler fastidieux au premier abord, effrayer le novice, mais ils rendent le travail du transcripteur beaucoup plus confortable et rigoureux. Ils permettent d'aller plus vite une fois qu'on les maîtrise, d'utiliser des formats de codage normalisés, de faire des analyses de fréquence et surtout d'aligner les transcriptions avec les enregistrements, ce dont on ne saurait aujourd'hui se passer. Grâce à ces outils informatisés, on se donne également les moyens d'entrer dans une communauté internationale de chercheurs qui mutualisent leur données via internet, de pérenniser les données et on apporte sa pierre à la construction des connaissances sur le langage. (Parisse & Morgenstern, 2010 : 203)

Maintenant que CIL-FLE est constitué, on ne peut s'attendre à ce qu'il offre de lui-même des vérités sur l'interlangue. L'analyse de ce corpus pourrait se faire en le réécoutant et en y relevant des faits linguistiques au fur et à mesure. CLAN possède toutefois un panel d'outils puissants qui peuvent grandement accélérer et optimiser le travail de l'analyste.

Conclusion générale

La constitution du corpus CIL-FLE avec ce qu'elle a comporté de réflexion sur les corpus oraux et les méthodologies de transcription, est une démarche originale dans son approche de l'interlangue des apprenants en FLE. Cette thèse, dont l'objectif était « la constitution d'un corpus oral d'interlangue », a abordé toutes les notions qu'impliquait cet intitulé. Nous avons donc consacré le premier chapitre à un retour sur l'histoire des corpus afin de comprendre les enjeux épistémologiques de l'oralité dans les corpus, et avons constaté que l'oralité avait suscité l'intérêt des chercheurs à des époques différentes selon le domaine de la linguistique, mais nous avons pu toutefois démontrer les liens entre les corpus et l'enseignement des langues qui constitue « un des piliers de la linguistique de corpus » selon les termes de Williams (2008 : 11). La linguistique de corpus oraux n'a néanmoins pu réellement se développer que grâce aux outils technologiques, enregistreurs et ordinateurs ; la technologie étant toutefois une condition nécessaire, mais non suffisante, comme en témoigne la situation actuelle des corpus oraux en France.

Dans le second chapitre, nous avons ensuite étudié la question de la constitution des corpus, problématique qui nous a amené à nous intéresser à la linguistique de corpus où nous avons détaillé les spécificités d'une telle approche en linguistique, puis à identifier les différents constituants d'un corpus afin d'en proposer une définition. Si les critères déterminants que nous avons retenus peuvent être discutés ou enrichis, nous avons proposé que toute constitution soit entamée dans une perspective de diffusion des corpus, dans l'attente d'une décision institutionnelle d'envergure qui permettra aux corpus oraux de mieux se développer en France.

Le troisième chapitre a détaillé les spécificités de constitution d'un corpus oral d'interlangue : l'enjeu majeur de cette question est celui de la transcription des données orales d'apprenants et les difficultés supplémentaires qu'elle implique par rapport à la transcription de la langue native. Le premier chapitre de cette thèse est donc historique, le second théorique et le troisième appliqué. Cette présentation logique ne représente pas la réalité, où nous avons commencé par nous former à CLAN avant de réellement réfléchir sur le corpus et sa linguistique. Les étapes sont d'autre part évidemment imbriquées les unes dans les autres : la révision des derniers transcripts soulevait de nombreux questionnement théoriques et les

travaux académiques confortaient ou contredisaient les choix de transcription tout au long du processus.

Nous nous proposons, pour conclure cette thèse, de revenir sur la constitution du corpus CIL-FLE. En premier lieu, nous remarquons que la plupart des thèses s'intéressant aux corpus, à l'interlangue ou aux deux adoptent l'une des méthodologies suivantes : soit un corpus est constitué pour les besoins analytiques d'un phénomène linguistique en particulier, soit un corpus extérieur est consulté. Les thèses qui, outre les résultats d'une analyse linguistique, offrent un corpus constitué selon des critères scientifiques sont rares, et le sont davantage lorsqu'il s'agit d'un corpus oral d'interlangue. Certes, nous aurions pu concentrer nos efforts sur le repérage de tel ou tel phénomène en interlangue en vue de l'analyser, mais nous croyons que le choix de proposer un corpus oral d'interlangue est une démarche au moins tout autant productive que l'analyse d'un phénomène particulier car nous osons espérer que CIL-FLE constituera un champ d'exploration pour les chercheurs intéressés par l'interlangue et permettra ainsi une série d'études plus générales. De plus, considérer que la constitution d'un corpus oral d'interlangue est un processus dénué d'analyse supposerait que la transcription est elle-même une représentation graphique supposée neutre du signal sonore. À propos de la valeur académique à accorder à ce type de démarche, Cappeau & Gadet écrivent :

> Durant de nombreuses années, une vision naïve des transcriptions avait cours, que l'on peut résumer ainsi : dans l'échelle des tâches que le linguiste accomplit, la transcription occupe une place précoce (en général après le recueil du corpus) et peu valorisée. Même au-delà du cas extrême mais encore existant où elle est déléguée à des étudiants (plus ou moins spécifiquement formés), elle n'est à peu près jamais considérée comme une activité de recherche à part entière : il y a peu de réflexions sur son établissement (on reconduit des modèles antérieurs), et surtout elle est trop souvent vite considérée comme définitive, sans dynamique et sans nécessité de retour. Il s'agirait seulement d'une phase ingrate et hélas nécessaire, qui précède le seul véritable travail du linguiste : l'analyse (et l'annotation). Divers travaux (…) ont plaidé pour réévaluer cette phase qui loin d'être une étape préalable à la description devrait en être une partie intégrante, pas spécialement localisée dans l'ordonnancement des tâches. (Cappeau & Gadet, 2012)

Le corpus que nous proposons aspire à être dynamique. À certains niveaux, les données sont malheureusement définitives, notamment en ce qui concerne les tâches imposées aux apprenants. Cette étape, jugée *a posteriori*, aurait sans doute bénéficié d'une meilleure structuration, ou d'un enrichissement par l'ajout de tâches communes comme la description d'une séquence filmique muette ou l'incitation à la production de phénomènes linguistiques

particuliers grâce à des supports dédiés. La poursuite du projet CIL dans sa globalité peut toutefois inclure d'autres types de données à l'avenir. Il serait aussi intéressant de surmonter les difficultés logistiques de la mise en place d'une collecte longitudinale en vue d'étudier les « itinéraires acquisitionnels et stades de développement » en FLE (Bartning & Schlyter, 2004). Quoi qu'il en soit, le caractère définitif n'est dommageable que lorsqu'il concerne les transcripts, et nous soumettons ceux de CIL-FLE aux points de vue de ceux qui s'y pencheront. Considérons ce qu'écrit Morgenstern :

> Les transcriptions sont marquées par un problème général : chaque chercheur a sa propre méthode pour encoder les données. Beaucoup ont développé un système assez sophistiqué pour représenter différents aspects des informations collectées, et leur corpus devient inutilisable pour les autres chercheurs. Une méthode standardisée devenait donc nécessaire afin de faciliter la mutualisation du travail entre chercheurs en acquisition du langage. (Morgenstern, 2009 : 198)

En l'absence d'une « méthode standardisée », nous avons tenu à ce que notre protocole de transcription n'entrave pas l'exploitation, la modulation ou l'adaptation des transcripts par une sophistication trop poussée. Comme nous l'avons vu, le choix de *tout* retranscrire aurait nécessairement conduit à l'annotation de tous les phénomènes phonétiques, prosodiques, lexicaux, syntaxiques, pragmatiques et les annotations évaluatives ou morphosyntaxiques auraient été tout aussi légitimes. L'enrichissement du corpus de ces annotations auraient de notre point de vue été un appauvrissement car il aurait limité l'exploitation du corpus en conduisant de telles analyses lors de la transcription, car ces annotations auraient bel et bien représenté une analyse qui, au final, ne serait que celle du transcripteur. Nous plaidons une dernière fois pour la nécessité de la réécoute des corpus oraux : un corpus oral n'est pas un corpus de transcripts à lire, mais un corpus à écouter, les transcripts n'étant là que pour faciliter les recherches, les analyses et les annotations que nous venons d'évoquer. Un protocole de transcription unanime restera de toute manière une utopie : aucun système de représentation graphique ne pourra être « fidèle » au signal sonore, en raison des différences de perceptions et du caractère intranscriptibles de la matière phonique. Blanche-Benveniste & Jeanjean écrivaient à ce propos :

> Transcrire de la langue parlée tient un peu du paradoxe : garder dans une représentation écrite certaines caractéristiques de l' « oralité » ; faire le « rendu » de la chose orale tout en restant dans ses habitudes de lecture établies depuis longtemps pour la chose écrite… On va se trouver tiraillé entre deux exigences : la fidélité à la

> chose parlée et la lisibilité de son rendu par écrit. (Blanche-Benveniste & Jeanjean, 1987 : 115)

À propos des perspectives du corpus, CIL-FLE a pour ambition de permettre une meilleure connaissance de l'interlangue grâce à des analyses directes, mais aussi grâce à des analyses contrastives, que ce soit des études non-natif/non-natif ou non-natif/natif, afin de spécifier les processus d'apprentissage liés aux communautés ou groupes linguistiques représentés. Il est envisageable de synthétiser les difficultés mais également les acquis de chaque communauté linguistique au cours de l'apprentissage du français. En cela, CIL-FLE ambitionne d'offrir aux chercheurs et enseignants des résultats qui leur permettront d'orienter les manuels et les cours de FLE en fonction du public visé.

De plus, CIL-FLE constituera un outil de consultation profitable aux enseignants quant à leurs vérifications des acquis de tel ou tel fait de langue, et de la nécessité éventuelle de remédiation. Il ne s'agit ici que de l'évocation des possibilités d'exploitation d'un corpus ; notre rôle est pour l'instant de s'assurer de sa capacité à être diffusé à deux niveaux : le premier concerne le corpus lui-même et le second est logistique. Nous espérons avoir constitué un corpus « diffusable » ; outre un protocole de transcription simplifié, nous avons à cet effet tenté de remplir les conditions juridiques nécessaires en anonymisant les données et en faisant signer aux apprenants un formulaire de consentement ; nous avons documenté le corpus afin qu'il constitue une base de travail définissable et, à défaut d'être représentatif, sa représentativité peut être évaluée ; enfin, ses formats sont des formats reconnus en linguistique de corpus et une version XML est présente afin d'assurer la pérennité des données. En ce qui concerne la logistique de diffusion, le consortium Consortium Corpus Oraux et Multimodaux (IRCOM) a été contacté en octobre 2014 afin que le corpus CIL soit déposé auprès de Speech and Language Data Repository (SLDR/ORTOLANG) et recensé dans l'inventaire de l'IRCOM. Une fois que de dépôt aura lieu, l'équipe LIDILE envisage également un dépôt sur la plateforme CHILDES.

De Fornel & Léon rapportent un cas intéressant : vers le milieu des années 1950, une séquence filmique de 18 secondes fut soumise à l'analyse de chercheurs en interaction. Cette même scène fut l'objet d'études de plusieurs analystes et, malgré la très courte durée de la scène, des années furent nécessaires pour en exploiter tout le potentiel :

> Pour la première fois en effet, plusieurs chercheurs se sont livrés à l'analyse du même fragment de données enregistrées. Cette première expérience sera ensuite répétée pour

> devenir partie intégrante de la méthode d'analyse des données enregistrées, d'ailleurs adoptée par l'analyse de conversation. (De Fornel & Léon, 2000 : 138)

C'est dans une telle perspective que le projet CIL a vu le jour et c'est un destin similaire à celui de la séquence filmique que nous lui espérons.

Références bibliographiques

Aarts Jan, « Does corpus linguistics exist? Some old and new issues », *Language and Computers* 40 (1), 2002, pp. 1–17.

Aarts Jan et Meijs Willem, *Corpus linguistics: new studies in the analysis and exploitation of computer compora*, vol. 47, Amsterdam, Rodopi, 1984.

Abeillé Anne *et al.*, « Un corpus français arboré: quelques interrogations », *Actes de Traitement Automatique des Langues Naturelles* 1, 2001, pp. 32‑43.

Abney Steven, « Statistical methods and linguistics », *The balancing act: Combining symbolic and statistical approaches to language*, 1996, pp. 1–26.

Abry Dominique, Boë Louis-Jean et Rakotofiringa Hippolyte, « Théodore Rosset et l'établissement de la phonétique expérimentale à Grenoble. », *Documents pour l'histoire du français langue étrangère ou seconde* (20), 1997, pp. 54‑76.

Aichele Dieter, « Quantitative Linguistik in Deutschland und Österreich », *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (16), 2005, pp. 16‑23.

Aijmer Karin et Altenberg Bengt, *English corpus linguistics: studies in honour of Jan Svartvik*, Londres, Longman Publishing Group, 1991.

ALPAC, « Language and Machines: Computers in Translation and Linguistics », 1416, 1966, <http://www.nap.edu/openbook.php?record_id=9547>, .

Anctil Dominic, *L'erreur lexicale au secondaire: analyse d'erreurs lexicales d'élèves de 3e secondaire et description du rapport à l'erreur lexicale d'enseignants de français*, Thèse, Montréal, Université de Montréal, 2010.

Anis Jacques, « Ordinateurs et traduction : survol d'un demi-siècle », *Langages* 28 (116), 1994, pp. 111‑122.

Antoniadis Georges et Chanier Thierry, « Tal (Traitement automatique des langues) et apprentissage des langues », *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication* 8 (2), 2005.

Archaimbault Sylvie et Léon Jacqueline, « La langue intermédiaire dans la traduction automatique en URSS (1954-1960) », *Histoire épistémologie langage* 19 (2), 1997, pp. 105–132.

Armstrong Nigel *et al.*, *La langue française au féminin: le sexe et le genre affectent-ils la variation linguistique?*, Paris, L'Harmattan, 2001.

Aston Guy, *Learning with corpora*, Houston, Athelstan, 2001.

Aston Guy, « The British National Corpus as a language learner resource », *in Proceedings of*

Teaching and Language Corpora, Lancaster, UCREL, 1996, pp. 178‑191.

Auchlin Antoine, « Mais heu, pis bon, ben alors voilà, quoi! Marqueurs de structuration de la conversation et complétude », *Cahiers de linguistique française* 2 (1), 1981, pp. 141–159.

Auroux Sylvain, *La raison, le langage et les normes*, Paris, PUF, 1998.

Auroux Sylvain, *La révolution technologique de la grammatisation. Introduction à l'histoire des sciences du langage*, Bruxelles, Mardaga, 1994.

Baayen Harald, *Analyzing linguistic data: A practical introduction to statistics using R*, Cambridge, Cambridge University Press, 2008.

Balthasar Lukas et Bert Michel, « La plateforme "Corpus de langues parlées en interaction" (CLAPI). Historique, état des lieux, perspectives. », *Lidil. Revue de linguistique et de didactique des langues* (31), 2005, pp. 13–33.

Bar-Hillel Yehoshua, « A demonstration of the nonfeasibility of fully automatic high quality translation », *Advances in computers* 1, 1960, pp. 158–163.

Bar-Hillel Yehoshua, « A quasi-arithmetical notation for syntactic description », *Language* 29 (1), a 1953, pp. 47–58.

Bar-Hillel Yehoshua, « Some linguistic problems connected with machine translation », *Philosophy of science* 20 (3), b 1953, pp. 217–225.

Baroni Marco et Kilgarriff Adam, « Large linguistically-processed web corpora for multiple languages », *in Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, [s. l.], Association for Computational Linguistics, 2006, pp. 87–90.

Bartning Inge, « L'apprenant dit avancé et son acquisition d'une langue étrangère. Tour d'horizon et esquisse d'une caractérisation de la variété avancée », *Acquisition et interaction en langue étrangère* (9), 1997, pp. 9–50.

Bartning Inge et Schlyter Suzanne, « Itinéraires acquisitionnels et stades de développement en français L2 », *Journal of French language studies* 14 (3), 2004, pp. 281–299.

Bauche Henri, *Le langage populaire*, Paris, Payot & Cie, 1920.

Baude O. *et al.*, *Corpus oraux, guide des bonnes pratiques 2006*, Paris, CNRS, 2006.

Baude Olivier et Abouda Lotfi, « Constituer et exploiter un grand corpus oral: choix et enjeux théoriques. Le cas des ESLO », *in Corpus en lettres et sciences sociales : des documents numériques à l'interprétation*, Paris, 2006, pp. 143‑150.

Bazillon Thierry, *Transcription et traitement manuel de la parole spontanée pour sa reconnaissance automatique*, Thèse, Le Mans, Université du Maine, 2011.

Béjoint Henri, « Informatique et lexicographie de corpus: les nouveaux dictionnaires », *Revue française de linguistique appliquée* 12 (1), 2007, pp. 7–23.

Benveniste Émile, « Nature du signe linguistique », *Acta Linguistica I*, 1939.

Benzitoun Christophe *et al.*, « Tu veux couper là faut dire pourquoi. Propositions pour une segmentation syntaxique du français parlé », *in 2ème Congrès Mondial de Linguistique Française*, [s. l.], EDP Sciences, 2010.

Bergounioux Gabriel, Baraduc Jean et Dumont Céline, « L'étude sociolinguistique sur Orléans (1966-1991): 25 ans d'histoire d'un corpus », *Langue française* 93 (1), 1992, pp. 74–93.

Bernardini Silvia, « Corpora in the classroom », *How to use corpora in language teaching* 12, 2004, p. 15.

Berrendonner A. et Reichler-Béguelin M. J., « Décalages: les niveaux de l'analyse linguistique », *Langue française* 81 (1), 1989, pp. 99–125.

Bertrand Roxane *et al.*, « Représentation, édition et exploitation de données multimodales: le cas des backchannels du corpus CID », *Cahiers de Linguistique.* 33 (2), 2009, pp. 183–212.

Biber Douglas, « Representativeness in corpus design », *Literary and linguistic computing* 8 (4), 1993, pp. 243–257.

Biber Douglas, « Using register-diversified corpora for general language studies », *Computational linguistics* 19 (2), 1993, pp. 219–241.

Biber Douglas, Conrad Susan et Reppen R., *Corpus linguistics: investigating language structure and use*, Cambridge, Cambridge University Press, 1998.

Biber Douglas et Finegan Edward, « On the exploitation of computerized corpora in variations studies », *in English corpus linguistics: studies in honour of Jan Svartvik*, Londres, Longman Publishing Group, 1991, pp. 204‑220.

Bilger Mireille, *Données orales: les enjeux de la transcription*, Perpignan, Presses universitaires de Perpignan, 2008.

Bilger Mireille *et al.*, « Transcription de l'oral et interprétation. Illustration de quelques difficultés », *Recherches sur le français parlé* (14), 1997, pp. 57–86.

Billières Michel et Gaillard Pascal, « Approche pluridisciplinaire de la perception de la parole », *in Données orales : les enjeux de la transcription*, Perpignan, Presses universitaires de Perpignan, 2008, pp. 173‑192.

Blanche-Benveniste Claire, *Approches de la langue parlée*, Paris, Ophrys, 2010.

Blanche-Benveniste Claire, « L'étude grammaticale des corpus de langue parlée en français », *Willians, G. (2005)*, 2005, pp. 47–66.

Blanche-Benveniste Claire, « Transcriptions et technologies », *Recherches sur le français parlé* (14), 1997, pp. 87–99.

Blanche-Benveniste Claire et Jeanjean Colette, *Le français parlé*, Paris, CNRS, 1987.

Blanc Olivier *et al.*, « Corpus oraux et chunking », *in 27èmes Journées d'Études sur la Parole*, Avignon, JEP'08, 2008.

Bloom Lois Masket, « Language Development: Form and Function in Emerging Grammars. », *University Microfilms International*, 1968.

Boulton Alex, « Esprit de corpus: Promouvoir l'exploitation de corpus en apprentissage des langues. », *Texte et Corpus.* (3), 2007, pp. 37–46.

Bourdoux Françoise *et al.*, « Guide d'utilisation de CLAN pour le projet Léonard », 2011.

Braine Martin, « The ontogeny of English phrase structure: The first phase », *Language*, 1963, pp. 1–13.

Branca-Rosoff Sonia *et al.*, « Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000) », *CFPP2000*, 2009.

Brown Roger, *A first language: The early stages.*, Cambridge, Harvard University Press, 1973.

Brunot Ferdinand, « Discours d'inauguration des Archives de la parole / Discours de M. Brunot », Université de Paris, Archives de la parole, 1911.

Bully Philippe, « Zipf, créateur de la linguistique statistique », *Communication et langages* 2 (1), 1969, pp. 23–28.

Burnard Lou, « Une introduction au British National Corpus dans son édition XML », *Texte et Corpus* (3), 2007, pp. 17‑34.

Burnard Lou, *Users Reference Guide British National Corpus*, Oxford, University Computing Service, 1995.

Burnard Lou et Baumann Syd, *TEI P5: Guidelines for Electronic Text Encoding and Interchange, 1.8. 0*, [s. l.], TEI Consortium, 2010.

Butler Christopher, *Statistics in linguistics*, New York, Blackwell, 1985.

Cailliau Frederik et Poudat Céline, « Caractérisation lexicale des contributions clients agents dans un corpus de conversations téléphoniques retranscrites », *in Proceedings of JADT*, 2008, pp. 267–275.

Callison-Burch Chris et Dredze Mark, « Creating speech and language data with Amazon's Mechanical Turk », *in Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 1–12.

Campione Estelle et Véronis Jean, « Etiquetage prosodique semi-automatique des corpus oraux », *in Actes de la conférence Traitement Automatique des Langues (TALN 2001)*, 2001, pp. 123–132.

Candea Maria, *Contribution à l'étude des pauses silencieuses et des phénomènes dits'd'hésitation'en français oral spontané. Etude sur un corpus de récits en classe de français.*, Thèse, Paris, Université Paris III, 2000.

Cappeau Paul, « Données erronées: quelles erreurs commettent les transcripteurs? », *Recherches sur le français parlé* (14), 1997, pp. 117–126.

Cappeau Paul et Gadet Françoise, « Transcrire c'est (déjà) analyser », *Travaux linguistiques du Cerlico* 25, 2012.

Cappeau Paul et Gadet Françoise, « Transcrire, ponctuer, découper l'oral. Bien plus que de simples choix techniques », *Cahiers de linguistique* (35/1), 2010, pp. 187–202.

Cappeau Paul et Gadet Françoise, « L'exploitation sociolinguistique des grands corpus. », *Revue française de linguistique appliquée* 12 (1), 2007, pp. 99–110.

Cappeau Paul et Gadet Françoise, « Où en sont les corpus sur les français parlés? », *Revue française de linguistique appliquée* 12 (1), 2007, pp. 129–133.

Cappeau Paul et Seijido Magalie, « Les corpus oraux en français », 2005.

Carton Francis, « L'apprentissage différencié des quatre aptitudes », *in Verbum: Didactique du Français Langue Étrangère*, Nancy, Presses universitaires de Nancy, 1995, pp. 63‑74.

Catach Nina, « La ponctuation », *Langue française* 45 (1), 1980, pp. 16–27.

Chambers Angela, « Les corpus oraux en français langue étrangère: authenticité et pédagogie », *Mélanges CRAPEL* 31, 2009, pp. 15–33.

Chambers Angela, « Integrating corpus consultation in language studies », *Language learning & technology* 9 (2), 2005, pp. 111–125.

Charaudeau Patrick, « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique », *Corpus* (8), 2009, pp. 37–66.

Chen Stanley F., « Aligning sentences in bilingual corpora using lexical information », *in Proceedings of the 31st annual meeting on Association for Computational Linguistics*, 1993, pp. 9–16.

Chevalier Jean-Claude et Encrevé Pierre, « La création de revues dans les années 60: matériaux pour l'histoire récente de la linguistique en France », *Langue française* 63 (1), 1984, pp. 57–102.

Chomsky Noam, « The master and his performance. Interview by Jozsef Andor », *Intercultural Pragmatics* 1 (1), 2004, pp. 93–111.

Chomsky Noam, *New Horizons in the Study of Language and Mind*, Cambridge, Cambridge University Press, 2000.

Chomsky Noam, *Generative grammar: its basis, development and prospects*, Kyoto, Gaikokugo Daigaku, 1988.

Chomsky Noam, *Aspects of the Theory of Syntax*, Cambridge, The MIT press, 1965.

Chomsky Noam, *Syntactic structures*, Berlin, Walter de Gruyter, 1957.

Chomsky Noam et Halle Morris, « Some controversial questions in phonological theory », *Journal of linguistics* 1 (2), 1965, pp. 97–138.

Clear Jeremy, « Corpus sampling », *New directions in English language corpora*, 1992, pp. 21–31.

Cobb Tom, « Is there any measurable learning from hands-on concordancing? », *System* 25 (3), 1997, pp. 301–315.

Cohen Jacob, « A coefficient of agreement for nominal scales », *Educational and psychological measurement* 20 (1), 1960, pp. 37–46.

Coppieters René, « Quelques réflexions sur la question des données: corpus et intuitions », *Recherches sur le français parlé* (14), 1997, pp. 21–41.

Corbin Pierre, « De la production des données en linguistique introspective », *Théories linguistiques et traditions grammaticales*, 1980, pp. 121–179.

Cori Marcel et David Sophie, « Les corpus fondent-ils une nouvelle linguistique? », *Langages* (3), 2008, pp. 111–129.

Cori Marcel et Léon Jacqueline, « La constitution du TAL. Etude historique des dénominations et des concepts », *Traitement Automatique des Langues* 43 (3), 2002, pp. 21–57.

Cowie Anthony Paul, *English dictionaries for foreign learners: A history*, Oxford, Oxford University Press, 1999.

Croft William, « La théorie de la typologie fonctionnelle dans son contexte historique et intellectuel », *Verbum* (3), 1998, pp. 289–307.

Crystal David et Quirk Randolph, *Systems of prosodic and paralinguistic features in English*, The Hague, Walter de Gruyter, 1964.

Dagiral Éric et Parasie Sylvain, « Vidéo à la une! L'innovation dans les formats de la presse en ligne », *Réseaux* (2), 2010, pp. 101–132.

Damourette Jacques et Pichon Édouard, *Des mots à la pensée: essai de grammaire de la langue française*, Paris, D'Artrey, 1930.

Darwin Charles, « A biographical sketch of an infant », *Mind* 2 (7), 1877, pp. 285–294.

Davies Mark, « The 385+ million word Corpus of Contemporary American English (1990-2008): Design, architecture, and linguistic insights », *International Journal of Corpus Linguistics* 14 (2), 2009, pp. 159–190.

Davis K. H., Biddulph R. et Balashek S., « Automatic recognition of spoken digits », *The Journal of the Acoustical Society of America* 24 (6), 1952, pp. 637–642.

Debaisieux Jeanne-Marie, « Des documents authentiques oraux aux corpus: un défi pour la didactique du FLE », *Mélanges CRAPEL* 31, 2009, pp. 36–56.

De Beaugrande Robert, « Large corpora and applied linguistics: HG Widdowson versus J. McH. Sinclair », 2000, <http://www.beaugrande.com/WiddowSincS.htm>, .

De Beaugrande Robert, « Theory and practice in applied linguistics: disconnection, conflict, or dialectic? », *Applied Linguistics* 18 (3), 1997, pp. 279–313.

De Fornel Michel et Léon Jacqueline, « L'analyse de conversation, de l'ethnomethodologie à la linguistique interactionnelle », *Histoire épistémologie langage* 22 (1), 2000, pp. 131–155.

Déjean Hervé et Gaussier Éric, « Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables », *Lexicometrica*, *Alignement lexical dans les corpus multilingues*, 2002, pp. 1–22.

Delais-Roussarie Élisabeth, « Les annotations des données orales », *in Données orales : les enjeux de la transcription*, Perpignan, Presses universitaires de Perpignan, 2008, pp. 156‑172.

Delais-Roussarie Élisabeth, *Constituer des corpus oraux: méthodes et outils*, Paris, CNRS, 2002.

Delais-Roussarie Élisabeth et Durand Jacques, *Corpus et variation en phonologie du français: méthodes et analyses*, Toulouse, Presses universitaires du Mirail, 2003.

De L'Épée Charles-Michel, *Institution des sourds et muets par la voie des signes méthodiques*, Paris, Nyon l'Aîné, 1776.

DELIC, « Présentation du Corpus de référence du français parlé », *Recherches sur le français parlé* 18, 2004, pp. 11–42.

Descamps Florence, Monnier François et Schnapper Dominique, *L'historien, l'archiviste et le magnétophone: de la constitution de la source orale à son exploitation*, Paris, Comité pour l'histoire économique et financière de la France, 2005 (Histoire économique et financière de la France).

Detey Sylvain *et al.*, « Ressources phonologiques au service de la didactique de l'oral: le projet PFC-EF », *Mélanges CRAPEL* 31, 2009, pp. 223–236.

Dister Anne et Simon Anne-Catherine, « La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé », *Arena Romanistica* 1 (1), 2008, pp. 54–79.

Doca Gheorghe, *Analyse psycholinguistique des erreurs faites lors de l'apprentissage d'une langue étrangère: applications au domaine franco-roumain*, Paris, Publications de la Sorbonne, 1981.

EAGLES, « Preliminary recommendations on spoken texts », *Expert Advisory Group on Language Engineering Standards*, 1996, <http://www.ilc.cnr.it/EAGLES/spokentx/spokentx.html>, .

EAGLES, « Recommendations for the Morphosyntactic Annotation of Corpora », *Expert Advisory Group on Language Engineering Standards*, 1996, <http://www.ilc.cnr.it/EAGLES/annotate/annotate.html>, .

Eaton Helen, *Semantic frequency list for English, French, German, and Spanish: a correlation of the first six thousand words in four single-language frequency lists*, Chicago, The University of Chicago Press, 1940.

Ellis Alexander J., « The existing phonology of English dialects », *Early English Pronunciation*, 1889.

Ernst Gerhard, *Gesprochenes Französisch zu Beginn des 17. Jahrhunderts: Direkte Rede in Jean Héroards « Histoire particulière de Louis XIII » (1605-1610)*, [s. l.], Niemeyer, 1985.

Evanini Keelan, Higgins Derrick et Zechner Klaus, « Using Amazon Mechanical Turk for transcription of non-native speech », *in Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, [s. l.], Association for Computational Linguistics, 2010, pp. 53–56.

Evanini Keelan et Zechner Klaus, « Using crowdsourcing to provide prosodic annotations for non-native speech », *in Twelfth Annual Conference of the International Speech Communication Association*, 2011, pp. 3069‑3072.

Faucett Lawrence *et al.*, « Interim report on vocabulary selection », *London: PS King*, 1936.

Fillmore Charles, « Corpus linguistics or computer-aided armchair linguistics », *in Directions in Corpus Linguistics: Proceedings of Nobel Symposium*, 1991, pp. 35–60.

Fischer Johann, « Une approche actionnelle dans l'enseignement des langues à travers les simulations globales et les études de cas », *Exploiting Internet Case Studies and Simulation Templates for Language Teaching and Learning* (8), 2009.

Fleiss Joseph L. et Cohen Jacob, « Measuring nominal scale agreement among many raters », *Psychological bulletin* 76 (5), 1971, p. 378.

Fligelstone Steve, « Some reflections on the question of teaching, from a corpus linguistics perspective », *ICAME Journal* 17, 1993, pp. 97–109.

Francis Nelson et Kucera Henry, *Computational analysis of present-day American English*, Providence, Brown University Press, 1967.

Francis W. Nelson, « Language corpora BC », *in Directions in Corpus Linguistics: Proceedings of Nobel Symposium*, [s. l.], Walter de Gruyter, 1991, pp. 17–31.

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

Fries Charles C., « Teaching and learning English as a Foreign Language. », 1945.

Fries Peter H., « Charles C. Fries, linguistics and corpus linguistics », *ICAME Journal* (34), 2008, pp. 89‑119.

Fuchs Catherine *et al.*, *Linguistique et traitements automatiques des langues*, Paris, Hachette, 1993.

Gadet Françoise, « L'œil et l'oreille à l'écoute du social », *in Données orales. Les enjeux de la transcription*, [s. l.], M. Bilger (ed), Presses universitaires de Perpignan, 2008, pp. 35‑48.

Galazzi Enrica, « Phonétique/Université/Enseignement à la fin du XIXe siècle », *Histoire épistémologie langage* 17 (1), 1995, pp. 95–114.

Gale William et Church Kenneth, « A program for aligning sentences in bilingual corpora », *Computational linguistics* 19 (1), 1993, pp. 75–102.

Gendner Véronique et Adda-Decker Martine, « Analyse comparative de corpus oraux et écrits français: mots, lemmes et classes morpho-syntaxiques », *Actes des 24èmes Journées d'Études sur la Parole (JEP), Nancy, France*, 2002, pp. 13‑16.

Geyken Alexander, « Quelques problèmes observés dans l'élaboration de dictionnaires à partir de corpus », *Langages* (3), 2008, pp. 77–94.

Gilliéron Jules et Edmont Edmond, *Atlas linguistique de la France*, vol. 2, Paris, Honoré Champion, 1902.

Gilmore Alex, « Authentic materials and authenticity in foreign language learning », *Language Teaching* 40 (02), 2007, pp. 97–118.

Giovannoni Dominique-Catherine et Savelli Marie-Josée, « Transcrire et orthographier le français parlé. De l'impossible copie à la falsification des données orales », *Recherches sur le français parlé* 10, 1990, pp. 19–37.

Goody Jack, *La raison graphique. La domestication de la pensée sauvage*, Paris, Éditions de Minuit, 1979.

Gougenheim Georges, Rivenc Paul et Sauvageot Aurélien, *L'élaboration du français élémentaire*, Paris, Didier, 1956.

Gougenheim Henri, « L'observation du langage d'un enfant royal au XVIIe siècle », *Revue de philologie française et romane* 43, 1931, pp. 1‑15.

Gouin François, *L'art d'enseigner et d'étudier les langues*, Paris, G. Fichbacher, 1880.

Granger Sylviane, « Corpus d'apprenants, annotation d'erreurs et ALAO: une synergie prometteuse », *Cahiers de lexicologie* 91, 2007, p. 117.

Granger Sylviane, « Uses of tenses by advanced EFL learners: evidence from an error-tagged computer corpus », *Language and Computers* 26, 1999, pp. 191–202.

Granger Sylviane, Hung Joseph et Petch-Tyson Stephanie, *Computer learner corpora, second language acquisition, and foreign language teaching*, Amsterdam, John Benjamins Publishing Company, 2002.

Granger Sylviane et Monfort Guy, « La description de la compétence lexicale en langue étrangère : perspectives méthodologiques », *Acquisition et interaction en langue étrangère* (3), 1994, pp. 55‑75.

Granger Sylviane et Tribble Chris, « Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning », *Learner English on computer*, 1998, pp. 199‑209.

Gross Maurice, « Notes sur l'histoire de la traduction automatique », *Langages* 7 (28), 1972, pp. 40‑48.

Gruenstein Alexander, McGraw Ian et Sutherland Andrew, « A self-transcribing speech corpus: collecting continuous speech with an online educational game », *in SLaTE Workshop*, 2009.

Guiraud Pierre, *Bibliographie critique de la statistique linguistique*, Utrecht, Spectrum, 1954.

Habert Benoît, « Des corpus représentatifs: de quoi, pour quoi, comment », *Linguistique sur corpus. Etudes et réflexions* (31), 2000, pp. 11–58.

Habert Benoît, Nazarenko Adeline et Salem André, *Les linguistiques de corpus*, Paris, Armand Colin, 1997.

Harris Zellig S., « Transfer grammar », *International Journal of American Linguistics* 20 (4), 1954, pp. 259–270.

Hartmann Reinhard Rudolf Karl, *The history of lexicography: papers from the Dictionary Research Centre Seminar at Exeter*, Amsterdam, John Benjamins Publishing Company, 1986.

Héroard Jean et Foisil Madeleine, *Journal de Jean Héroard, médecin de Louis XIII*, Paris, Fayard, 1989.

Hoek Léo, *La marque du titre*, Berlin, De Gruyter Mouton, 1981.

Hornby Albert Sydney, Gatenby Edward Vivian et Wakefield A. H., *Idiomatic and syntactic English dictionary*, [s. l.], Institute for Research in Language Teaching, 1942.

Howe Jeff, « The rise of crowdsourcing », *Wired magazine* 14 (6), 2006, pp. 1–4.

Hunston Susan, *Corpora in applied linguistics*, Cambridge, Cambridge University Press, 2002.

Hutchins John, « Current commercial machine translation systems and computer-based translation tools: system types and their uses », *International Journal of Translation* 17 (1)‑(2), 2005, pp. 5–38.

Hutchins John, « The history of machine translation in a nutshell », *Retrieved December* 20, 2005.

Hutchins John, « ALPAC: the (in) famous report », *Readings in machine translation*, 2003, p. 131.

Hutchins John et Somers Harold, *An introduction to machine translation*, New York, Academic Press, 1992.

Ide Nancy, Reppen Randi et Suderman Keith, « The American National Corpus », *in Proceedings of the Fourth International Language Resources and Evaluation Conference (LREC)*, Lisbonne, 2002, pp. 1681‑1684.

Ingram David et Le Normand Marie-Thérèse, « A diary study on the acquisition of Middle French: A preliminary report of the early language acquisition of Louis XIII », *in Proceedings of the 20th annual Boston University conference on language development*, 1996, pp. 352‑63.

James Carl, *Errors in language learning and use: Exploring error analysis*, Londres, Routledge, 2013.

Järvinen Timo, « Annotating 200 million words: the Bank of English project », *in Proceedings of the 15th conference on Computational linguistics*, 1994, pp. 565–568.

Johansson Stig et Hofland Knut, « Towards an English-Norwegian parallel corpus », *Creating and using English language corpora*, 1994, pp. 25–37.

Johns Tim, « Should you be persuaded: Two samples of data-driven learning materials », *English language research journal* 4, 1991, pp. 1–16.

Kaeding Friedrich Wilhelm, *Häufigkeitswörterbuch der deutschen Sprache*, [s. l.], Selbstverlag des Herausgebers; ES Mittler & Sohn, 1898.

Karlgren Bernhard, *The Early History of the Chou Li and Tso Chuan Texts*, [s. l.], HW Tullbergs, 1932.

Kennedy Graeme, *An introduction to corpus linguistics*, Londres, Longman Publishing Group, 1998.

Kennedy Graeme, « Preferred ways of putting things with implications for language teaching », *in Directions in corpus linguistics. Proceedings of Nobel Symposium*, 1992, pp. 335–378.

Kilgarriff Adam, « Comparing corpora », *International journal of corpus linguistics* 6 (1), 2001, pp. 97–133.

Kilgarriff Adam et Grefenstette Gregory, « Introduction to the special issue on the web as corpus », *Computational linguistics* 29 (3), 2003, pp. 333–347.

Kilgarriff Adam et Tugwell David, « Sketching words », *Lexicography and natural language processing: a festschrift in honour of BTS Atkins*, 2002, pp. 125–137.

Koehn Philipp, « Europarl: A parallel corpus for statistical machine translation », *in MT summit*, 2005.

Labov William, « How I got into linguistics, and what I got out of it », *Historiographia Linguistica* 28 (3), 2001, pp. 455–466.

Labov William, « Qu'est-ce qu'un fait linguistique? », *Marges linguistiques* 1, 2001, pp. 1–44.

Labov William, *What is a linguistic fact?*, Lisse, Peter de Ridder Press, 1975.

Labov William, *Language in the inner city: Studies in the Black English vernacular*, vol. 3, Philadelphie, University of Pennsylvania Press, 1972.

Labov William, « The Social Stratification of (R) in New York City Department Stores », *Sociolinguistics Patterns*, 1972, pp. 43‑54.

Labov William, « The Logic of Non-Standard English. », *Georgetown Monographs on Languages and Linguistics* (22), 1969.

Labov William, *The social stratification of English in New York City*, Washington, Center for Applied Linguistics, 1966.

Laks Bernard, « Pour une phonologie de corpus », *Journal of French Language Studies* 18 (01), 2008, pp. 3–32.

Laks Bernard, « Langage et pratiques sociales », *Actes de la recherche en sciences sociales* 46 (1), 1983, pp. 73–97.

Landau Sidney, *Dictionaries: The Art and Craft of Lexicography*, Cambridge, Cambridge University Press, 2001.

Landis J. Richard et Koch Gary G., « The measurement of observer agreement for categorical data », *Biometrics*, 1977, pp. 159–174.

Landragin Frédéric, « Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits », *Corpus* 10, 2011, pp. 61‑80.

Landure Corinne et Boulton Alex, « Corpus et autocorrection pour l'apprentissage des langues. », *Asp* 57, 2010, pp. 11–30.

Laporte Éric, « Exemples attestés et exemples construits dans la pratique du lexique-grammaire », *in Observations et manipulations en linguistique : entre concurrence et complémentarité*, Paris, Peeters, 2008, pp. 11‑32.

Lebraty Jean-Fabrice, « Vers un nouveau mode d'externalisation: le crowdsourcing », *in 12ème conférence de l'AIM*, Lausanne, 2007.

Lebrun Yvan, « Le linguiste et les nombres », *Revue belge de philologie et d'histoire* 46 (3), 1968, pp. 771–778.

Lecherbonnier Bernard, *Pourquoi veulent-ils tuer le français?*, Paris, Albin Michel, 2005.

Leech Geoffrey, « New resources, or just better old ones? The Holy Grail of representativeness », *Language and Computers* 59 (1), 2006, pp. 133–149.

Leech Geoffrey, « Adding Linguistic Annotation », *Developing Linguistic Corpora: a Guide to Good Practice*, 2004.

Leech Geoffrey, « The state of the art in corpus linguistics », *in English corpus linguistics:*

studies in honour of Jan Svartvik, Londres, Longman Publishing Group, 1991, pp. 8‑29.

Leech Geoffrey, « Corpora and theories of linguistic performance », *Directions in Corpus Linguistics: Proceedings of Nobel Symposium*, 1991, pp. 105–122.

Lee Winnie Yuk-chun, « Authenticity revisited: Text authenticity and learner authenticity », *Elt Journal* 49 (4), 1995, pp. 323–328.

Legallois Dominique, « La colligation: autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique? », *Corpus* (11), 2012, pp. 31‑54.

Legallois Dominique et François Jacques, « La linguistique fondée sur l'usage: parcours critique », *Travaux de linguistique* (1), 2011, pp. 7–33.

Lemaire Benoît, « Limites de la lemmatisation pour l'extraction de significations », *in JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles*, 2008, pp. 725‑732.

Léon Jacqueline, « Aux sources de la «Corpus Linguistics» : Firth et la London School », *Langages* (3), 2008, pp. 12–33.

Léon Jacqueline, « Claimed and unclaimed sources of corpus linguistics », *Henry Sweet Society Bulletin* 44, 2005, pp. 36–50.

Léon Jacqueline, « Le CNRS et les débuts de la traduction automatique en France », *La revue pour l'histoire du CNRS* (6), 2002.

Léon Jacqueline, « Conceptions du'mot'et débuts de la traduction automatique », *Histoire épistémologie langage* 23 (1), 2001, pp. 81–106.

Léon Jacqueline, « Les débuts de la traduction automatique en France (1959-1968): à contretemps? », *Modèles Linguistiques* 19 (2), 1998, pp. 55–86.

Leroy Sarah et Muni Toke Valelia, « Une date dans la description linguistique du nom propre: l'Essai de grammaire de la langue française de Damourette et Pichon », *Actes des sessions de linguistique et de littérature* 27, 2007.

Loffler-Laurian Anne-Marie, *La traduction automatique*, Villeneuve d'Ascq, Presses universitaires Septentrion, 1996.

Luzzati Daniel, « Corpus, d'hier à aujourd'hui, progrès quantitatifs et progrès qualitatifs », *Cahiers de linguistique* 32 (2), 2009, pp. 97‑112.

MacWhinney B., « The CHILDES system », *American Journal of Speech-Language Pathology* 5 (1), 1996, p. 5.

MacWhinney Brian, « The CHILDES Project. Tools for Analyzing Talk – Electronic Edition. Part 1: The CHAT Transcription Format », 2014.

MacWhinney Brian, « The CHILDES Project. Tools for Analyzing Talk – Electronic Edition. Part 2: The CLAN Programs », 2014.

MacWhinney Brian et Snow Catherine, « The Child Language Data Exchange System », *Journal of Child Language* 12 (02), 1985, pp. 271‑295.

Manning Christopher D., « Probabilistic syntax », *Probabilistic linguistics*, 2003, pp. 289–341.

Manning Christopher et Schütze Hinrich, *Foundations of statistical natural language processing*, vol. 59, Cambridge, MIT Press, 1999.

Marcoccia Michel, « Les smileys: une représentation iconique des émotions dans la communication médiatisée par ordinateur », *Les émotions dans les interactions communicatives. Presses Universitaires de Lyon*, 2000, pp. 249–263.

Marge Matthew, Banerjee Satanjeev et Rudnicky Alexander J., « Using the Amazon Mechanical Turk for transcription of spoken language », *in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference*, 2010, pp. 5270–5273.

Martinet André, *La prononciation du français contemporain*, Genève, Droz, 1945.

Mauranen Anna, « Spoken corpus for an ordinary learner », *How to Use Corpora in Language Teaching. Amsterdam: Benjamins*, 2004, pp. 89–105.

Mayaffre Damon, « Rôle et place du corpus en linguistique. Réflexions introductives », *in Rôle et place des corpus en linguistique JETOU 2005*, Université de Toulouse-Le Mirail, 2005, pp. 5‑17.

McCarthy Dorothea, « Language development in children. », 1946.

McEnery Tony et Hardie Andrew, *Corpus linguistics: method, theory and practice.*, Cambridge, Cambridge University Press, 2011.

McEnery Tony et Wilson Andrew, *Corpus linguistics: an introduction*, Édimbourg, Edinburgh University Press, 2001.

McEnery Tony et Wilson Andrew, « Teaching and language corpora (TALC) », *ReCALL* 9 (01), 1997, pp. 5–14.

McEnery Tony et Xiao Richard, « What corpora can offer in language teaching and learning », *Handbook of research in second language teaching and learning. London: Routledge*, 2011, pp. 364–380.

McEnery Tony et Xiao Zhonghua, « Parallel and comparable corpora: What are they up to? », 2007.

Meunier-Crespo Mariette, « La constitution d'un corpus, parcours initiatique en linguistique », 2008.

Meyer Charles F., *English corpus linguistics: An introduction*, Cambridge, Cambridge University Press, 2002.

Meyer Charles F. et Nelson Gerald, « Data collection », *The handbook of English linguistics*,

2006, pp. 93–113.

Michel J. B. *et al.*, « Quantitative analysis of culture using millions of digitized books », *Science* 331 (6014), 2011, pp. 176–182.

Miller Wick et Ervin Susan, « The development of grammar in child language », *Monographs of the Society for Research in Child Development*, 1964, pp. 9–34.

Milner Jean-Claude, *Introduction à une science du langage*, Paris, Seuil, 1989.

Mindt Dieter, « English corpus linguistics and the foreign language teaching syllabus », *Using corpora for language research*, 1996, pp. 232–247.

Mondada Lorenza, « La transcription dans la perspective de la linguistique interactionnelle », *Données orales, les enjeux de la transcription*, 2008, pp. 78–109.

Mondada Lorenza, « Pour une linguistique interactionnelle », *Marges linguistiques* 1, 2001, pp. 142‑162.

Mondada Lorenza, « Les effets théoriques des pratiques de transcription », *Linx. Revue des linguistes de l'université Paris X Nanterre* (42), 2000, pp. 131–146.

Morgenstern Aliyah, *L'enfant dans la langue*, Paris, Presses Sorbonne Nouvelle, 2009.

Mougeon Raymond, Nadasdi Terry et Rehner Katherine, « État de la recherche sur l'appropriation de la variation par les apprenants avancés du FL2 ou FLE », *Acquisition et interaction en langue étrangère* (17), 2002, pp. 7–50.

Muni Toke Valelia, « Moi a pas mal du tout : le traitement des données orales "hors norme" par Damourette et Pichon (1930-1950) », (Dés)organisation de l'oral, Université Rennes 2, 2011.

Muni Toke Valelia, « La grammaire nationale selon Damourette et Pinchon: l'invention du locuteur », *Information grammaticale* (115), 2007, pp. 52–53.

Muni Toke Valelia et Habert Benoît, « Que faire des exemples singuliers? », *Vers une histoire générale de la grammaire française? matériaux et perspectives*, 2011, p. 69.

Murison-Bowie Simon, « Linguistics corpora and language teaching », *Annual review of applied linguistics* 16, 1996, pp. 182–199.

Murray K. M. Elisabeth, *Caught in the web of words*, New Haven, Yale University Press, 2001.

Niklas-Salminen Aïno, *La lexicologie*, Paris, Armand Colin, 2005.

Oakes Michael, *Statistics for corpus linguistics*, Édimbourg, Edinburgh University Press, 1998.

Ochs Elinor, « Transcription as theory », *Developmental pragmatics*, 1979, pp. 43–72.

Ogden Charles, *The system of basic English*, [s. l.], Harcourt, Brace and company, 1934.

Olson David R., *The world on paper*, [s. l.], Cambridge University Press, 1996.

Ooi Vincent, *Computer corpus lexicography*, [s. l.], Edinburgh University Press, 1998.

Pallaud Berthille, « Erreurs d'écoute dans la transcription de données orales », *Revue PArole* (22)‑(24), 2002, pp. 267–294.

Pallaud Berthille et Henry Sandrine, « Amorces de mots et répétitions: des hésitations plus que des erreurs en français parlé », *Le poids des mots*, 2003, pp. 848–858.

Palmer H.E., « Second interim report on English collocations », *Institute for Research in English Teaching, Tokyo*, 1933.

Paoli Bruno, « Du rôle fondateur d'al-Khalîl en métrique arabe », *Langues et Littératures du Monde Arabe* 7, 2007, pp. 1–11.

Parisse Christophe et Morgenstern Aliyah, « Transcrire et analyser les corpus d'interactions adulte-enfant », *Acquisition du langage et interaction*, 2010, pp. 201–222.

Pearson Jennifer, *Terms in context*, vol. 1, Amsterdam, John Benjamins publishing company, 1998.

Perry Marvin *et al.*, *Western Civilization: Ideas, Politics, and Society*, [s. l.], Houghton Mifflin Harcourt Publishing Company, 2008.

Péry-Woodley Marie-Paule, *Une pragmatique à fleur de texte: approche en corpus de l'organisation textuelle*, 2000.

Péry-Woodley Marie-Paule, « Quels corpus pour quels traitements automatiques? », *Traitement automatique des langues* 36 (1)‑(2), 1995, pp. 213–232.

Péry-Woodley M. P. *et al.*, « ANNODIS: une approche outillée de l'annotation de structures discursives », 2009.

Pinkham Jessie et Smets Martine, « Traduction automatique ancrée dans l'analyse linguistique », *in Proc. TALN*, 2002, pp. 287–296.

Preyer William, *Die Seele des Kindes, Beobachtungen über die geistige Entwicklung des Menschen in den ersten Lebensjahren Leipzig*, [s. l.], Grieben's Verlag, 1884.

Psutka J. *et al.*, « Automatic transcription of Czech language oral history in the MALACH project: Resources and initial experiments », *in Text, Speech and Dialogue*, 2006, pp. 219–234.

Puren Christian, « Histoire des méthodologies de l'enseignement des langues », 1988, <http://www.aplv-languesmodernes.org/IMG/pdf/puren_histoire_methodologies.pdf>, .

Quemada Bernard, *Les Dictionnaires du français moderne, 1539-1863: étude sur leur histoire, leurs types et leurs méthodes*, [s. l.], Didier, 1968.

Queneau Raymond, *Bâtons, chiffres et lettres*, [s. l.], Gallimard, 1965.

Quirk Randolph, « Towards a description of English usage », *Transactions of the Philological Society* 59 (1), 1960, pp. 40–61.

Quirk Randolph et Crystal David, *A comprehensive grammar of the english language*, Cambridge, Cambridge University Press, 1987.

Quirk Randolph *et al.*, *A grammar of contemporary English*, [s. l.], Oxford University Press, 1972.

Rastier François, « Enjeux épistémologiques de la linguistique de corpus », *G. Williams (éd.), La linguistique de corpus. Rennes: PUR*, 2005, pp. 31–45.

Rayson Paul et Garside Rayson, « Comparing corpora using frequency profiling », *in Proceedings of the workshop on Comparing Corpora*, 2000, pp. 1–6.

Riegel Martin, Pellat Jean-Christophe et Rioul René, *Grammaire méthodique du français*, Paris, Quadrige/PUF, 2009.

Römer Ute, « Corpora and language teaching », *Corpus Linguistics. An International Handbook* 1, 2008, pp. 112–130.

Rosset Théodore, « Du rôle de la phonétique dans l'enseignement des langues vivantes », *Les Langues Modernes*, 1909.

Sacks Harvey, *Lectures on conversation*, vol. 1, [s. l.], Blackwell, 1995.

Sacks H., Schegloff E. A. et Jefferson G., « A simplest systematics for the organization of turn-taking for conversation », *Language*, 1974, pp. 696–735.

Sankoff David, Lessard Rejean et Truong Nguyen Ba, « Computational linguistics and statistics in the analysis of the Montreal French Corpus », *Computers and the Humanities* 11 (4), 1977, pp. 185–191.

Saussure Ferdinand, « Cours de linguistique générale », *Payot, Lausanne-Paris*, 1916.

Schaeffer-Lacroix Eva, *Corpus numériques et production écrite en langue étrangère. Une recherche avec des apprenants d'allemand.*, 2009.

Scheer Tobias, « Le corpus heuristique: un outil qui montre mais ne démontre pas », *Corpus* (3), 2004.

Schmied Josef, « Second-language corpora », *Sidney Greenbaum (ed.)*, 1996, pp. 182–196.

Schweitzer Charles, *Enseignement direct de l'allemand, première année*, [s. l.], Armand Colin, 1904.

Selinker Larry, « Interlanguage », *IRAL-International Review of Applied Linguistics in Language Teaching* 10 (1)‑(4), 1972, pp. 209–232.

Shannon Claude, *Communication theory of secrecy systems*, [s. l.], AT & T, 1949.

Shin H. B. et Bruno R. R., « Language Use and English-Speaking Ability, 2000 », US

Department of Commerce, Economics and Statistics Administration, US Census Bureau, 2003.

Sinclair John, « Corpus and Text - Basic Principles », *Developing Linguistic Corpora: a Guide to Good Practice*, 2004, <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm>, .

Sinclair John, « Preliminary recommendations on corpus typology », *EAGLES*, 1996, <http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html>, .

Sinclair John, *Corpus, concordance, collocation*, [s. l.], Oxford University Press, 1991.

Sinclair John, « Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary », *Computational linguistics* (16), 1990, pp. 184‑186.

Sinclair John, « Uncommonly common words », *Learners' Dictionaries*, 1989, pp. 135–52.

Sinclair John *et al.*, *English collocation studies: The OSTI report*, [s. l.], Continuum Intl Pub Group, 2004.

Sinclair John McHardy, *How to use corpora in language teaching*, vol. 12, Amsterdam, John Benjamins Publishing Company, 2004.

Somers Harold et Jones Danny, « La génération de textes multilingues par un utilisateur monolingue », *Meta* 37 (4), 1992, pp. 647–656.

Sperberg-McQueen Michael, « The Text Encoding Initiative: Electronic Text Markup for Research », 1994, pp. 35‑55.

Steinberger Ralf *et al.*, « The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages », *Arxiv preprint cs/0609058*, 2006.

Stern William et Stern Clara, « Die Kindersprache », 19711, 1907.

Svartvik Jan, *Directions in corpus linguistics*, Berlin, Mouton de Gruyter, 1992.

Svartvik Jan, « Corpus linguistics comes of age Jan Svartvik », *in Directions in corpus linguistics: proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, 1992.

Taine Hippolyte, « Note sur l'acquisition du langage chez les enfants et dans l'espece humaine », *Revue philosophique* 1, 1876, pp. 5–23.

Teubert Wolfgang, « Corpus Linguistics: An Alternative », *Semen* (27), 2010.

Teubert Wolfgang, *Text corpora and multilingual lexicography*, Amsterdam, John Benjamins Publishing Company, 2007.

Teubert Wolfgang, « My version of corpus linguistics », *International Journal of Corpus Linguistics* 10 (1), 2005, pp. 1–13.

Teubert Wolfgang, « Corpus linguistics and lexicography », *International Journal of Corpus*

Linguistics, 2001, pp. 125–153.

Thorndike Edward, *The teacher's word book*, New York, Teachers College, Columbia University, 1921.

Thorndike Edward et Lorge Irving, *The teacher's word book of 30,000 words. New York: Teachers College, Columbia University*, [s. l.], Bureau of Publications, 1944.

Tognini-Bonelli Elena, *Corpus linguistics at work*, vol. 6, Amsterdam, John Benjamins Publishing Company, 2001.

Tsui Amy B M, « What teachers have always wanted to know - and how corpora can help », *How to use corpora in language teaching* 12, 2004, p. 39.

VanRullen Tristan, Blache Philippe et Portes Cristel, « Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales », *in Traitement Automatique des Langues Naturelles (TALN)*, 2005, pp. 41‑48.

Váradi Tamás, « The linguistic relevance of corpus linguistics », *in Proceedings of the Corpus Linguistics 2001 Conference. UCREL Technical Papers*, 2001, pp. 587–593.

Véronis Jean, « Annotation automatique de corpus: panorama et état de la technique », *Ingénierie des langues*, 2000, pp. 111–129.

West Michael, *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*, Londres, Longmans, Green, 1953.

West Michael, « Speaking vocabulary in a foreign language », *The Modern Language Journal* 14 (7), 1930, pp. 509–521.

Wichmann Anne *et al.*, *Teaching and language corpora*, Londres, Longman Publishing Group, 1997.

Widdowson Henry, « On the limitations of linguistics applied », *Applied linguistics* 21 (1), 2000, pp. 3–25.

Widdowson Henry G., « Context, Community, and Authentic Language », *TESOL quarterly* 32 (4), 1998, pp. 705–716.

Williams Geoffrey, « La linguistique de corpus: Une affaire préposition-nelle », *Texto*, 2006, pp. 151–158.

Williams Geoffrey, *La linguistique de corpus*, Rennes, Presses universitaires de Rennes, 2005.

Williams Geoffrey, « Sur les caractéristiques de la collocation », *Actes de TALN*, 2001, pp. 9–16.

Williams Geoffrey, *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique*, Thèse, Nantes, Université de Nantes, 1999.

Wright Joseph, *The English dialect grammar*, Oxford, Henry Frowde, 1905.

Wright Joseph, *The English Dialect Dictionary*, Oxford, Henry Frowde, 1898.

Xiao Richard, « Can corpora contribute to linguistic theory? », *in Handbooks of Linguistics and communication Science*, Berlin, Mouton de Gruyter, 2006.

Yngve Victor, « On getting a word in edgewise », *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, 1970, pp. 567‑577.

Yngve Victor, « Syntax and the problem of multiple meaning », *Machine translation of languages. John Wiley & Sons, New York*, 1955, pp. 208–226.

Zechner Klaus, « What did they actually say? agreement and disagreement among transcribers of non-native spontaneous speech responses in an English proficiency test. », *in Proceedings of the ISCA SLaTE-2009 Workshop*, 2009, pp. 25–28.

Zipf George, « Human behavior and the principle of least effort. », 1949.

Annexes

Annexe 1 : Installation de CLAN et utilisation du logiciel pour une consultation simplifiée

Le corpus est les fichiers annexes du corpus sont disponibles sur la clé USB jointe. Il s'agit de la version non anonymisée du corpus. Le contenu est le suivant :

- Dossier Logiciels (logiciels nécessaires au fonctionnement du corpus)
- Dossier Corpus (contenant les fichiers de transcription, d'entretien, de lecture et de productions écrites)
- Dossier Documentation (contenant les autorisations d'exploitation et les fiches signalétiques des apprenants)
- Dossier XML : contenant la version XML des transcriptions

Afin d'utiliser le corpus :

- Dans le dossier « Logiciels » : Installer clanwin.exe pour les utilisateurs de WINDOWS, clan.dmg pour les utilisateurs de Macs
- Installer le logiciel QuickTimeInstaller.exe pour les utilisateurs de WINDOWS

L'ensemble des fichiers qui concernent un apprenant sont nommés de manière similaire, par exemple :

- eng_ca_re_83_f_11.wav
- eng_ca_re_83_f_11_lecture.wav
- eng_ca_re_83_f_11.cha
- eng_ca_re_83_f_11.docx
- eng_ca_re_83_f_11.jpg

Le premier fichier est le fichier sonore de l'entretien, le second le fichier sonore de la tâche de lecture, le troisième le fichier de transcription, le quatrième la saisie sous WORD de la tâche de production écrite et le dernier un scan de la production écrite de l'apprenant.

Une fois le logiciel installé, afin d'écouter un transcript :

- Ouvrir le fichier de transcription eng_ca_re_83_f_11.cha
- Fermer la petite fenêtre de commandes si elle s'affiche
- Cliquer sur « Mode » dans la barre d'outils
- Cliquer sur « Continuous playback »

Pour rechercher un passage en particulier et le réécouter :

- Ouvrir le fichier de transcription
- Cliquer sur « Edit »

- Cliquer sur « Find »
- Rechercher un ou plusieurs termes du passage recherché
- Placer le curseur sur la ligne de transcription concernée
- Appuyer sur la touche F4 qui lira uniquement la ligne concernée

Afin d'installer le lexique français pour CLAN et lancer des commandes :

- Dans le dossier Logiciels, copier le dossier « fra », et le coller dans le dossier d'installation de CLAN, dans le dossier C://TALKBANK/CLAN
- Ouvrir un fichier CLAN
- Ouvrir la fenêtre de commandes : cliquer sur Window, puis sur Commands
- Indiquer dans « Working », le chemin d'accès au corpus
- Indiquer dans « Mor lib », le chemin d'accès au dossier « fra »

Les commandes peuvent alors être lancées dans la fenêtre.

Annexe 2: Texte de la lecture imposée aux apprenants et consignes pour la production écrite

Lire à haute voix le texte ci-contre. Avant d'être enregistré en train de lire le texte, vous pouvez le regarder un moment.

Le rendez-vous

Antoine veut me voir ; et moi je veux le voir aussi. C'est un ami d'enfance. Nous nous rencontrons une fois tous les trois mois. Rendez-vous, donc, pour ce samedi matin, 11 heures, dans un café du centre-ville.

Je n'aime pas arriver en avance à un rendez-vous, et surtout je n'aime pas attendre un homme dans un lieu public. Je suis en avance, tant pis. Je m'achète le journal et je m'installe à l'intérieur du café. Un garçon s'approche :

-Bonjour Madame, que désirez-vous ?

-Un thé, s'il vous plaît.

-Nature, citron ou au lait ?

-Nature, s'il vous plaît.

Le garçon m'apporte mon thé bien chaud. J'ouvre le journal à la page économique. Je décide de la lire toute entière avant de regarder ma montre. Il est 11 heures précises, j'ai le droit d'attendre officiellement. Le thé n'est pas mauvais, il fait beau dehors, et je suis contente de revoir la bonne tête d'Antoine.

À 11h10, la bonne tête d'Antoine a perdu quelques points.

À 11h20, je commence à m'impatienter.

"Mais où est-il ? Qu'est-ce qui a bien pu lui arriver ? " Mon thé est bu.

À 11h25 je hais Antoine. S'il arrivait à cet instant, je n'aurai rien à lui dire, sauf peut-être de s'en aller.

Il est déjà 11H30 et toujours pas d'Antoine. Ou bien Antoine a oublié, ou il a trouvé au dernier moment quelque chose de plus excitant à faire, personne ne m'aime.

Enfin j'aperçois une silhouette d'homme aux cheveux bouclés, avec un imperméable. Ça va mieux, je m'entends répondre aux excuses d'Antoine :

-Mais ça n'a aucune importance, raconte-moi plutôt comment tu vas.

Consignes pour la production écrite des apprenants

Relisez le texte « Un rendez-vous ».

1. De quoi parle-t-il ?
2. Imaginez sur une ou deux page(s) le dialogue entre Antoine et son amie.

Annexe 3: Autorisation d’exploitation signée par les apprenants

Autorisation d’exploitation scientifique

Les études sur les langues étrangères (menées dans des universités en France et à l'étranger) sont précieuses i) pour la connaissance du langage et de son acquisition, et ii) pour le développement de méthodes d'apprentissage des langues. Pour faire des progrès dans ces domaines, il est fondamental d'avoir accès à des productions d'apprenants diversifiées (entretiens oraux spontanés, lecture de textes, productions écrites). Ces productions peuvent ensuite être transcrites, analysées, et même diffusées, sous la forme d'enregistrements sonores, de transcriptions ou d'éditions numériques; mais cela se fera toujours en préservant l’anonymat des candidats.

Afin de pouvoir utiliser et diffuser les données collectées dans le cadre du Master Linguistique et Didactique des Langues ALE et FLE de l'Université de Rennes 2, nous vous demandons de signer une autorisation officielle.

Je soussigné(e) _____ autorise tout étudiant de Rennes 2, mais également tout universitaire français ou étranger :

1. à effectuer un enregistrement des productions sonores suivantes :
   a) un entretien oral entre l’enquêteur et moi-même
   b) une tâche de lecture
2. à saisir numériquement les productions écrites remises sous forme manuscrite
3. à transcrire orthographiquement les productions orales sonores ci-dessus mentionnées
4. à éditer à des fins scientifiques ou pédagogiques les transcriptions orthographiques, les données sonores et les productions écrites, en préservant mon anonymat.
5. à diffuser et à représenter les transcriptions orthographiques, les données sonores et les productions écrites, en préservant mon anonymat.

Signature

Date :

Nom :

Adresse :

Annexe 4: Fiche signalétique de l'apprenant

Nom Prénom :

Sexe :

Date de naissance :

Age :

Lieu de naissance (ville et pays) :

Domiciles successifs (si villes ou pays différents) en indiquant le nombre d'années passées à chaque endroit :

Domicile actuel :

Niveau d'instruction / études du locuteur (préciser jusqu'à quel âge il a été scolarisé, dans quel type de structure, etc.) :

Profession actuelle du locuteur :

Situation familiale personnelle :

Loisirs :

Profil linguistique 1 / Langue cible

Anglais Français

Niveau de compétence (auto-évaluation) :

- Oral : A1 A2 B1 B2 C1 C2
- Écrit : A1 A2 B1 B2 C1 C2

Profil linguistique 2 / Langue première

Langue première :

Précisez la variété de la langue première si cela est pertinent :

Profil linguistique 3 / Langue(s) étrangère(s) autre que la langue cible (auto-évaluation) :

Langue étrangère :

Niveau écrit :

Niveau oral :

Renseignements complémentaires, à remplir par l’enquêteur :

Date de l’enregistrement :

Durée approximative :

Lieu :

Annexe 5: Exemple de transcription

Transcripts de l'entretien ara_ge_fr_80_f_08

@Begin
@Languages: fra
@Participants: FLE Student, ENQ Investigator
@Options: bullets
@ID: fra|CIL|FLE|28;|male|arabe||Student||
@ID: fra|CIL|ENQ|||||Investigator||
@Media: ara_ge_fr_80_f_08 audio *ENQ: bonjour xxx.
*FLE: bonjour.
*ENQ: est+ce+que tu peux te présenter s'il+te+plaît nom prénom âge ton nom ton prénom ton âge où est+ce+que tu habites.
*FLE: euh moi je m' appelle xxx.
*FLE: euh j' ai euh vingt huit ans.
*FLE: j' habite à Damas Bab_Touma.
*FLE: euh je travaille dans un banque euh Byblos_Bank.
*FLE: et j' étudie maintenant la langue française au centre culturel.
*ENQ: tu étudies le français depuis combien de temps ?
*FLE: j' ai étudié depuis mon enfance la langue française à l' école.
*FLE: et j' ai continué euh quelques classes au centre.
*FLE: après ça j' ai commencé depuis deux mois euh la langue euh la français des affaires.
*FLE: et euh je veux continuer euh jusque mars.
*FLE: pour prendre un diplôme de euh de qu'est+ce+ que on dit &c pas centre du bureau &industri industriel et commercial euh de Paris. *ENQ: et tu apprends le français pour des raisons professionnelles ou euh?
*FLE: oui pour des raisons professionnelles.
*ENQ: parce+que tu travailles dans une banque c'est pour ça ? *FLE: oui c'est pour ça parce+que ya'ni@s:ara tu sais la matière euh le travail au banque euh.
*FLE: en France est très euh est mieux est qu'est+ ce+que je peux dire est très bien en monde ya'ni@s:ara.
*FLE: pour ça donc euh je pense de continuer mes études en français euh.
*ENQ: est+ce+que tu as déjà été en France ?
*FLE: non.
*ENQ: et comment est+ce+que tu pratiques la langue française tu rencontres des français ou tu regardes des films tu. *FLE: oui je [x 2] euh j' ai un ami euh français nous parlons euh toujours.
*FLE: et euh quelquefois je vois quelques programmes en français.
*FLE: je ya'ni@s:ara et de [x 2] mes études au centre.
*FLE: je [x 2] lis un+peu et.
*ENQ: quand tu regardes la télévision française tu peux comprendre de façon continue ?
*FLE: non pas de façon continuée ya'ni@s:ara ano@s:ara euh.
*FLE: pardon je [x 2] dis quelques mots ya'ni@s:ara.
*ENQ: c'est pas grave.
*FLE: pas de [x 2] ya'ni@s:ara je comprends pas tout mais je comprends par+exemple je comprends le sujet qu'est+ce+que ils parlent. *FLE: mais pas tous les mots mais en+général je [x 2] comprends qu'est+ce+qui se passe.
*ENQ: est+ce+que tu [x 2] rencontres des difficultés particulières dans la langue française que tu pourrais définir qu'est+ce+que tu trouves difficile ou facile dans cette langue?
*FLE: je trouve euh je pense que lirer est plus facile beaucoup plus facile de [x 2] voir euh un film ou un programme.
*FLE: de lire un livre ou de lire quelque+chose.
*FLE: et aussi de parler est plus facile beaucoup plus facile de voir un programme ou un film en français.
*ENQ: s' exprimer c'est plus facile que.
*FLE: oui ya'ni@s:ara quand je parle avec quelqu'un en français avec mon ami je [x 2] comprends ya'ni@s:ara à+peu+près tout mais quand je vois un film peut+être pas puisqu' il+y+a beaucoup des [x 2] mots des nouvelles mots et euh.
*FLE: en+plus que mon ami euh essayer essaye de [x 2] parler avec moi en langue facile.
*ENQ: est+ce+que tu connais d' autres langues ?
*FLE: oui l' anglais mais euh moins que euh la langue française.

*ENQ: tu l' as appris avant ou après le [x 2] français ?
*FLE: non non après le français.
*FLE: ya'ni@s:ara <j' ai commencé à> [x 2] apprendre des à l' école la langue anglais depuis le sixième.
*FLE: et après ça j' ai pris quelques cours au centre.
*FLE: mais la langue française j' ai commencé depuis les [x 2] petites classes.
*FLE: petit jardin et grand jardin et jusqu'à maintenant aussi.
*ENQ: est+ce+que tu utilises la langue française euh au travail ?
*FLE: euh quelquefois pas toujours quelquefois.
*ENQ: au au+sein+de de l' administration ou avec les clients ?
*FLE: non il+y+a quelques clients et avec euh ya'ni@s:ara il+y+a quelques emails.
*FLE: il+y+a quelques euh sujets que importants qu' on doit lire xxx pour l' éducation euh en français.
*ENQ: la langue principale utilisée reste l' arabe ?
*FLE: l' arabe oui.
*ENQ: et en seconde langue c'est plutôt le français ou l' anglais qu'est+ce+qui est le plus important dans ton milieu professionnel ?
*FLE: ici en Syrie c'est <l' anglais> [x 2] maintenant est [x 2] plus important.
*ENQ: même si c'est une société libanaise comme Byblos ? *FLE: oui oui mais euh je pense que euh l' esprit de [x 2] des banques libanaises est français ya'ni@s:ara.
*FLE: l' origine ya'ni@s:ara ils ont pris euh l' éducation française et euh ya'ni@s:ara il+y+a beaucoup qui [x 2] a étudié la [x 2] les l' étude de banque bancaire en France.
*FLE: des euh directeurs mais &ic &ic ici en Syrie c'est plus utilisé la langue anglais.
*ENQ: est+ce+que vous parlez dans votre famille toi et ta famille en français est+ce+que c'est un milieu francophone où il+y+a des membres ça veut dire est+ce+que il+y+a des membres de ta famille qui parlent français ou [x 2] pas ?
*FLE: oui [x 2] il+y+a on parle tout ya'ni@s:ara tous les membres de ma famille parlent pas mes pères ma mère et mon père.
*FLE: mais pour nous les euh mes frères et mes soeurs et moi nous avons étudié la langue française.
*ENQ: vous parlez le français entre vous ?
*FLE: non pas toujours j' ai essayé quelquefois de parler avec ma soeur pour euh être mieux mais euh.
*ENQ: pour améliorer.
*FLE: &pé pour améliorer oui mais &c ça ne marche pas elle n' a elle n' est pas marché euh.
*FLE: c'est difficile un+peu de parler.
*ENQ: ce n' est pas naturel.
*FLE: oui [x 2].
*ENQ: euh qu'est+ce+que tu as poursuivi comme études Georges ?
*FLE: euh j' ai étudié gestion [x 2].
*FLE: euh dans l' université d' économie et euh.
*ENQ: à Damas ?
*FLE: à Damas oui.
*FLE: je pense de continuer euh qu'est+ce+qu' on dit euh quelque+chose en finance [x 2] internationale.
*ENQ: tu as étudié le la gestion tu as terminé quand ?
*FLE: j' ai terminé en deux mille six.
*ENQ: et tu as tout+de+suite commencé à travailler ?
*FLE: oui après quelques mois trois ou quatre mois j' ai commencé à travailler au banque.
*ENQ: tu peux me rappeler ton âge j' ai pas.
*FLE: vingt huit.
*ENQ: tu penses euh après avoir obtenu le diplôme de gestion poursuivre et faire un diplôme aussi en France ou ?
*FLE: oui peut+être oui.
*ENQ: mais ce n' est pas donc uniquement pour euh pour la banque c'est aussi [x 2] pour un niveau personnel que tu veux peut+être continuer là+bas.
*FLE: oui.
*ENQ: tu [x 2] as vécu à Damas et à Beyrouth ?
*FLE: oui à Zahlé j' ai vécu à Zahlé oui et à Damas.
*ENQ: ce sont deux pays arabophones est+ce+que [x 2] tu peux nous parler des est+ce+que il+y+a une grande différence entre ces deux villes ou c'est la même mentalité les mêmes.
*FLE: non ya'ni@s:ara c'est ils utilisent la même langue mais la différence un+peu plus que les méthodes à l' école au Liban c'était en français.
*FLE: et la langue française au Liban surtout dans les écoles des soeurs ou dans les écoles chrétiennes est [x 2] euh plus importante que la langue anglaise.

*FLE: c'était quand nous étions à l' école maintenant la langue anglais est aussi très important ya'ni@s:ara ils ont commencé à utiliser. *FLE: mais ici euh en Syrie c'était la langue anglais ya'ni@s:ara la langue plus important et la langue.
*FLE: plus euh utilisée plus [x 2] enseignée on dit ?
*ENQ: plus enseignée oui.
*FLE: enseignée à l' école.
*ENQ: plus répandue <que la langue> [x 2] française.
*FLE: mais maintenant la langue française je pense que ya'ni@s:ara ici en Syrie ils ont commencé à donner la langue français comme [x 2] la langue anglais.
*FLE: oui euh mais pas depuis beaucoup de des années. *FLE: euh et c'était ya'ni@s:ara je pense une [x 2] <très important > [x 2] pas euh très important ya'ni@s:ara step@s:eng.
*ENQ: ah d'accord une [x 2] décision très importante.
*FLE: oui une décision très &im.
*ENQ: une démarche.
*FLE: une marche très important.
*ENQ: démarche.
*FLE: oui démarche très important.
*ENQ: tu as trouvé une différence de difficulté entre ces deux langues l' anglais le français qu'est+ce+qui est le plus facile le plus difficile ?
*FLE: pour moi la langue anglais la langue français pardon la langue xxx français est plus facile.
*FLE: la langue xxx est plus difficile parce+que tu sais quand euh tu étudies quelque+chose depuis l' enfance c'est plus facile. *FLE: et mais aussi la langue anglais est maintenant ya'ni@s:ara tu peux dire que à+peu+près toutes les films qu' on voit sont en anglais il+y+a beaucoup des choses qu' on l' utilise en [x 2] anglais. *FLE: c'est une langue très une langue internationale plus la langue que la langue française.
*FLE: mais pour moi je [x 2] vois que la langue française est plus facile pour moi.
*ENQ: parce+que tu as commencé.
*FLE: plus parce+que j' ai commencé plus jeune et oui. *ENQ: et pour un touriste étranger en Syrie est+ce+que un touriste francophone mais qui ne parle pas anglais pourrait trouver [x 2] facilement son chemin ou obtenir des renseignements dans la rue ?
*FLE: oui euh je n' ai pas compris.
*ENQ: si un touriste étranger qui est en Syrie <un touriste> [x 2] français qui ne parle pas anglais est+ce+qu' il peut rencontrer des gens dans la rue qui parlent français qui peuvent l' aider ou qui ou c'est plutôt l' anglais pour [x 2] obtenir des renseignements le [x 2] diriger dans sa route.
*FLE: c'est plutôt l' anglais mais il+y+a il+y+en+ a beaucoup aussi qui euh qui utilisent la langue français.
*ENQ: tu peux facilement rencontrer quelqu'un qui parle français dans la rue pour t' aider.
*FLE: oui xxx pas trop difficile.
*FLE: mais pas comme l' anglais tu sais la langue anglais est euh. *ENQ: quelle est ta [x 2] responsabilité exacte dans la banque qu'est+ce+que tu euh quel est ton travail c'est en c'est directement avec les clients ou tu travailles dans l' administration ?
*FLE: non je suis un guichetier on dit ?
*ENQ: oui.
*FLE: guichetier oui.
*FLE: depuis deux années euh à+peu+près et je travaille directement avec les clients.
*FLE: avec le cash avec euh.
*ENQ: tu as suivi une formation à Beyrouth ?
*FLE: oui pour &dep ya'ni@s:ara &pend j' ai suivi euh une un stage en Beyrouth.
*FLE: pendant trois mois après ça j' ai venu ici en Syrie et j' ai commencé oui.
*ENQ: et euh est+ce+que tu pourrais nous [x 2] parler <de la> [x 2] vie en Syrie en+général en+ce+qui+concerne la [x 2] vie le soir le les sorties les c'est en comparaison avec le Liban comme tu as tu connais les deux pays.
*FLE: maintenant la Syrie est [x 2] ya'ni@s:ara est [x 2] plus ya'ni@s:ara est &amé améliorée la vie de nuit on dit ya'ni@s:ara est plus est mieux.
*FLE: que euh avant parce+que il+y+en+a beaucoup des [x 2] clubs qui sont ouverts il+y+a beaucoup des discothèques des restaurants. *FLE: il+y+a et surtout tu sais quand il+y+a de [x 2] l' argent quand il+y+a de travail donc euh c'est [x 2] plus &f c'est très facile qu' il+y+a plus des [x 2] choses pour amuser.
*FLE: et pour ça ici en Syrie ils ont commencé par+exemple le [x 4] &deman domaine de banques ils ont commencé depuis cinq années euh pas depuis très longtemps donc euh maintenant est mieux. *FLE: beaucoup plus mieux donc si tu veux aller danser il+y+a beaucoup des discothèques.
*FLE: si tu veux aller dans un restaurant surtout dans les anciens maisons c'est très joli c'est très sympa et.

*FLE: il+y+a beaucoup aussi des [x 2] nouvelles ya'ni@s:ara toujours tu entends qu' il+y+a un projet ici il+y+a un nouvel restaurant ici il+y+a un nouvel hôtel.
*ENQ: et il+y+a une [x 3] amélioration économique en Syrie.
*FLE: oui très vite ya'ni@s:ara xxx.
*ENQ: la croissance économique.
*FLE: la croissance économique oui [x 2] il+y+a du travail il+y+a. *ENQ: et euh qu'est+ce+que tu pourrais dire <à un> [x 2] touriste français pour lui conseiller de venir en Syrie qu'est+ce+que il peut faire en Syrie ?
*FLE: euh en Syrie il+y+a tout il+y+en+a tout vraiment ya'ni@s:ara tu veux des [x 2] euh la [x 2] qu'est+ce+que on dit des euh des anciennes des anciens places comme Palmyra comme Bosra comme Ma'loula il+y+a.
*FLE: tu veux euh tu peux voir les [x 2] mélanges des [x 2] religions les mélanges des [x 3] groupes des il+y+a des groupes des.
*ENQ: des groupes ?
*FLE: oui des par+exemple il+y+a arman@s:ara il+y+en+a euh.
*ENQ: des Arméniens.
*FLE: des Arméniens il+y+a des xxx il+y+a des kurdi@s:ara il+y+a des arabes il+y+a seryan@s:ara il+y+en+a beaucoup ya'ni@s:ara tu peux voir tout.
*ENQ: les Syriaques les Kurdes les oui il+y+a beaucoup d' ethnies.
*FLE: beaucoup d' ethnies oui.
*FLE: euh il+y+a maintenant comme j' ai dit il+y+a la plage la [x 2] mer.
*FLE: euh il+y+a aussi comme il+y+a la désert il+y+a les forêts il+y+a les [x 2] montagnes.
*ENQ: et &com comment ça se fait que tu [x 2] nous as parlé de Palmyre de Bosra c'est quoi ce sont des [x 3] anciennes villes ? *FLE: des anciennes villes oui ce sont des très anciennes villes très importantes dans l' histoire de [x 2] l' être humain.
*FLE: tu sais que ici en Syrie c'est la première alphabet a commencé au Syrie la première euh qu'est+ce+que on dit culture ?
*ENQ: civilisation ?
*FLE: la [x 2] première civilisation elle a commencé la première musique la première note de musique.
*FLE: euh ils ont continué il+y+a plusieurs civilisations qui sont venues ici.
*FLE: tu peux voir beaucoup des [x 2] euh tu peux voir dans un même place des [x 2] choses pour plusieurs civilisations.
*FLE: comme Palmyra tu peux quand tu entends la l' histoire souvent de Palmyra comme était très important après ça elle a la les Romains ont détruire cette ville et.
*FLE: tu peux aussi les [x 2] premières langues <la langue> [x 2] à Ma'loula ils parlent la langue xxx.
*ENQ: araméenne?
*FLE: c'est la araméen c'est la [x 2] peut+être <la plus> [x 2] ancienne langue euh du monde.
*FLE: euh le [x 2] première évangile il est écrit par cette langue et il était parlé du de hind@s:ara jusqu'à la Syrie il est commencé ici c'est euh.
*ENQ: ça c'est pour euh la civilisation en Syrie et pour euh les autres tourismes comme euh le tourisme dans les est+ce+que il+y+a des infrastructures pour euh des hôtels qui [x 2] peuvent. *ENQ: les transports les tout ce+qui est nécessaire au touriste qu'est+ce+que tu en penses ?
*FLE: ya'ni@s:ara autre que les anciens civilisations il+y+a les hôtels sont très importants je peux nommer quelques hôtels qui sont par+exemple Four_Seasons il est ouvert depuis des &an deux années. *FLE: il+y+en+a un autre qui veut ouvrir il+y+a la plage euh qui il+y+a plusieurs places euh plusieurs hôtels très importants euh. *FLE: au mer Méditerranée euh pour les touristes ya'ni@s:ara tu sais euh les touristes voulons venir pas seulement voir les anciens civilisations ils vont amuser aussi.
*FLE: donc il+y+a euh <des très importants> [x 2] places à la mer des très importants places en Damas les hôtels les restaurants. *FLE: c'est très joli les nouvelles restaurants mais dans les anciens maisons c'est très joli les discothèques qui sont dans les anciens maisons aussi.
*FLE: les la nouvelle ville de Damas c'est très [x 2] joli aussi tu peux voir la différence que la Damas a euh réservé ? *ENQ: a conservé.
*FLE: a conservé les &anc les anciens maisons les anciens places sa [x 2] ancienne image.
*FLE: il+y+a des [x 2] plusieurs places que les très euh progressées ?
*FLE: il+y+a plusieurs euh.
*FLE: par+exemple tu peux euh cette année la Syrie était la le pays de la culture arabe.
*FLE: il+y+en+a beaucoup des [x 2] parties@s:eng je sais pas qu'est+ce+que on dit en français.
*ENQ: des fêtes ?
*FLE: des fêtes pour très importants chanteurs de du monde arabe et de [x 3] la monde aussi.
*ENQ: des concerts.

*FLE: <des concerts> [x 2] pas seulement des chanteuses aussi des pas seulement des chanteurs aussi des musiciens international de plusieurs pays.
*FLE: euh j' ai allé à plusieurs concerts c'était vraiment très formidable parce+que tu vois le [x 2] la culture de [x 2] toutes les pays.
*FLE: ya'ni@s:ara il+y+en+a du grec il+y+en+a de l' Italie il+y+en+a du Suède il+y+en+a de [x 2] la France il+y+en+a beaucoup. *FLE: et tu peux dire que aussi les centres culturels des [x 2] pays aussi ils travaillent beaucoup ya'ni@s:ara.
*FLE: tu vois que dans le centre culturel français par+exemple il+y+en+a beaucoup des [x 2] choses des euh beaucoup des je sais pas si c'est beaucoup des projets ou beaucoup des euh affaires. *FLE: beaucoup des activités xxx centre culturel de Russe le centre culturel euh d' Angleterre.
*ENQ: et d'après ce+que tu as dit tu peux donc voir la les anciennes civilisations même dans les lieux modernes on a conservé une image ancienne même dans les restaurants les lieux de tourisme moderne tu parles souvent des euh restaurants euh la cuisine syrienne est intéressante ?
*FLE: oui très intéressante la cuisine syrienne vraiment très intéressante et c'est euh je pense c'est la très bien cuisine au monde pour moi.
*FLE: xxx je vois pas que qu' il+y+a+pas seulement moi ya'ni@s:ara je
[x 2] euh.
*FLE: je connais beaucoup de personnes nous dans notre maison ici à Bab_Touma nous euh qu'est+ce+qu' on dit nous loue euh des chambres pour les étrangères qui viennent étudier la langue arabe aussi je vais parler de la langue arabe après.
*FLE: mais ils disent vraiment que la [x 3] cuisine syrienne est très bonne est très euh important euh.
*FLE: pour la langue arabe aussi la Syrie la première pays qui euh que les euh que les étrangères viennent pour apprendre la langue arabe.
*FLE: ils viennent dans euh.
*ENQ: c'est la première destination pour les apprenants de la langue arabe.
*FLE: oui [x 2] parce+que la notre langue euh qu'est+ce+que familiale qu'est+ce+qu' on dit notre langue qu' on parle euh en social. *ENQ: la langue de tous les jours tu veux dire la langue euh ah le dialecte syrien.
*FLE: le dialecte syrien. *ENQ: la langue parlée.
*FLE: la langue parlée est très proche de la langue écrit ou la langue euh pas comme les autres pays arabes.
*FLE: il+y+en+a très grandes différences par+exemple en Tunisie en Maroc.
*FLE: il+y+a une grande différence xxx euh entre la dialecte et la langue euh lire et la langue dans les livres.
*ENQ: alors qu' en Syrie c'est la langue le dialecte et la langue parlée la langue écrite pardon sont assez proches ?
*FLE: oui [x 2] c'est très proche.
*ENQ: d'accord et donc euh ce sont des appréciations d' étrangers ils ont apprécié la langue syrienne la pardon la cuisine syrienne.
*FLE: oui.
*ENQ: ils n' ont pas trouvé que c'était euh qu' elle était trop grasse trop euh tu sais en Syrie il+y+a de la viande crue il+y+a beaucoup de on mange beaucoup de choses un+peu.
*FLE: mais ce+qui est bien je pense dans la cuisine syrienne qu' il+y+a une grande variété.
*FLE: tu peux voir les la cuisine qu' il+y+a qui utilise la viande en grande quantité aussi.
*FLE: il+y+a un grande nombre de [x 4] cuisines.
*FLE: qui n' utilisent pas utilisent selon l' huile.
*FLE: tu sais en [x 2] avant ils ne [x 3] mangent pas ils mangent la viande mais pas comme maintenant peut+être ils mangent la viande pendant une seule fois par semaine ou une seule fois donc par mois.
*FLE: donc il+y+a une [x 2] très grande variété de la cuisine qui n' utilise pas la viande.
*FLE: il+y+a une grande mélange.
*FLE: xxx ils ont dit.
*ENQ: tu as parlé de la cuisine à l' huile on cuisine comment en Syrie à l' huile aussi ?
*ENQ: pardon ?
*ENQ: tu as parlé de la cuisine à l' huile de cuisiner à l' huile on cuisine comment en Syrie à l' huile ?
*FLE: la cuisine en Syrie ?
*ENQ: les matières grasses c'est quoi c'est l' huile le beurre qu'est+ce+qu' on utilise pour euh comme matières grasses ? *FLE: pour matières grasses ah moi j' ai compris que c'est très lourd par+exemple en l' utilisant la viande il+y+a l' huile qui qu' on utilise pas la viande on utilise l' huile.
*FLE: mais il+y+en+a aussi qu' on utilise pas euh qu'est+ce+que je veux dire il+y+a.
*ENQ: alsamne@s:ara ? *FLE: quoi ?
*ENQ: alsamne@s:ara ?
*FLE: alsamne@s:ara.

*ENQ: oui ça n' existe pas en français.
*ENQ: j' ai su que tu réfléchissais à ça mais il+n'y+a+pas de il+n'y+a+pas de mot en français pour dire samne@s:ara ça n' existe pas.
*ENQ: on ne peut ça n' existe pas.
*FLE: ça n' existe pas.
*ENQ: ils ne connaissent pas.
*FLE: ils doivent dire samne@s:ara donc.
*ENQ: ils ne connaissent pas.
*FLE: oui ils ne connaissent pas.
*ENQ: <on peut> [x 2] rien dire et euh quelles sont les ressources économiques de la Syrie ?
*FLE: les ressources hala@s:ara je veux dire aussi une autre chose [x 2] de la Syrie autre que la les.
*FLE: que la sécurité est très important ici en Syrie.
*FLE: il+y+a en deux mille cinq ou deux mille quatre c'était la troisième &pa pays du monde en &sécu &sé sécurité.
*FLE: par+exemple tu sors à quatre heures du matin chaque [x 2] jour tu vois les il+y+en+a beaucoup qui sont euh qui [x 2] sont dans la rute dans la route.
*FLE: et qui euh va et vient et euh surtout les étrangers peut+être dans leur pays je pense que ils ne peuvent pas sortir tout seuls. *FLE: il+y+en+a les des filles qui sortent et euh c'est très important je pense.
*ENQ: il+n'y+a+pas de risques.
*FLE: non il+n'y+a+pas de risques non [x 2].
*ENQ: et pour euh les étrangères il faut quand+même avoir prendre des précautions si elles veulent venir à Damas ou.
*FLE: oui je pense que.
*ENQ: pour la tenue vestimentaire par+exemple qu'est+ce+qu' elles doivent qu'est+ce+que les étrangers doivent porter ou ils peuvent s' habiller comme ils veulent ici en Syrie ?
*FLE: non en+général les étrangers respectent les [x 2] respectent notre euh culture notre euh.
*ENQ: mentalité ?
*FLE: mentalité oui.
*FLE: ya'ni@s:ara ils sait je xxx ils en+général tous qui sont venus chez
[x 2] nous ils connaît notre mentalité et euh comment ils doivent faire.
*ENQ: donc tu peux nous me parler maintenant de des ressources économiques syriennes ?
*ENQ: la Syrie vit de la du parce+que c'est un pays du Moyen_Orient il vit du pétrole du tourisme de l' agriculture <quelles sont ses > [x 2] richesses économiques ?
*FLE: je pense que euh le pétrole est le premier ressource.
*FLE: euh c'est très important pas comme la Golfe le Golfe mais il+y+en+a beaucoup des [x 2] compagnies qui des [x 2] entreprises qui viennent pour [x 2] euh rechercher le pétrole.
*FLE: et il+y+a le domaine touristique c'est [x 2] très important. *FLE: il+y+a aussi le domaine euh le les [x 2] transit qui passent de la Syrie c'est très important euh le la place de la Syrie.
*FLE: à la moitié des à la Méditerranée euh pas beaucoup.
*ENQ: sa position géographique ?
*FLE: sa position géographique aussi.
*ENQ: c'est un carrefour.
*FLE: oui.
*ENQ: et euh tu [x 2] me parlais du [x 2] pétrole et euh du [x 2] tourisme mais il+n'y+a pas d' agriculture ?
*FLE: il+y+a d' agriculture oui.
*ENQ: ce n' est pas un pays désertique où il fait trop chaud pour l' agriculture on peut.
*FLE: non oui c'est très important euh surtout en oliviers ?
*FLE: oui en l' huile peut+être.
*FLE: moi j' ai entendu quelque+chose je sais pas si c'est vrai. *FLE: que la Syrie euh elle [x 2] a le gouvernement syrienne elle a un projet de faire de la Syrie <la première> [x 2] pays de [x 2] faire l' huile de d' olivier.
*FLE: euh il+y+a aussi les citrons.
*ENQ: le premier producteur.
*FLE: le premier &produc producteur de l' huile.
*FLE: aussi il+y+a les <les citrons> [x 2] qu'est+ce+qu' on dit.
*ENQ: les citronniers.
*FLE: les citronniers euh à la &pla à la mer.
*FLE: aussi le il+y+a le raisin il+y+a ya'ni@s:ara tu sais c'est un pays de sur la Méditerranée donc il est très important en agriculture. *FLE: mais je ne pas dire je pense que la pétrole est euh donne plus de [x 2] l' argent et le premier ressource économique pour l' économie syrienne et le domaine touristique aussi.

*FLE: l' agriculture est très important.
*FLE: mais euh aussi ya'ni@s:ara si on veut comparer avec des autres pays par+exemple la France ou les autres je pense que la Syrie est euh on peut pas dire que la Syrie exporte <beaucoup des> [x 2] produits euh d' agriculture.
*ENQ: d'accord et le climat ?
*FLE: le climat c'est moyen non c'est [x 2] moyen.
*FLE: c'est moyen comme toutes les pays euh de la Méditerranée.
*FLE: ça dépend aussi des [x 2] places il+y+a la déserte.
*FLE: il+y+a aussi les places froids dans les montagnes il+y+en+a aussi les euh où il+y+a de l' humidité comme Homs comme euh la plage.
*ENQ: c'est varié.
*FLE: c'est varié mais c'est en+général c'est moyen.
*ENQ: tempéré.
*FLE: c'est tempéré oui.
*FLE: et on a notre et on a le saison de l' hiver le saison du printemps l' été.
*FLE: ya'ni@s:ara c'est comme toutes les tu sais les [x 2] pays de la Méditerranée sont comment est son climat.
*FLE: le meilleur <de tout le> [x 2] monde ya'ni@s:ara.
*FLE: tu sais que par+exemple pendant l' été il+n'y+a pas de des [x 2] pluies.
*FLE: dans euh des autres pays par+exemple peut+être à l' été il commence à pluie il+y+a des tempêtes.
*ENQ: en été il fait quand+même chaud ici.
*FLE: il [x 2] fait chaud oui.
*ENQ: les températures peuvent monter jusqu'à ?
*FLE: ici en Damas ?
*ENQ: oui.
*FLE: je sais pas quarante cinq cinquante.
*ENQ: il peut y avoir quarante cinq.
*FLE: oui.
*ENQ: bon merci.
*FLE: de rien.
@End

Annexe 6 : Liste des catégories morphosyntaxiques de la base de données CLAN

adj	adj	adjectif
adv	adv	adverbe
adv:int	adv:int	adverbe interrogatif
adv:neg	adv:neg	adverbe de négation
adv:place	adv:place	adverbe de lieu
adv:yn	adv:yn	adverbe oui/non
co	co	communicateur ou interjection
conj	conj	conjonction
det	det	determinant (articles définis et indéfinis)
det:dem	det:dem	déterminant démonstratif
det:gen	det:gen	déterminant général (autres déterminants que les
det:poss	det:poss	déterminant possessif
n	n	nom
n:let	n:let	lettre
n:prop	n:prop	nom propre
num	num	numéro
on	on	onomatopée
pct	pct	ponctuation
prep	prep	préposition
prep:art	prep:art	préposition-article
pro	pro	pronom (général)
pro:dat	pro:dat	pronom personnel datif
pro:dem	pro:dem	pronom démonstratif
pro:int	pro:int	pronom interrogatif
pro:obj	pro:obj	pronom personnel objet direct
pro:refl	pro:refl	pronom réfléchi
pro:rel	pro:rel	pronom relatif
pro:subj	pro:subj	pronom personnel sujet
pro:y	pro:y	pronoms y, en
unk	unk	catégorie indéfinie (xxx, yyy)
v	v	verbe (conjugué)
v&INF	v:aux	verbe auxiliaire
v&PP	v:exist	verbe d’existence (être et il+y+a)
v&PPRE	v:inf	verbe infinitif
v:aux	v:mdl	verbe modal (je fais cuire un gâteau)
v:aux&INF	v:mdllex	verbe modal lexical (je fais un gâteau)
v:aux&PP	v:poss	verbe d’appartenance (verbe avoir lexical)
v:exist	v:pp	verbe participe passé
v:exist&INF	v:ppre	verbe participe présent
v:exist&PP		
v:inf		
v:mdl		
v:mdl&INF		
v:mdl&PP		
v:mdl&PPRE		
v:mdllex v:mdllex&INF		
v:mdllex&PP		
v:mdllex&PPRE		
v:poss		
v:poss&INF		
v:poss&PP		

Annexe 7: Liste des ajouts au lexique de CLAN

La liste suivante répértorie l'ensemble des items que nous avons manuellement ajoutés à la base de données lexicale de CLAN. Certains n'y étaient pas présents car la base n'est pas exhaustive, d'autres concernent des locutions que nous voulons inclure.

aleppin {[scat adj]}
amharique {[scat adj][gn invar]}
amharique {[scat n] [gn masc]} "amharique&_MASC"
andin {[scat adj]}
anglophone {[scat adj]}
arabophone {[scat adj]}
araméen {[scat adj][gn invar]} "araméen&_MASC"
araméenne {[scat adj][gn invar]} "araméen&_FEM"
araméen {[scat n] [gn masc]} "araméen&_MASC"
apprenant {[scat n] [gn masc] [anim yes]} "apprenant"
apprécie {[scat v]}
apprécies {[scat v]}
arc+de+triomphe {[scat n] [gn masc]} "arc+de+triomphe&_MASC"
a+t+il {[scat v:poss]}
au+début {[scat adv]}
au+fur+et+à+mesure {[scat adv]}
au+milieu {[scat adv:place]}
au+milieu+d' {[scat prep:art]} "au+milieu+de"
au+milieu+de {[scat prep:art]} "au+milieu+de"
au+milieu+des {[scat prep:art]} "au+milieu+de"
au+milieu+du {[scat prep:art]} "au+milieu+de"
au+niveau {[scat adv]}
au+niveau+d' {[scat prep:art]} "au+niveau+de"
au+niveau+de {[scat prep:art]} "au+niveau+de"
au+niveau+des {[scat prep:art]} "au+niveau+de"
au+niveau+du {[scat prep:art]} "au+niveau+de"
au+pair {[scat adv]}
autour+de {[scat prep]}
autre+part {[scat adv]}
avant+qu' {[scat conj]} "avant+que"
avant+que {[scat conj]}
badminton {[scat n] [gn masc]} "badminton&_MASC"
blog {[scat n] [gn masc]} "blog&_MASC"
bolero {[scat n] [gn masc]} "bolero&_MASC"
brocoli {[scat n] [gn masc]} "brocoli&_MASC"
café+gourmand {[scat n] [gn masc] [tn singonly]} "café+gourmand&_MASC&_SING"
cafés+gourmands {[scat n] [gn masc] [tn plonly]} "cafés+gourmands&_MASC&_PL"
calin {[scat adj]}
calin {[scat n] [gn masc]} "calin&_MASC"
cantonais {[scat adj]}

catégoriser {[scat v]} "catégoriser"
centre+ville {[scat n] [gn masc]} "centre+ville&_MASC"
challenge {[scat n] [gn masc]} "challenge&_MASC"
chambre+de+commerce {[scat n] [gn fem] [tn singonly]} "chambre+de+commerce&_FEM&_SING"
chambres+de+commerce {[scat n] [gn fem] [tn plonly]} "chambres+de+commerce&_FEM&_PL"
charleston {[scat n] [gn masc]} "charleston&_MASC"
clope {[scat n] [gn masc]} "clope&_MASC"
coloc {[scat n]} "coloc"
colocation {[scat n] [gn fem]} "colocation&_FEM"
colombien {[scat adj]}
comme+ci+comme+ça {[scat adv]}
convoncatoire {[scat adj]}
coréen {[scat adj][gn invar]} "coréen&_MASC"
coréenne {[scat adj][gn invar]} "coréen&_FEM"
créativité {[scat n] [gn fem]} "créativité&_FEM"
cursus {[scat n] [gn masc] [tn singpl]} "cursus&_MASC&_SINGPL"
demi+heure {[scat n] [gn fem]} "demi+heure&_FEM"
design {[scat n] [gn masc]} "design&_MASC"
designer {[scat n] [gn masc]} "designer&_MASC"
de+temps+en+temps {[scat adv]}
de+toute+façon {[scat adv]}
disco {[scat n] [gn masc]} "disco&_MASC"
du+coup {[scat adv]}
dvd {[scat n] [gn masc] [tn singpl]} "dvd&_MASC&_SINGPL"
email {[scat n] [gn masc]} "email&_MASC"
en+ce+moment {[scat adv]}
en+face {[scat adv:place]}
en+face+d' {[scat prep:art]} "en+face+de"
en+face+de {[scat prep:art]} "en+face+de"
en+face+des {[scat prep:art]} "en+face+de"
en+face+du {[scat prep:art]} "en+face+de"
en+gros {[scat adv]}
en+tant+qu' {[scat pro:rel]}
exam {[scat n] [gn masc]} "exam&_MASC"
fac {[scat n] [gn fem]} "fac&_FEM"
finaliser {[scat v]} "finaliser"
foie+gras {[scat n] [gn masc]} "foie+gras&_MASC"
foot {[scat n] [gn masc]} "foot&_MASC"
frisbee {[scat n] [gn masc]} "frisbee&_MASC"
fruits+de+mer {[scat n] [gn masc]} "fruit+de+mer&_MASC"
fusionnel {[scat adj]}
galérer {[scat v]}
géochimie {[scat n] [gn fem]} "géochimie&_FEM"
géochimique {[scat adj][gn invar]}
gym {[scat n] [gn fem]} "gym&_FEM"
halte+garderie {[scat n] [gn fem]} "halte+garderie&_FEM"
hip+hop {[scat n] [gn masc]} "hip+hop&_MASC"
hispanophone {[scat adj][gn invar]}
hobbie {[scat n] [gn masc]} "hobbie&_MASC"
hongkongais {[scat adj]}

institutionnel {[scat adj]}
intergouvernemental {[scat adj]}
il+n'y+a {[scat v:exist]}
il+n'y+a+pas {[scat v:exist]}
il+n'y+a+plus {[scat v:exist]}
il+n'y+a+rien {[scat v:exist]}
il+n'y+avait+pas {[scat v:exist]}
il+n'y+en+a+pas {[scat v:exist]}
il+n'y+avait+rien {[scat v:exist]} "il+n'y+a+rien&IMPF"
il+y+aura {[scat v:exist]}
il+y+avait {[scat v:exist]} "il+y+a&IMPF"
jogging {[scat n] [gn masc]} "jogging&_MASC"
karaoké {[scat n] [gn masc]} "karaoké&_MASC"
labo {[scat n] [gn masc]} "labo&_MASC"
latino {[scat adj]}
latino+américain {[scat adj]}
libanais {[scat adj]}
libano+syrien {[scat adj]}
logiciel {[scat adj]}
logiciel {[scat n] [gn masc]} "logiciel&_MASC"
logistique {[scat adj]}
logistique {[scat n] [gn fem]} "logistique&_FEM"
management {[scat n] [gn masc]} "management&_MASC"
marketing {[scat n] [gn masc]} "marketing&_MASC"
master {[scat n] [gn masc]} "master&_MASC"
mayo {[scat n] [gn fem]} "mayo&_FEM"
meilleures {[scat adj]} "meilleur&_FEM&PL"
mise+en+scène {[scat n] [gn fem] [tn singonly]} "mise+en+scène&_FEM&_SING"
mises+en+scène {[scat n] [gn fem] [tn plonly]} "mises+en+scène&_FEM&_PL"
moitié+moitié {[scat adv]}
mon+Dieu {[scat co]}
mot+à+mot {[scat adv]}
multinational {[scat adj]}
multinationale {[scat n] [gn fem]} "multinationale&_FEM"
narcotrafic {[scat n] [gn masc]} "narcotrafic&_MASC"
nigérian {[scat adj][gn invar]} "nigérian&_MASC"
nigériane {[scat adj][gn invar]} "nigériane&_FEM"
n'est-ce+pas {[scat co]}
n'importe+quel {[scat adj]}
n'importe+quelle {[scat adj]}
n'importe+quelles {[scat adj]}
n'importe+quels {[scat adj]}
n'importe+où {[scat adv]}
ok {[scat co]}
palestinien {[scat adj]}
panda {[scat n] [gn masc]} "panda&_MASC"
par+ci+par+là {[scat adv]}
par+coeur {[scat adv]}
par+hasard {[scat adv]}
parapente {[scat n] [gn masc]} "parapente&_MASC"

pharmacologiste {[scat n] [gn masc]} "pharmacologiste&_MASC"
piercing {[scat n] [gn masc]} "piercing&_MASC"
plus+ou+moins {[scat adv]}
postdoc {[scat n] [gn masc]} "postdoc&_MASC"
privatisation {[scat n] [gn fem]} "privatisation&_FEM"
pro {[scat adj]}
pub {[scat n]}
promo {[scat n] [gn fem]} "promo&_FEM"
protège {[scat v]}
pénaliste {[scat n]}
préhispanique {[scat adj][gn invar]}
près+de+la {[scat prep:art]}
psycho {[scat n] [gn fem]} "psycho&_FEM"
psychopédagogie {[scat n] [gn fem]} "psychopédagogie&_FEM"
péruvien {[scat adj]}
radiologique {[scat adj][gn invar]}
rap {[scat n] [gn masc]} "rap&_MASC"
rebonjour {[scat co]}
rebonsoir {[scat co]}
reggae {[scat n] [gn masc]} "reggae&_MASC"
rennais {[scat adj]}
reprogresser {[scat v]} "reprogresser"
resto {[scat n] [gn masc]} "resto&_MASC"
resto_U {[scat n] [gn masc] [tn singonly]} "resto_U&_MASC&_SING"
restos_U {[scat n] [gn masc] [tn plonly]} "restos_U&_MASC&_PL"
retravailler {[scat v]}
revalidation {[scat n] [gn fem]} "revalidation&_FEM"
réceptionniste {[scat n]}
réorienter {[scat v]}
s'il+vous+plaît {[scat co]}
salle+de+bain {[scat n] [gn fem] [tn singonly]} "salle+de+bain&_FEM&_SING"
salles+de+bain {[scat n] [gn fem] [tn plonly]} "salles+de+bain&_FEM&_PL"
salsa {[scat n] [gn fem]} "salsa&_FEM"
sauf+qu' {[scat conj]} "sauf+que"
sauf+que {[scat conj]}
saxo {[scat n] [gn masc]} "saxo&_MASC"
scénographie {[scat n] [gn fem]} "scénographie&_FEM"
semi {[scat adj]}
socio {[scat adj]}
soi+même {[scat pro]} "soi+même"
squash {[scat n] [gn masc]} "squash&_MASC"
stressé {[scat adj]}
stéréotype {[scat n] [gn masc]} "stéréotype&_MASC"
sud+américaine {[scat adj]}
surfer {[scat v]}
suspicieux {[scat adj]}
syro+libanais {[scat adj]}
taoïsme {[scat n] [gn masc]} "taoïsme&_MASC"
tartiflette {[scat n] [gn fem]} "tartiflette&_FEM"
tofu {[scat n] [gn masc]} "tofu&_MASC"

transnational {[scat adj]}
trip+hop {[scat n] [gn masc] [tn singonly]} "trip+hop&_MASC&_SING"
trois_D {[scat adj][gn invar]}
trois_D {[scat n] [gn fem]} "trois_D&_FEM"
ultimate {[scat n] [gn masc]} "ultimate&_MASC"
viking {[scat adj]}
vénézuélien {[scat adj]}
waouh {[scat co]}
y+a+t+il {[scat v:exist]}
à+cause+de+la {[scat prep:art]} "à+cause+de"
à+cause+des {[scat prep:art]} "à+cause+de"
à+cause+du {[scat prep:art]} "à+cause+de"
à+ce+moment+là {[scat adv]}
à+l'aise {[scat adv]}
à+l'inverse {[scat adv]}
à+la+fin {[scat adv]}
à+part {[scat adv]}
étasunien {[scat adj]}
éthiopien {[scat adj]}
étrange {[scat adj][gn invar]}

RÉSUMÉ

Les méthodologies de constitution de corpus linguistiques ont été amplement étudiées, mais sont moins abondantes quand il s'agit de corpus oraux ; ces méthodologies sont encore plus rares en ce qui concerne l'interlangue orale. Le projet CIL (Corpus Inter Langue), en cours de finalisation à l'Université Rennes 2 et sous la supervision de l'équipe d'accueil LIDILE (EA 3874), vise à la constitution d'un corpus de productions écrites et orales d'apprenants en FLE et ALE. Cette thèse concerne le corpus oral de FLE du projet global (CIL-FLE). Partant du constat que l'intérêt des linguistes pour la langue orale a systématiquement été en retard par rapport à celui porté à la langue écrite, nous nous intéressons dans un premier temps à l'étude de l'oralité dans différents domaines de la linguistique d'un point de vue historique et épistémologique. Le second chapitre est consacré à la linguistique de corpus de manière générale et au corpus en tant qu'objet linguistique en particulier. En ce qui concerne la linguistique de corpus, nous tentons de présenter les différentes méthodologies auxquelles les linguistes ont recours lorsqu'il s'agit de consulter des données : introspection, élicitation ou consultation de données authentiques. Le concept de corpus est ensuite analysé selon un ensemble de critères définitoires que nous étudions en détail, afin de proposer une définition du corpus linguistique. Le troisième et dernier chapitre est la mise en application des constats théoriques dans la constitution du corpus CIL-FLE : nous détaillons les constituants du corpus, les protocoles de collecte et d'archivage. C'est au protocole de transcription que nous nous intéressons en particulier, en insistant sur les difficultés de la transcription de l'interlangue. Le corpus CIL-FLE, qui représente environ 105000 mots, représente le fruit de ce travail et sera ainsi détaillé.

Mots-clés : corpus, corpus FLE, corpus oral, corpus d'apprenants, linguistique de corpus, parole spontanée, transcription, transcription de l'interlangue.

Abstract

The need to design linguistic corpora to support research in linguistics has triggered the development of numerous studies exploring various approaches and methodologies regarding good practices for written corpus building. Fewer studies are available when it comes to spoken data and those that concern the interlanguage of learners are even rarer. The CIL project (Corpus Inter Langue), under completion at the University of Rennes2 and supervised by a research team specialising in the fields of linguistics and pedagogy (LIDILE), aims at building a large corpus of written and spoken productions in EFL and in FFL. This phd dissertation mainly focuses on the FFL (French as a Foreign Language) corpus (CIL-FLE).

The first chapter of the thesis is dedicated to the study of oral speech as a linguistic object from both a historical and an epistemological perspective. The second chapter tackles the question of corpus linguistics generally speaking as well as the concept/ notion of corpus as a linguistic object. Regarding corpus linguistics, we will review and explore the diverse approaches and methods that are used so as to carry out research enquiries: introspection, elicitation or consultation of authentic data. The concept of corpus is then analysed according to/following a series of criteria which we will closely examine in order to propose a definition of the linguistic corpus. The third and last chapter will implement the former theoretical findings through the description of the CIL corpus design. Thus, corpus constituents, transcription and archiving protocols will be described in detail. We are particularly interested in the transcription protocol and we will insist on the difficulties encountered when attempting to transcribe learners 'data. Finally, the CIL-FLE corpus, which contains approximately 105 000 words and was developed all along this phd, will be described.

Key words: corpora, FFL, oral speech, learner corpora, corpus linguistics, spontaneous speech, transcription, interlanguage transcription.