Najib ARBACH
Université Rennes 2, LIDILE EA 3874

Les fondements épistémologiques de la linguistique des corpus oraux

Article reçu le 15.01.2018 / Modifié le 26.03.2018 / Accepté le 05.05.2018

Résumé

Toute étude des phénomènes oraux nécessite, de fait, la consultation d'un ensemble d'exemples oraux (constitués ou non sous forme de corpus structuré), et cette linguistique du corpus oral (même si la pratique a précédé la terminologie), en tant que stricte méthodologie, s'est constituée dans différentes branches de la linguistique. Autrement dit, la langue orale s'est imposée en tant que sujet d'étude sérieux et légitime par étapes successives totalement hétérogènes géographiquement, chronologiquement et thématiquement. Ce sont ces étapes qui sont détaillées dans le présent article afin d'offrir une vision générale de plusieurs branches linguistiques qui possèdent entre elles un point commun essentiel : l'étude de l'oral. Ces branches qui ont permis à la linguistique des corpus oraux de se constituer sont l'acquisition du langage chez les enfants, la lexicographie, la sociolinguistique et enfin la phonétique et son articulation avec l'enseignement des langues.

Mots-clés : corpus oral, linguistique de corpus, langue parlée, corpus de référence, enseignement des langues.

Abstract

Any study of an oral phenomena necessarily requires the consultation of a set of oral examples that may or may not have the form of a structured corpus. The linguistics of an oral corpus, (even if the practice preceded the terminology,) as a strict methodology, was constituted in different branches of linguistics. In other words, oral language has emerged as a serious and legitimate subject of study in successive stages which are heterogeneous geographically, chronologically, and thematically. These steps are detailed in this paper to offer a general view of several linguistic disciplines that share an essential common point: the study of the spoken language. The branches that have enabled the linguistics of oral corpora to be constituted are the acquisition of language, lexicography, sociolinguistics and finally phonetics and its articulation with the language teaching science.

Key Words : Spoken corpora, oral corpus linguistics, spoken language, monitor corpus, teaching science.

Pour citer cet article :

ARBACH, N. (2018). Les fondements épistémologiques de la linguistique des corpus oraux. *Action Didactique*, 1, 31-54. http://univ-bejaia.dz/pdf/ad1/Arbach.pdf

Pour citer le numéro :

Amokrane, S. et Cortier, C. (dir.). (2018). Oral et oralité: perspectives didactiques, anthropologiques ou littéraires [numéro thématique]. *Action Didactique*, 1. http://univ-bejaia.dz/ad1

Introduction

Les grammaires classiques ont longtemps exclu la langue orale et considéré que seule la langue écrite était digne d'être étudiée, analysée ou enseignée. Nous envisageons donc, dans cet article, de comprendre quels ont été les événements ou les courants de pensée qui ont amené les linguistes à s'intéresser à la langue orale. En effet, jusqu'à la fin du XIX^{e} siècle, 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 XX^{e} siècle 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, soit une linguistique non normative, dont l'objectif est la description du langage dans toutes ses dimensions, dont la langue parlée.

Le processus n'est pas linéaire. Jusque dans les années 1980, soit soixante-dix ans après Saussure, la langue parlée était encore considérée comme une langue inférieure ou déclassée par rapport à la langue écrite, ou comme une « langue populaire », indigne d'être étudiée. Blanche-Benveniste et Jeanjean (1987, p. 12) illustrent ce 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 »).

Nous présenterons donc les circonstances dans lesquelles la langue orale s'est constituée en tant qu'objet d'étude dans plusieurs branches de la linguistique. Ensuite, nous analyserons la situation actuelle en ce qui concerne les corpus oraux français généralistes ou de référence, et nous proposerons une explication du retard qu'accuse la France en la matière. Les branches de la linguistique auxquelles nous nous intéresserons sont l'acquisition du langage chez les enfants, la lexicographie, la sociolinguistique et enfin la phonétique et son articulation avec l'enseignement des langues. Une présentation purement chronologique des corpus oraux n'aurait en effet pas été pertinente, compte tenu de l'éclatement (temporel et géographique) des pratiques qu'a connu chaque discipline en la matière. C'est pourquoi nous avons opté pour une présentation thématique. Nous précisons aussi, d'emblée, que nous n'envisageons pas d'effectuer un historique de l'oralité pour chacune de ces disciplines, mais uniquement de remonter à l'origine des *corpus oraux*. Cette

précision est nécessaire car nous verrons que l'étude de l'oral n'est pas systématiquement passée par la constitution d'un corpus oral.

En guise de préliminaire, nous soulignerons l'importance de la technique en ce qui concerne la linguistique des corpus oraux. En effet, les grammairiens et théoriciens, jusqu'à la fin du XIX[e] siècle, étaient fortement limités dans l'étude de la langue orale par l'impossibilité d'enregistrer le continuum sonore. Ce point sera le premier sur lequel nous allons nous pencher avant de proposer l'historique des corpus oraux qui constitue le cœur de cet article.

1. Langue orale et technologies

L'invention d'appareils d'enregistrement fut un moteur des études de la langue orale. En 1911, le grammairien et historien de la langue française Ferdinand Brunot inaugure « Les Archives de la Parole »[1]. En ce qui concerne le recueil et l'archivage des données sonores, c'est une première institutionnelle en France qui s'était inspirée du « Phonogrammarchiv » de Vienne[2]. Ferdinand Brunot s'impliqua lui-même dans la collecte des données, et enregistra patois, dialectes et accents dans différentes régions de France.

Mais les linguistes furent, dans un premier temps, majoritairement réfractaires au recours aux nouvelles technologies, tel que le rappelle Bonu : « L'oubli de la technologie est provoqué par la prépondérance accordée dans le traitement historique à la théorisation et à l'orientation textuelle et formalisante des courants dominants. » (Bonu, 2014, p. 8). Ces hésitations concernent les premiers phonographes, les enregistreurs portables, mais aussi l'informatique, qui permet la pérennisation des données et leur manipulation à grande échelle. En ce qui concerne le flux sonore en particulier, l'alignement du son avec sa transcription bouleverse l'approche même d'un corpus oral ; il est devenu possible d'interroger le corpus sur un fait linguistique particulier et d'en relever toutes les occurrences, sans que ceci ne signifiât pour autant la création immédiate d'un grand nombre de corpus, en raison de transcriptions alors très peu exploitables (Blanche-Benveniste et Jeanjean, 1987, p. 47).

Si la langue parlée a longtemps été négligée à cause d'une certaine forme de conservatisme des scientifiques d'avant le XX[e] siècle, ce conservatisme a également concerné l'adoption des nouvelles technologies. Celles-ci évoluant

[1] Le site est disponible à l'adresse suivante : http://gallicadossiers.bnf.fr/ArchivesParole/
[2] 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.

et se diffusant par à-coups, la constitution des corpus oraux a suivi le même chemin discontinu, plus ou moins irrégulier selon les différentes disciplines. Une fois que la technique a été en mesure de permettre la fixation des données sonores et de les analyser, il a fallu attendre que les linguistes ancrent son utilisation dans leurs recherches, dont les résultats « modifi[aient], ou tent[aient] de modifier, l'objet même de la discipline » (Bonu, 2014, p. 8). En effet, tel que nous allons le voir dans les branches de la linguistique que nous allons présenter, le rôle de la technologie est tout autant qualitatif que quantitatif.

2. Langue orale et acquisition du langage chez l'enfant

Bien avant toute technologie, l'acquisition du langage chez l'enfant est l'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 passe par l'analyse d'un corpus oral.

Au XIX^{e} siècle, les expériences du type « baby books » ou « diary's note », dans lesquelles 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 phonétiquement 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 et Stern, 1907), le premier journal entièrement consacré au langage de l'enfant.

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 et Wilson, 2001, p. 3). Les Américains 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^{e} 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, annonçant en cela les critères de constitution des corpus scientifiques modernes (Arbach, 2015). L'objectif formel des corpus transversaux est l'établissement de normes dans l'acquisition du langage grâce à de larges études quantitatives et comparatives.

Après 1957, les chercheurs privilégièrent les études longitudinales aux études transversales comparatives, pour deux raisons. La première fut la démocratisation du magnétophone[3] qui permit le suivi d'un enfant plus facilement qu'auparavant. La seconde fut la parution de *Syntactic structures* (Chomsky, 1957) qui amena les chercheurs à orienter leurs travaux sur la naissance de la syntaxe ; pour cela, ils eurent besoin de corpus longitudinaux qui leur permettent de suivre l'évolution de la syntaxe d'un seul et même enfant et donc, théoriquement, de comprendre le processus universel d'acquisition du langage. Moins d'enfants sont enregistrés, mais le suivi est plus long. Ce type de collecte s'étala jusqu'à la moitié des années 1970 environ ; depuis, la situation est une synthèse des deux approches, car les chercheurs désirent avoir à disposition des corpus à la fois longitudinaux (pour comprendre les processus d'acquisition) et transversaux (pour vérifier, comparer et compléter leurs résultats).

3. Langue orale et lexicographie

En lexicographie, 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 lexicographie est par définition un recensement qui nécessite inévitablement

[3] 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.

un corpus de travail, que celui-ci soit revendiqué en tant que tel ou pas (Geyken, 2008, p. 77).

La constitution de corpus dits représentatifs (Arbach et Ali, 2013) est donc une nécessité pour les lexicographes modernes, et la représentativité implique d'inclure des données orales dans le corpus du lexicographe. C'est dans les années 1980 que cela s'est pleinement concrétisé avec le projet « Collins Birmingham University International Language Database » (COBUILD). Initié à l'université de Birmingham, son objectif était la constitution d'un corpus de référence pour la langue anglaise et son exploitation pour la création d'un dictionnaire pour apprenants. La première matérialisation des objectifs du projet est la parution du dictionnaire *Collins COBUILD English Language Dictionary* en 1987, qui reflétait la réalité du langage anglais grâce à la prise en compte de données orales.

Le corpus COBUILD a en outre continué à évoluer jusqu'à nos jours, pour devenir l'actuel Bank of English (BoE). 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 la standardisation des données.

4. Langue orale et étude de la variation

L'étude de la variation est un autre champ d'étude où la constitution de corpus, ou du moins la consultation d'exemples authentiques s'est imposée 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. D'un point de vue théorique, ni le structuralisme, ni le générationalisme n'attribuèrent une place réelle à l'étude des variations individuelles et sociales dont se préoccupent la dialectologie et la sociolinguistique.

Jusqu'aux années 1960, l'étude des variations se faisait uniquement par le biais d'une sociolinguistique externe qui regroupait, selon Delais-Roussarie et Durand (2003, p. 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 ». Mais seul le facteur géographique est cité et étudié, et l'étude de la variation concernait donc les champs de la dialectologie. Avec les travaux de William Labov, dans les

années 1960 et 1970, naît une sociolinguistique interne qui place au centre de ses intérêts l'analyse des variations individuelles des locuteurs en se basant sur des facteurs sociaux et démographiques, et non plus exclusivement géographiques. Outre la dialectologie et la sociolinguistique, la prise en compte des données orales a également concerné l'analyse conversationnelle.

4.1. Langue orale et dialectologie

La dialectologie est une science qui nécessite la collecte de corpus conséquents mais qui, d'après Francis (1991, p. 23), ne se développa qu'au début du XIX^{e} siècle, vers 1820, soit en retard par rapport à la lexicographie : ceci est 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 à l'intérêt accordé à la linguistique historique et à la linguistique comparative d'une part, mais aussi au romantisme dominant de l'époque : l'attention se porte sur le « langage même des hommes », et les linguistes commencent à s'intéresser aux termes omis par les lexicographes.

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. 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.

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 pour recueillir des informations sur la prononciation dans les régions (Delais-Roussarie et Durand, 2003, p. 18). Les résultats obtenus ont permis l'élaboration de *l'Atlas linguistique de la France* en 1902, « un travail de pionniers qui reste un grand ouvrage de référence » (Delais-Roussarie et Durand, 2003, p. 18), mais la mort de Gilliéron en 1926 laisse la place vide et la dialectologie en France s'en trouve retardée (Chevalier et Encrevé, 1984, p. 66).

En Angleterre et aux États-Unis, de très nombreuses études similaires ont été menées dès la fin du XIX^{e} siècle. Mais après la Seconde Guerre mondiale, la discipline n'a pas périclité comme en France, mais mué. De Fornel et Léon (2000, p. 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 ».

Cette 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 et Jeanjean, qui indiquent qu'à l'époque, les auteurs intéressés par le français parlé urbain ne trouvaient aucun cercle pour les accueillir (Blanche-Benveniste et Jeanjean, 1987, p. 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 CFPP2000 (Corpus de Français Parlé Parisien des années 2000), qui est un projet de l'équipe SYLED à l'Université Sorbonne nouvelle, Paris 3[4]. Sur le site du CFPP2000, le corpus est présenté comme la première étape d'un projet qui concerne les formes du français parlé dans l'agglomération parisienne.

4.2. Langue orale et sociolinguistique

Les liens entre corpus oraux et sociolinguistique sont exclusivement illustrés par les travaux de William Labov, dont l'importance 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.

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 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, mais ouvre la voie aux corpus oraux.

[4] Il faudrait aussi citer le Multicultural Paris French (MPF), qui est projet sociolinguistique qui vise à comparer entre le français contemporain de Paris et l'anglais contemporain de Londres. Toutefois, ce corpus est un projet créé dans les universités de Londres, à l'image d'autres corpus francophones mais non français, tel que nous allons le voir dans la section suivante.

Tout au long de sa carrière, Labov a initié et perfectionné les méthodologies de recueil des données orales en vue de leur analyse, en insistant sur l'importance des données réelles, enregistrées en situation et loin de tout contexte artificiel de laboratoire. Outre l'apport de Labov à la sociolinguistique en général[5], les approches méthodologiques de la sociolinguistique labovienne ont été d'un apport précieux pour les protocoles de constitution des corpus oraux modernes.

Mais en ce qui concerne la France[6], Blanche-Benveniste et Jeanjean (1987, p. 85) parlent de « la curieuse aliénation du milieu linguistique français » et Cappeau et 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 et Gadet, 2007, p. 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. Ainsi, les corpus sociolinguistiques en France sont-ils relativement rares ; 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 :

> 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 (Bergounioux, Baraduc et Dumont, 1992, p. 74)

Le projet est donc lancé et il est pionnier dans sa volonté de constituer un corpus cohérent et documenté d'enregistrements oraux. Les données furent

[5] À ce propos, cf. le numéro thématique de *Langage* consacré à l'œuvre de William Labov, « Hétérogénéité et variation : Labov, un bilan », numéro coordonné par Françoise Gadet, *Langage*, n° 108, 1993.
[6] Nous parlons bien des corpus constitués en France, et non des corpus francophones par des équipes étrangères, tel le MPF ou le corpus Sankoff-Cedergren, constitué au Canada.

collectées en cinq semaines en 1969. 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. Bien que les motivations de la constitution d'ESLO1 fussent 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.

4.3. Corpus et analyse conversationnelle

Le domaine de l'analyse conversationnelle nous intéresse pour les méthodologies qu'il a employées 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. Champ d'analyse ayant émergé aux États-Unis avec les travaux de Harvey Sacks, qui est assistant au centre d'études scientifiques du suicide à Los Angeles en 1963. 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 et 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 et Léon, 2000, p. 133)

En ce qui concerne les filiations de l'analyse conversationnelle, De Fornel et Léon (2000, p. 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, p. 142), cette époque était aussi celle des premières grammaires de l'oral et des premiers grands corpus de données orales authentiques. L'apport de la linguistique conversationnelle se situe ainsi dans ses enquêtes de terrain, en raison de l'exigence de travail sur des données orales collectées dans leur contexte social d'énonciation.

En ce qui concerne la France, Balthasar et Bert (2005, p. 2) affirment que ce sont les structures dont est issu l'actuel laboratoire ICAR[7] 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[8]. Les corpus ont été numérisés afin de garantir leur pérennité et leur diffusion.

5. Corpus et enseignement des langues

L'enseignement des langues est l'une des pierres angulaires de la linguistique de corpus. Autrement dit, la constitution de manuels de langues, de dictionnaires pour apprenants, de grammaires exhaustives ou simplifiées (en ALE plus particulièrement et en didactique, sont autant d'objectifs qui ont été à l'origine de la constitution de corpus parfois extrêmement élaborés. Dans la majorité des cas, et en raison de la nature même des objectifs primaires de ces corpus (l'enseignement des langues), les données orales ont été incluses dans les bases de données, analysées, étudiées. L'enseignement des langues est donc un facteur essentiel dans la valorisation de la langue parlée dans différentes branches de la linguistique. Les liens entre l'enseignement des langues et l'oral dans les corpus sont toutefois nés avec la phonétique en Europe, puis aux États-Unis avec la volonté d'implémenter des méthodes d'enseignement rapides et efficaces.

5.1. Langue orale et phonétique dans le cadre de l'enseignement des langues

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'oral pour des raisons évidentes. En second lieu, la phonétique ne saurait reposer sur l'introspection ou les intuitions du chercheur ; l'empirisme n'est pas, pour un phonéticien, un choix, mais une obligation car il s'intéresse à l'étude des variations individuelles, et son

[7] ICAR : Interactions, Corpus, Apprentissages, Représentations. UMR 5191 CNRS, Université de Lyon.
[8] CLAPI : Corpus de LAngue Parlée en Interaction. L'adresse de la base est la suivante : http://clapi.univ-lyon2.fr/

matériau de travail est nécessairement un corpus de données authentiques. 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. Ces liens sont autrement plus complexes que ne le laisseront entendre les lignes qui suivent, mais les limites de cet article nous imposent de ne pas les évoquer.

Sur le plan théorique, la phonétique naît dans les milieux pédagogiques, du fait des travaux d'enseignants phonéticiens préoccupés par l'enseignement d'une langue orale, en réaction aux méthodologies classiques : ce sont en effet les nouveaux besoins sociaux[9] de la fin du XIX^{e} siècle qui expliquent le progrès de la méthode orale dans l'enseignement des langues vivantes, méthode qui porta les phonéticiens à la pointe du combat contre les pratiques traditionnelles. L'apport de la phonétique à l'enseignement des langues est « inestimable » (Galazzi, 1995, p. 95).

Les figures les plus marquantes sont celles du phonéticien Wilhelm Viëtor (1850-1918) 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 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 grammaire-traduction de l'époque.

En 1886, Paul Passy fonde 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

[9] Concernant ces besoins, Puren (1988, p. 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 ».

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 »[10]. Passy avouait toutefois 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 l'une des premières études de la langue parlée faite à l'aide de vérifications instrumentales. Mais 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, p. 111-112) rapporte l'essoufflement de la phonétique après la mort des deux enseignants : 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 ayant amené à la constitution d'un corpus oral selon des critères scientifiques. Nous précisons que ce ne sont pas les enquêtes en dialectologie ou en phonologie qui ont fait défaut, mais la constitution de corpus oraux qui aurait dû en découler. Nous avons par exemple vu que Martinet (1945), pour son enquête sur la prononciation dans un camp d'officiers prisonniers, s'était reposé sur un questionnaire écrit. De même, lorsque Martinet et Walter publient le *Dictionnaire de la prononciation française dans son usage réel* (Martinet et Walter, 1973), il s'agit encore une fois d'une enquête phonologique qui ne repose pas sur un corpus oral, tel que l'explique Henriette Walter :

> Le tout premier problème à résoudre pour réaliser ce projet était sans aucun doute celui de la constitution du corpus à soumettre à l'enquête. La solution finalement adoptée a reposé sur la confrontation des prononciations figurant dans les dictionnaires et traités d'orthoépie alors en usage. (Walter, 2009, p. 147).

Nous avons ainsi un autre exemple de l'analyse de l'oral par le truchement de l'écrit[11], et ces biais n'ont été évités qu'avec les projets *Phonologie du Français*

[10] L'expression est rapportée par Enrica Galazzi (1995, p. 98).

[11] Nous n'avons cité les travaux de Martinet qu'à titre d'exemple, puisque nous nous intéressons au corpus oral en tant qu'objet scientifique. Mais il n'est pas utile, ici, de recenser toutes les enquêtes de dialectologie ne s'étant pas reposées sur un corpus oral. Cf. à ce sujet Pop (1950) ou Auroux (1979).

Contemporain (PFC) et *Interphonologie du français contemporain* (IPFC) dans les années 2000. Tout au long du XX^e^ siècle, l'élan de la phonétique appliquée à la didactique des langues n'a permis la constitution d'aucun corpus oral en France.

5.2. Langue orale et listes simplifiées pour l'enseignement des langues

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, p. 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. En ce qui concerne l'Allemagne vaincue, son état est catastrophique en 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 durant les années 1920[12]. 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

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

l'anglais, nous comprenons que la réflexion sur la didactique des langues fut principalement anglo-saxonne durant trois décennies.

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 de « manufactured, doctored, lop-sided, unnatural, peculiar, and even bizarre examples »[13], 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 :

> 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, p. 335)

La démarche est donc d'établir, pour une langue donnée, une liste de vocabulaire simplifiée, qui concentrerait en un nombre restreint le lexique le plus usité 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, p. 335), qui affirme que « la linguistique de corpus

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

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 BoE, où il est écrit que « à peu près 90% de l'anglais oral et écrit est constitué d'environ 3 500 mots ».

Le premier travail sur corpus en vue d'élaborer une liste de vocabulaire simplifié fut celui d'Edward Thorndike[14] aux États-Unis, qui publia *The teacher's word book* (1921). Ces travaux bénéficièrent d'une politique volontaire et d'investissements institutionnels. Sous la houlette de la Carnegie Corporation of New York[15], se 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, ou Harold Palmer, Edward Thorndike et Lawrence Faucett en Chine, qui publieront *Interim report on vocabulary selection* en 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. Ce sont ces travaux, ou encore le Basic English, qui inspirèrent *Le Français fondamental*, qui nous intéresse particulièrement pour les raisons que nous allons voir.

L'élaboration du français fondamental est la première étude de statistique lexicale menée sur un corpus de conversations spontanées enregistrées en français[16]. Quand *L'élaboration du français fondamental* paraît en 1956 (Gougenheim, Rivenc et Sauvageot, 1956), la langue française avait perdu de son rayonnement passé et n'est plus la langue universelle du XVIII^{e} siècle[17] ; 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 répondait à un élargissement de la nécessité de

[14] Thorndike (1874-1949) était un psychologue américain, qui travailla essentiellement sur l'intelligence animale et la pédagogie de l'enseignement et de l'éducation.
[15] 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.
[16] Une polémique naquit du projet, accusé par certains, à droite, de détériorer le niveau de la langue et de la culture françaises 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, p. 208).
[17] Gougenheim, Rivenc et Sauvageot (1956) nuancent 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.

diffusion de la langue française[18] et les auteurs du *Français fondamental* s'étaient inspiré, dans leurs motivations et dans leur démarche méthodologique, du *Basic English*. Leur travail « a constitué, avec de nombreux tâtonnements, une doctrine, nous dirons presque une science, des langues de base » (Gougenheim, Rivenc et Sauvageot, 1956, p. 11). De nombreuses études et manifestations scientifiques se sont penchées sur le *Français fondamental*[19], il nous intéresse ici car il a permis la constitution du premier corpus oral en France.

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, Rivenc et Sauvageot, 1956, p. 61)

Jusqu'au *Français fondamental*, l'étude de l'oral se faisait en France au moyen d'artifices : les exemples oraux parfois forgés de Damourette et Pichon (voir *infra*), l'étude du français populaire par le biais de lettres par Henri Frei (1929)[20] ou l'étude de la prononciation du français par Martinet par le biais de l'écrit. 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 et Jeanjean au milieu des années 1980 attestent que, 30

[18] Un grand nombre des anciennes colonies et protectorats français avait adopté la langue française comme langue nationale, les facultés et écoles françaises voyaient un grand nombre d'étudiants étrangers venir y faire leurs études et un grand nombre de techniciens étrangers venait accomplir des stages professionnels dans les entreprises françaises en raison de l'essor économique et industriel d'après-guerre,

[19] Voir les actes du Colloque « Français Fondamental, corpus oraux, contenus d'enseignement. 50 ans de travaux et d'enjeux », École Normale Supérieure - Lettres et Sciences Humaines (Lyon), décembre 2005 et notamment le volume 36 de la revue *Documents* SIHFLES « De quelques enjeux et usages historiques du Français fondamental », https://journals.openedition.org/dhfles/1178 ; ainsi que Galazzi (2008).

[20] 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.

ans après, l'oral n'était pas encore linguistiquement légitime, ce qui amplifie le caractère pionnier et audacieux des auteurs du *Français fondamental*.

6. Les corpus oraux de référence, le retard de la France

Historiquement, les prémisses de l'étude de l'oral pour l'élaboration d'une grammaire générale sont à situer en France avec les travaux de Damourette et Pichon, qui ont rédigé l'*Essai de grammaire de la langue française* en 1930. Les auteurs 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, ou en notant des remarques entendues dans la rue ou durant les consultations médicales de Pichon, qui était médecin. 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 ».

Après Damourette et Pichon, nous avons vu qu'il y eut peu de corpus oraux en France, au moins jusqu'aux années 1980. En revanche, aux États-Unis, les travaux fondateurs de Charles C. Fries (1887-1967) auront un impact plus important. Il s'était principalement intéressé à l'enseignement de la langue anglaise en tant que langue première et seconde, en publiant notamment *Teaching and learning English as a Foreign Language* (1945), et *The structure of English : An introduction to the construction of English sentences* (1952). 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[21], 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 ». Fries voyait au contraire dans la langue parlée « le langage réel » qu'il lui fallait étudier et analyser, contrairement aux 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.

C'est donc, encore une fois, l'enseignement des langues qui a constitué un prétexte pour la valorisation des données orales. 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 et Pichon. Longtemps

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

déconsidéré en France, l'oral est aujourd'hui reconnu en tant qu'objet d'étude légitime et sérieux, mais la France n'a toujours pas résolu le problème de la disponibilité des données orales ; 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, 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 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. Pour illustrer notre propos sur la disponibilité des données, nous citerons les trois plus importants corpus 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[22] et le Corpus de Référence du Français Parlé (CRFP). Comme le montre l'inventaire de Cappeau et 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 remonte à 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. Dès les années 1980, Blanche-Benveniste et Jeanjean (1987, p. 4) écrivaient 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 resta inchangée 20 ans plus tard, comme le notent Cappeau et 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 et Gadet, 2007, p. 130)

[22] 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.

De nombreuses initiatives institutionnelles ou d'équipes que présentent Cappeau et Gadet (2007, p. 130-131) tentent de combler ce retard, mais considérons ses raisons qui pourraient 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.

Dans d'autres cas, le corpus est constitué de façon optimale pour une recherche précise ; dans ce cas de figure, les transcriptions et les annotations trop spécifiques le rendent inexploitables pour 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 (2006, p. 3), soit des corpus inexploitables pour des raisons juridiques, personnelles ou scientifiques. À un autre niveau, un certain conservatisme français qui subsiste encore vis-à-vis de la langue parlée est à souligner. 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 16e au 20e 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, p. 37)

Debaisieux (2009, p. 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'auteure, « 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. Si des avancées institutionnelles sont donc tangibles, nous ne saurions cependant 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.

Conclusion

Cet article nous a permis de vérifier que 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 parfois communs sont toutefois à souligner. Nous avons aussi pu constater que dans la plupart des domaines, l'intérêt pour la langue orale est né en Europe, mais de nombreux facteurs historiques ont fait en sorte que les travaux des pionniers n'ancrèrent pas l'étude de la langue parlée dans la tradition européenne : ce furent souvent les Anglo-saxons qui 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.

Pour citer à nouveau Blanche-Benveniste et Jeanjean (1987, p. 85), « la curieuse aliénation du milieu linguistique français » ne peut plus être excusée par l'obstacle technologique. De nos jours où la France est à l'avant-garde des moyens technologiques et humains, comment justifier alors 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 se sont dotés d'un ou de plusieurs corpus oraux de référence.

Références bibliographiques

Arbach, N. (2015). *Constitution d'un corpus oral de FLE : enjeux théoriques et méthodologiques* (Thèse de doctorat). Université Rennes 2, Rennes, France.

Arbach, N., et Ali, S. (2013). Aspects théoriques et méthodologiques de la représentativité des corpus, *Corela. Cognition, représentation, langage*, [en ligne], HS-13, 1-16. Récupéré de : https://journals.openedition.org/corela/3029.

Auroux, S. (1979). La catégorie du parler et la linguistique. *Romantisme*, 9(25), 157-178.

Balthasar, L., et Bert, M. (2005). La plateforme "Corpus de langues parlées en interaction" (CLAPI). Historique, état des lieux, perspectives. *Lidil,*

Revue de linguistique et de didactique des langues, 31, 13–33.

Baude, O. (ed.), (2006). *Corpus oraux, guide des bonnes pratiques*. Paris : CNRS.

Bergounioux, G., Baraduc, J., et Dumont, C. (1992). L'étude sociolinguistique sur Orléans (1966-1991) : 25 ans d'histoire d'un corpus. *Langue française*, 93 (1), 74–93.

Blanche-Benveniste, C., et Jeanjean, C. (1987). *Le français parlé. Transcription et édition*. Paris : Didier Érudition, Institut national de la Langue française.

Bonu, B. (2014). "L'autre" révolution technologique en sciences du langage : les cas du phonographe et du magnétophone à cassette. *Dossiers d'HEL, SHESL, Linguistiques d'intervention. Des usages socio-politiques des savoirs sur le langage et les langues*. Récupéré de : https://halshs.archives-ouvertes.fr/halshs-01115046/document

Boulton, A. (2007). Esprit de corpus : Promouvoir l'exploitation de corpus en apprentissage des langues. *Texte et Corpus* (3), 37–46.

Cappeau, P., et Gadet, F. (2007). Où en sont les corpus sur les français parlés ? *Revue française de linguistique appliquée*, 12 (1), 129–133.

Cappeau, P., et Seijido, M. (2005). *Les corpus oraux en français*, Récupéré de : http://www.culturecommunication.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France

Chomsky, N. (1957). *Syntactic structures*. Berlin : Walter de Gruyter.

Chevalier, J.C., et Encrevé, P. (1984). La création de revues dans les années 60 : matériaux pour l'histoire récente de la linguistique en France. *Langue française,* 63 (1), 57–102.

Darwin, C. (1877). A biographical sketch of an infant. *Mind*, 2 (7), 285–294.

Debaisieux, J.-M. (2009). Des documents authentiques oraux aux corpus : un défi pour la didactique du FLE. *Mélanges CRAPEL*, 31, 36–56.

De Beaugrande, R. (2000). *Large corpora and applied linguistics : HG Widdowson versus J.McH. Sinclair*, Récupéré du site de l'auteur : http://www.beaugrande.com/WiddowSincS.htm

De Fornel, M., et Léon, J. (2000). L'analyse de conversation, de l'ethnomethodologie à la linguistique interactionnelle. *Histoire épistémologie langage*, 22 (1), 131–155.

Delais-Roussarie, E., et Durand, J. (2003). *Corpus et variation en phonologie du français : méthodes et analyses*. Toulouse : Presses Univ. du Mirail.

Francis, W. N., (1991). Language corpora BC. Dans J. Svartvik (dir.), *Directions*

Corpus Linguistics: Proceedings of Nobel Symposium (p. 17-31), Berlin : Walter de Gruyter.

Frei, H. (1929). *La grammaire des fautes*. Genève : Slatkine.

Fries, C.C. (1945). *Teaching and learning English as a Foreign Language*. New York : Harcourt Brace.

Fries, C.C. (1952). *The structure of English : An introduction to the construction of English sentences*. New York : Harcourt Brace.

Fries, P. H. (2008). Charles C. Fries, linguistics and corpus linguistics. *ICAME Journal*, (34), 89‑119.

Gougenheim, G., Rivenc, P., et Sauvageot, A, (1956). *L'élaboration du français élémentaire*. Paris : Didier.

Kennedy, G. (1991). Preferred ways of putting things with implications for language teaching. Dans J. Svartvik (dir.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium* (p. 335-378), Berlin : Walter de Gruyter.

Galazzi, E. (1955). Phonétique/Université/Enseignement à la fin du XIX[e] siècle. *Histoire épistémologie langage*, 17 (1), 95–114.

Galazzi, E. (2008). 1950 : où est passée l'oralité ? Français Fondamental, corpus oraux, contenus d'enseignement. 50 ans de travaux et d'enjeux. *Le français dans le monde, recherches et applications*, 43, 12-26.

Geyken, A. (2008). Quelques problèmes observés dans l'élaboration de dictionnaires à partir de corpus. *Langages*, 3, 77–94.

Martinet, A. (1945). *La prononciation du français contemporain*. Paris : Librairie Droz.

Martinet, A. et Walter H. (1973). *Dictionnaire de la prononciation française dans son usage réel.* Paris : France-Expansion.

McEnery, A. M., et Wilson, A. (2001). *Corpus linguistics : an introduction.* Édimbourg : Edinburgh University Press.

Mondada, L. (2000). « Les effets théoriques des pratiques de transcription », *Linx. Revue des linguistes de l'université Paris X Nanterre*, 42, 131–146.

Pop, S. (1950). *La dialectologie : aperçu historique et méthodes d'enquêtes linguistiques*. Louvain-la-Neuve : Presses Universitaires de Louvain.

Preyer, W. (1884). *Die Seele des Kindes*. Berlin : Grieben's Verlag.

Puren, C. (1988). *Histoire des méthodologies de l'enseignement des langues.* Paris : Nathan-CLE international. Récupéré de : http://www.aplv-languesmodernes.org/IMG/pdf/puren_histoire_methodologies.pdf.

Stern, C., et Stern, W. (1907). *Die kindersprache.* Oxford : Barth.

Taine, H. (1876). Note sur l'acquisition du langage chez les enfants et dans l'espèce humaine. *Revue Philosophique de la France et de l'Étranger*, 1, 5–23.

Walter H. (2009). André Martinet et la linguistique appliquée. *La linguistique*, 2, 145-152.

AUTEUR

Najib ARBACH est docteur en sciences du langage de l'Université Rennes 2 (2005) et ATER à l'Université Sophia-Antipolis de Nice. Il est rattaché à l'unité de recherche *Linguistique Ingénierie et Didactique des Langues* (LIDILE), EA 3874, Université Rennes 2. Sa thèse de doctorat est intitulée *Constitution d'un corpus oral de FLE : enjeux théoriques et méthodologiques*. Ses principales publications sont :
- Arbach, N., et Ali, S. (2013). « Aspects théoriques et méthodologiques de la représentativité des corpus », *Corela. Cognition, représentation, langage*, (HS-13).
- Arbach, N. (2015). *Constitution d'un corpus oral de FLE : enjeux théoriques et méthodologiques* (Thèse de doctorat). Université Rennes 2, Rennes.