Corpus et appropriation de L1 et L2

Alex Boulton, Emmanuelle Canut, Emmanuelle Guerin, Christophe Parisse, Henry Tyne

Linx
Revue des linguistes de l'université Paris X Nanterre

68-69 | 2013
Corpus et apprentissage du français

Corpus et appropriation de L1 et L2

Alex Boulton, Emmanuelle Canut, Emmanuelle Guerin, Christophe Parisse et Henry Tyne

revues.org

Édition électronique
URL : http://linx.revues.org/1475
DOI : 10.4000/linx.1475
ISSN : 2118-9692

Éditeur
Université Paris Ouest – département Sciences du langage

Édition imprimée
Date de publication : 19 novembre 2013
Pagination : 9-32
ISSN : 0246-8743

Référence électronique
Alex Boulton, Emmanuelle Canut, Emmanuelle Guerin, Christophe Parisse et Henry Tyne, « Corpus et appropriation de L1 et L2 », *Linx* [En ligne], 68-69 | 2013, mis en ligne le 29 novembre 2015, consulté le 02 octobre 2016. URL : http://linx.revues.org/1475 ; DOI : 10.4000/linx.1475

Corpus et appropriation de L1 et L2

Alex Boulton, *Université de Lorraine, ATILF*
Emmanuelle Canut, *Université de Lille 3, Savoirs Textes Langages*
Emmanuelle Guerin, *Université d'Orléans, Laboratoire Ligérien de Linguistique*
Christophe Parisse, *Université Paris Ouest Nanterre La Défense, MoDyCo*
Henry Tyne, *Université de Perpignan Via Domitia, CRESEM*

1. Introduction

Les corpus ont pris une place importance dans bien des secteurs de la linguistique et de la linguistique appliquée depuis un certain nombre d'années. Cet article se donne pour objectif de présenter une vue d'ensemble de la place des corpus et de la linguistique de corpus dans le domaine de l'appropriation (acquisition, apprentissage et enseignement) de la ou des langue(s). Dans un premier temps il sera question d'aborder les enjeux pour la description linguistique avant d'aborder la question de l'utilisation ou l'application des corpus. Nous consacrerons du temps à l'acquisition et à l'enseignement de la langue première (L1) comme seconde (L2), tout en faisant émerger les aspects singuliers de chaque approche comme les aspects partagés. Il sera question dans cet article non seulement de mettre en avant des comparaisons entre domaines d'utilisation et d'exploitation des corpus autour de l'appropriation du langage et des langues, mais il sera également question de pointer les problèmes ou les difficultés liées à la méthodologie sur corpus pour le traitement de certains domaines. Les différentes définitions et utilisations du terme « corpus » qui sont évoquées ici seront illustrées dans les articles qui forment ce numéro.

2. Qu'est-ce qu'un corpus ?

Pour le commun des mortels, un corpus est une collection d'objets, souvent de documents, de textes, d'œuvres, etc. En linguistique de corpus[1], un corpus est généralement considéré comme une collection électronique de textes pouvant être interrogés à l'aide d'un logiciel de recherche (Gilquin & Gries, 2009). Si cette définition ne vaut ni pour tous les linguistes, ni pour tous les corpus, elle indique néanmoins ce qu'est un corpus pour bon nombre de linguistes aujourd'hui : une collection de données langagières que l'on peut interroger à l'aide de techniques et technologies dédiées. Mais au-delà de l'importance du corpus comme collection en vue de l'observation de données, il importe de savoir ce qui constitue cette collection : que contient un corpus ? Et, plus précisément, que contient un corpus de données en langue ? Le *Trésor de la langue française informatisé* (TLFi) se réfère tout d'abord au *Larousse encyclopédique* de 1961 : le corpus est un « ensemble d'énoncés servant de base à l'analyse linguistique ». Dans sa définition actuelle, le TLFi parle d'un « ensemble de textes établi selon un principe de documentation exhaustive, un critère thématique ou exemplaire en vue de leur étude linguistique », avec comme exemple « le corpus des textes parus d'un journal, d'une revue ; un corpus littéraire ; le corpus du vocabulaire français ». Il est donc intéressant de noter que le contenu d'un corpus n'est pas arrêté puisqu'il va de sa taille la plus réduite (ensemble d'énoncés) à une taille conséquente (ensemble de textes).

La linguistique de corpus s'est développée en partie grâce aux avancées technologiques permettant de travailler sur des masses de données de plus en plus importantes mais aussi avec des outils de recherche appropriés (Tognini-Bonelli, 2010), ce qui explique sa définition actuelle de collection électronique de textes interrogeables à l'aide d'un logiciel. On parle aussi de « bases de données » ou encore de « bases textuelles » ou même d'« archives » pour certains recueils électroniques contenant des éléments qui ne formeraient pas un tout cohérent au même titre qu'un corpus. C'est néanmoins la méthodologie de travail sur corpus qui peut être privilégiée dans ces cas : autrement dit, un certain type de démarche, avec des outils spécifiques, peut valoir à une étude de revendiquer une approche en linguistique de corpus. Ceci peut être le cas notamment en linguistique diachronique où l'on travaille régulièrement à partir de données formant un ensemble relativement restreint et incomplet (Combettes, 2014).

Si la légitimité de la linguistique de corpus comme *discipline* est discutée par certains, c'est en partie pour des raisons historiques : le choix des orientations et des visées théoriques, mais aussi la faible quantité de données en jeu dans les premiers travaux. Chomsky (1979) parle de « collecte de papillons » pour qualifier le travail sur données attestées réunies ; plus tard (cité dans Aarts, 2001), il dira de la linguistique de corpus qu'elle « n'existe pas ». Labov (1971), quant à lui, tout en critiquant l'introspection comme méthode d'investigation, reconnaît les limites des collections de données (orales). Et Milroy (1987 : 5), à propos des données récoltées sur le terrain,

[1] Parfois aussi linguistique *sur* corpus (par ex. Bilger, 2000). Voir aussi Bilger et Cappeau (ce volume).

reconnaît que celles-ci sont insuffisantes pour saisir de façon adéquate la grammaire de la langue, même si dans la tradition descriptiviste, la découverte de langues « nouvelles » (ou en tout cas non décrites) se fait de façon empirique et ne semble pas poser problème. Or, comme le souligne Tognini-Bonelli (2010), il arrive un moment dans les années 1990 où la linguistique de corpus commence à s'affirmer comme véritable discipline, tant les méthodes et les techniques ont évolué (McEnery et al., 2006 : 8).

Il en va de même de la quantité de données : si dans les années 1980 on considérait encore que les corpus étaient trop petits pour décrire *la* grammaire d'une langue, les choses ont évolué considérablement et il existe aujourd'hui bon nombre de descriptions et de grammaires qui se prévalent d'être faites à partir de corpus (à commencer par les travaux sur l'anglais, dont ceux de Biber et al., 1999).

En ce qui concerne le nombre de mots que contient un corpus, on les compte souvent aujourd'hui en millions voir en milliards d'occurrences dans certains cas (ce qui n'exclut pas de travailler sur des corpus plus petits comme nous le verrons). Ainsi, par exemple, le *British National Corpus* (BNC, http://www.natcorp.ox.ac.uk/) constitué vers 1990, avec ses 100 millions de mots, constitue aujourd'hui un corpus qui n'impressionne plus par sa taille (alors qu'il passait pour un géant à l'époque aux côtés du corpus Brown d'un million de mots – voir Kučera & Francis, 1967) ; c'est même un petit corpus comparé à d'autres projets (cf. les corpus WaCKy, par exemple – http://wacky.sslmit.unibo.it/doku.php ; Baroni & Bernadini, 2006).

Mais la dimension quantitative n'est pas la seule mesure légitime ; en effet, un corpus bien ciblé au niveau de son contenu peut être plus petit tout en étant pertinent pour certains types d'analyses (Koester, 2010). Par ailleurs, d'autres considérations s'appliquent à la description des données réunies sous forme de corpus, comme :

– la disponibilité des textes (cf. le cas de la linguistique historique) ;

– le nombre de locuteurs : on ne peut avoir des millions de mots de productions d'enfants si le public sur lequel on travaille ne dépasse pas les 2 ou 3 locuteurs, par exemple ;

– le problème que pose la transcription, qui est une opération très chronophage pour les corpus oraux.

Concernant le travail sur la langue parlée, certains (surtout en France) font du mot *corpus* un quasi-synonyme de *corpus de données orales*, étant donné l'héritage des enquêtes (à commencer par la dialectologie, mais incluant les travaux sur les interactions et sur la grammaire de l'oral, par exemple). Pour d'autres (cf. Cappeau & Gadet, 2007a, 2007b, 2010), le développement de la linguistique de corpus, couplé au développement des technologies permettant de stocker mais aussi de recueillir des textes, a un peu mis de côté la langue orale. Certes, les moyens technologiques dont nous disposons aujourd'hui servent à mettre en avant des études naguère impossibles car portant sur des masses de données ne pouvant être traitées manuellement. Mais les grands corpus d'aujourd'hui doivent leur grandeur essentiellement à la composante

écrite et non orale. En effet, la collecte, transcription et mise en forme de données orales sont coûteuses et la composante orale des corpus existants se limite bien souvent à des transcriptions d'émissions de radio ou de télévision (voir par exemple les corpus BYU pour l'anglais, l'espagnol et le portugais – http://corpus.byu.edu/ ; voir aussi le corpus CREA pour l'espagnol – http://corpus.rae.es/creanet.html). Néanmoins, Forchini (2012) trouve que les scripts des films et de séries en anglais (américain) sont proches de la langue parlée spontanée. Cependant, la question de la transcription des données reste un problème, tant au niveau de la fiabilité des transcriptions qu'au niveau des conventions (qui sont rarement les mêmes d'une étude à l'autre, puisque fonction de la visée de recherche). Une autre dimension à prendre en compte lorsque l'on travaille sur de l'oral est celle de la méthode de recueil des données : à l'heure actuelle, les démarches dites « écologiques » sont mises en avant afin de privilégier des recueils de données les plus authentiques possibles (Gadet et al., 2012 ; Tyne et al., 2014).

3. Corpus et acquisition

Si l'un des buts du travail sur corpus est de démontrer l'existence de faits langagiers avérés, dans le cadre de l'acquisition du langage (L1 comme L2), il est possible d'utiliser les corpus pour construire des normes descriptives portant sur l'acquisition en général, des cadres ou des parcours-types expliquant comment se développe le langage chez les enfants ou les apprenants. De plus, l'utilisation de corpus dans le cas de l'acquisition de L1 ne fournit pas seulement des données sur l'acquisition du langage, elle représente également un modèle de la connaissance langagière différent du modèle social du langage. On pourra parfois constater que les tendances et productions sont très différentes d'un enfant à l'autre, d'un apprenant à l'autre, tout en dégageant des propriétés générales. Cette approche de la connaissance linguistique d'un individu (L1 ou L2), qui se différencie de l'approche du langage en général, pour tous les individus, et qui fait partie des propositions de la linguistique cognitive (Goldberg, 2006), ne peut être démontrée ou étudiée sans la notion de suivi de corpus, qu'il s'agisse d'enfants ou d'adultes.

3.1. Corpus et acquisition de L1

L'étude de corpus d'acquisition de L1 a débuté dans les années 1960 avec en particulier les travaux pionniers de Braine (1963), Brown (1973), et Bloom (1970) et tous les travaux qui ont fait suite. Auparavant, la plupart des travaux dans le domaine de l'acquisition de langage étaient basés sur le recueil de journaux (voir Morgenstern & Parisse, 2007 ; Delefosse, 2010). Braine, Brown et Bloom restent aujourd'hui des références incontournables du domaine de l'acquisition du langage grâce à la qualité de leurs travaux, tous basés sur des corpus dits longitudinaux. Il s'agit de corpus qui comprennent des enregistrements successifs d'un même enfant sur une période assez longue pour voir son langage se développer. Les fréquences de recueil de données sont variables en fonction du type d'étude et de l'âge de l'enfant. La plupart des premiers travaux sont basés sur des fréquences de l'ordre d'un enregistrement d'une heure toutes les deux semaines ou tous les mois. Les données sont recueillies en interaction

naturelle entre l'enfant et des partenaires langagiers, ce qui permet une étude directe du développement du langage mais qui pose des problèmes spécifiques, aussi bien pratiques que théoriques. Les interlocuteurs peuvent être les parents de l'enfant comme des investigateurs scientifiques. Dans ce dernier cas, on utilise souvent une situation de jeu, ce qui permet à l'enfant de produire un langage naturel et créatif avec un partenaire inconnu (ou peu connu).

Les problèmes ou difficultés du travail sur corpus de L1 ne sont pas seulement liés au recueil de données. Ils sont pratiques comme théoriques et touchent aussi bien la technique, la linguistique, que l'épistémologie des sciences du langage et des sciences cognitives en général (Canut & Vertalier, 2008). Toutes ces questions sont riches en enseignements et permettent d'ouvrir des fenêtres sur la compréhension du domaine de recherche et sur les conclusions à tirer des résultats obtenus. Ces questions ne sont d'ailleurs pas uniquement posées dans les corpus d'acquisition de L1.

La disponibilité de corpus d'enfants en situation d'acquisition de L1 est fondamentale pour le travail de recherche et d'évaluation du langage. Il faut en effet être conscient du coût de tels corpus. Filmer ou enregistrer et surtout transcrire est une opération qui prend beaucoup de temps, en particulier dans le cas de jeunes enfants dont le langage est en développement et qui peut de ce fait être difficile à transcrire. Par exemple, le corpus COLAJE qui contient plus de 160 heures de transcriptions (Morgenstern & Parisse, 2012) représente environ 3 à 4 ans passés uniquement à transcrire. C'est pour cela que se sont développées des initiatives pour diffuser les corpus d'acquisition de langage. La plus importante d'entre elles est celle du site CHILDES (http://childes. psy.cmu.edu) qui, hébergé aux États-Unis, a une portée tout à fait internationale. Ce site propose des corpus dans plus de 32 langues pour un total de plus de 44 millions de mots. Il contient et met à disposition 15 corpus d'acquisition de français L1 pour un total d'environ 2,8 millions de mots (enfants et adultes compris), ce qui représente environ un millier d'heures de langage transcrit. Certains corpus bilingues ou de L2 sont également disponibles. Tous ces corpus peuvent être utilisés librement à condition de citer les auteurs qui les ont fournis. Les conventions d'usage et de dépôt des corpus sont décrites sur le site TalkBank (http://talkbank.org/share/) et basées sur une licence Creative Commons (http://creativecommons.org/licenses/by-nc-sa/3.0/). D'autres initiatives existent, mais n'ont pas encore aujourd'hui la même ampleur comme la partie « enfants » du projet TCOF (http://cnrtl.fr/corpus/tcof/ – André & Canut, 2010) ou le site du projet COLAJE (http://colaje.scicog.fr/).

3.2. Corpus et acquisition de L2

En ce qui concerne l'acquisition de L2, comme pour l'acquisition de L1, le travail d'analyse porte sur des données collectées auprès d'apprenants. Les premiers travaux sur l'acquisition des L2 s'intéressent surtout à l'influence de la L1 : l'apprenant appliquerait des règles de sa L1, créant ainsi des productions « erronées ». Mais lorsque se développe l'étude systématique des erreurs dans les productions d'apprenants, on commence à voir des éléments difficilement attribuables à la seule influence de la L1 ;

il s'agit de formes propres au « dialecte idiosyncrasique » (Corder, 1971) ou « système approximatif » (Nemser, 1971) que développe l'apprenant. On parle alors d'« interlangue » (Corder, 1971 ; Selinker, 1972). Cette « variété de l'apprenant » est ainsi prise comme objet d'étude *per se*, au même rang que toute autre langue naturelle (Adjémian, 1976), ce qui implique nécessairement le développement d'une méthodologie d'observation des données attestées.

Dans le domaine de l'étude de l'acquisition en L2, une large partie de ce qui peut être évoqué plus généralement comme « problèmes » ou « difficultés » pour la prise en compte de l'oral s'applique également, qu'il s'agisse de l'acte d'enregistrer, de transcrire, de coder, ou bien d'étudier des données (densité, « erreurs », question de la norme, etc.). De plus, dans l'histoire du développement des travaux sur l'acquisition des L2, la question de la « sollicitation » des données (Corder, 1973) paraît centrale pour la prise en compte des données en amont de la collecte. Cet aspect est d'autant plus important que les travaux des acquisitionnistes ciblent les types de production des apprenants les plus à même de rendre compte du développement de l'interlangue. On pense en particulier aux nombreuses études portant sur des tâches comme la narration à partir de stimuli tels que le livre pour enfants *Grenouille, où es-tu ?* (« *Frog, where are you ?* » – Mayer, 1969) ou le film *Les temps modernes*.

Si le recours aux données sollicitées (pratique plus présente dans les travaux en L2 qu'en L1 du fait de la complexité des données « naturelles » en L2 notamment – voir plus loin) pose problème pour qui voudrait une démarche plus écologique, elles s'imposent encore comme méthode de constitution de corpus de L2 (du moins c'est le cas pour l'oral). Premièrement, ces données permettent d'interroger de manière directe le système linguistique (approche plus ou moins complète) que développe l'apprenant sans passer par des tests ou des questions faisant confiance aux intuitions des apprenants. Deuxièmement, elles permettent de cibler (surtout en quantité) certains éléments linguistiques précis qui ne seraient peut-être pas présents si on devait attendre qu'ils soient produits naturellement dans le discours des apprenants (voir plus loin la question de la densité). Troisièmement, il est essentiel d'avoir des productions comparables d'un apprenant à l'autre, d'un groupe à l'autre. Mais il existe aussi la possibilité de travailler sur des données plus écologiques, notamment pour l'étude des écrits en L2, où l'on se sert de productions existantes (des devoirs comme les dissertations, par exemple) pour observer des phénomènes à l'écrit (Dubois et al., 2014).

3.3. Acquisition et travail sur données attestées : questions de recueil et de transcription

Si les données écrites ont contribué (et contribuent encore) à l'étude de l'acquisition, l'oral constitue un domaine privilégié dans la mesure où (surtout en L1) c'est le premier code à se développer chez l'enfant qui acquiert le langage[2]. Le recueil

[2] Si dans beaucoup de contextes c'est aussi le cas pour l'acquisition de L2, ce n'est pas nécessairement toujours ainsi, notamment dans des contextes d'apprentissage de langues proches où les formes écrites sont, pour des apprenants maîtrisant déjà le code écrit dans leur L1, plus accessibles que les formes orales. Pour ce qui relève du travail sur l'écrit, les questions de recherche concernent à la fois

de corpus pour l'étude de l'acquisition vise la plupart du temps à recueillir des données les plus naturelles possibles, tout en acceptant bien souvent des contraintes imposées par le protocole de recherche (ciblage de certains types de productions, utilisation d'entretiens, etc.). Pour cela, l'enregistrement audio utilisé dès les premiers corpus des années 1960 est parfaitement adapté. Il y a 40-50 ans, il était déjà possible de réaliser un enregistrement audio très discret (voir dans Wells, 1985). Cependant, l'audio est souvent insuffisant pour réaliser une étude correcte et surtout complète du développement du langage, en particulier dans le domaine de l'acquisition des L1 où l'interaction enfant-adulte est souvent la cible des études. D'une part, sans support visuel il est souvent difficile de réaliser des transcriptions fines (à moins de disposer d'un carnet de bord très détaillé, et encore…). D'autre part, l'étude du développement du langage ne porte pas seulement sur les parties verbales des interactions, mais aussi sur le pointage, les mouvements du visage, etc. (voir ci-dessous pour ces deux points ; voir aussi Canut et al., ce volume). Le recueil vidéo est devenu très bon marché depuis quelques années et très aisé à intégrer dans les transcriptions de corpus depuis l'avènement des caméras numériques qui suppriment les difficultés liées à la numérisation.

Mais on pourrait se poser la vieille question, qui a fait l'objet de bien des réflexions chez les sociolinguistes, qui est celle du paradoxe de l'observateur : comment peut-on enquêter, enregistrer ou filmer alors que la seule présence du microphone ou de la caméra altère le côté « naturel » des données ? On pourrait aussi se demander dans quelle mesure les enfants (dans le cas de l'étude de l'acquisition de L1 surtout) ne seraient pas trop perturbés par un dispositif d'enregistrement lourd. Pour ce qui est des apprenants de L2, si la même question se pose, elle est incluse dans la question plus générale de la « mise en scène » du recueil de données : on est plus ou moins obligé de faire parler en L2 pour avoir des données car les interactions « naturelles » des apprenants ont plutôt lieu dans leur L1 (ou quelque autre langue de communication communément utilisée par le groupe d'apprenants).

L'utilisation de corpus présuppose un échantillonnage de données langagières étudiées. Une des difficultés du domaine est de mesurer la représentativité de ces échantillons (Tomasello & Stahl, 2004). En effet, si on étudie, dans le cadre de l'acquisition de la L1, un phénomène qui se produit une fois par jour, sachant qu'un enfant peut avoir en moyenne 10 heures par jour d'interaction langagière, dans le cas d'observations ayant lieu une heure par mois, on n'aura qu'une chance sur 300 d'observer ce phénomène (10 x 30 jours). Si un phénomène se produit 30 fois par jour (c'est-à-dire 3 fois en une heure), on n'aura qu'une chance sur 10 de l'observer dans un recueil effectué une fois par mois. Il convient donc lors de la réalisation d'une étude de corpus de connaître ou de mesurer à l'avance la fréquence des phénomènes étudiés afin d'ajuster l'échantillonnage au mieux. On constate donc que le travail sur corpus peut être coûteux dans le cas de l'étude de phénomènes rares. La difficulté de l'observation des phénomènes peu fréquents est particulièrement problématique dans l'observation de ce qu'on appelle en général le « non typique », c'est-à-dire tous les phénomènes qui ne correspondent pas à la cible

---

la maîtrise de la langue mais aussi (et surtout parfois) la maîtrise des règles d'orthographe de cette langue (voir Dubois et al., 2014 ; voir aussi Dubois et al., ce volume).

adulte attendue, cette cible représentant souvent la « norme ». Dans cet ensemble on trouve en particulier les incertitudes sémantiques (emploi d'un mot pour un autre), ou les sur-généralisations (emploi d'une forme en utilisant une règle générale qui ne s'applique pas dans un cas précis : par exemple, la production de « prendé » sur le modèle des verbes du premier groupe au lieu de « pris »). Le terme « non typique » sera préféré ici à « erreur » (cf. ci-haut) car ce dernier sous-entend la référence à une norme ou un résultat valide, ce qui n'existe pas encore chez l'enfant qui développe sa L1. Au contraire, toute production est le résultat d'un processus sous-jacent, qui peut être temporaire chez le jeune enfant mais qui n'est ni juste ni faux. Lorsque l'enfant ou l'apprenant produit un élément non typique, il peut être repris ou explicitement corrigé par un adulte ou un enseignant. A ce moment, il peut ou non selon les circonstances tenir compte de cette information pour faire évoluer son propre système langagier (Clark, 2009). Il peut aussi ignorer la correction explicite, et son système peut évoluer avec l'usage.

La problématique très vaste et très instructive de la production non typique pose le problème de la représentativité des corpus de langage spontané qui ne sont pas toujours assez importants pour permettre l'observation de l'exceptionnel. Par exemple, Tomasello et Stahl (2004) expliquent que dans le cas de sur-généralisations rares, un corpus très peu dense ne permet pas de mettre en opposition deux théories. Ceci est aussi vrai pour les créations non standard des enfants. Elles sont peu fréquentes et leur étude systématique est complexe dans un corpus de langage spontané. C'est pour cela que certains corpus sont basés sur des productions sous forme d'entretiens semi-dirigés ou dirigés. On met l'enfant dans une situation précise qui amène à un type de langage, ou on procède par interactions dirigées avec un adulte, par exemple avec un système de questions et de réponses ou de situations expérimentales. Pour ce qui relève du domaine de l'acquisition de la phonologie, en revanche, le recours à des données « naturelles » paraît tout à fait envisageable (voir Liégois et al., ce volume).

Quand on regarde les données utilisées pour étudier l'acquisition, les corpus qu'elles forment sont souvent de taille modeste, malgré la richesse des données et leur pertinence. De plus, les données ne sont pas nécessairement comparables d'une étude à l'autre et le nombre de participants enregistrés est souvent réduit (ne portant que sur quelques apprenants, par exemple) ce qui ne permet pas d'observer de grandes tendances chiffrées comme cela peut être le cas dans une démarche ordinaire aujourd'hui en linguistique de corpus. Or, le domaine de recherche sur l'acquisition se distingue clairement de la linguistique de corpus en ce sens qu'il attache une grande importance au sujet parlant dans les productions (bien plus qu'un simple support permettant d'accéder à une masse de données pour décrire *la* grammaire, par exemple, même si des généralisations sont possibles). Ceci paraît d'autant plus important que les études portant sur l'acquisition attachent une très grande importance aux métadonnées : qui parle (quel âge ?), où, quand, avec qui, etc. ? Pour les millions de mots que comptent la majorité des grands corpus modernes, les détails des locuteurs eux-mêmes ainsi que du terrain d'enquête sont souvent inconnus ou tout simplement considérés comme non pertinents. Mais pour les études traitant de l'acquisition, ces éléments paraissent d'autant plus importants que les analyses se trouvent liées au milieu d'enquête en ce

qu'il y a un développement logique à partir de la position théorique adoptée, en passant par l'élaboration d'une enquête et la réflexion autour du terrain et la représentation des données.

En ce qui concerne la prise en compte des données orales, se pose la question du rapport à la transcription car tout corpus oral nécessite par définition une transcription, c'est-à-dire la traduction du son ou de l'image dans un ensemble de codes prédéfinis (ce qui permet bien souvent de recourir à des analyses qui sont essentiellement les mêmes que pour les données écrites, avec les mêmes outils). La question de la transcription, ainsi que les problèmes qu'elle présente ne sont pas spécifiques aux données d'apprenants (L1 comme L2). Toutefois, la question de la transcription se pose particulièrement, tout comme celle du rapport à la norme (*a fortiori* écrite), dans l'élaboration du corpus. Pour la partie langage d'un enregistrement, le code prédéfini est très souvent la glose orthographique, ou parfois la transcription phonétique ou phonologique intégrale, mais aussi l'intonation, la valeur sémantique et pragmatique. Beaucoup d'autres codes existent, par exemple la description des pointages, des gestes co-verbaux, des centres d'attention, de la situation, des objets manipulés, etc. La particularité de ce processus est qu'il est largement sujet à incertitude et qu'il existe des différences parfois importantes d'un codeur à l'autre. Si ce même problème existe pour la transcription des données en général, comme l'ont mis en avant les travaux précurseurs du *Groupe aixois de recherches en syntaxe* (GARS – voir Blanche-Benveniste & Jeanjean, 1987), il apparaît comme particulièrement problématique pour ce qui relève de la transcription de la parole en acquisition, qu'il s'agisse d'enfants acquérant leur L1 ou qu'il s'agisse de locuteurs non natifs acquérant une L2. Le type de transcription dépend essentiellement du type d'analyse à effectuer.

Plus particulièrement se pose la question de la transcription des formes non typiques ou « erronées » (cf. la prise en compte de la variation et la transcription de formes non standard en général – voir ci-dessous) dont la représentation transcrite peut être problématique dans la mesure où on met en jeu plus d'un système linguistique (et plus d'un système orthographique éventuellement lors des transcriptions). En effet, pour des apprenants de L2 peu avancés, dont les productions sont très marquées par des influences ou des transferts de la L1, ou bien des bilingues qui font de l'alternance codique ou des mélanges de langues, on peut se demander quelle est la langue de base dans certains cas : quelle est donc la norme, et que faire des formes qui ne la respectent pas ? Si le recours (autant que faire se peut) aux formes standard de la langue cible semble pour un certain nombre de raisons s'imposer (cf. le risque de faire passer les apprenants pour de « mauvais » parleurs ; le risque de non-repérage automatisé de formes ou d'étiquetage erroné ou partiel), de nombreuses questions se posent encore pour la transcription. Comment, par exemple, transcrire l'extrait suivant produit par un apprenant anglophone de français L2 : « des filles qui sont déjà si [grɑ̃] et déjà plus [aʒe] ». Faut-il transcrire *âgés* sans accord de genre, considérant que la présence de deux attributs sans accord est une preuve de masculinisation de *filles* ? Faut-il restituer l'accord dans la logique de l'orthographe standard ? Ou doit-on systématiquement recourir à des représentations phonétiques

(comme nous l'avons fait ici pour l'exemple) ou quelque autre forme de codage au risque de rendre d'éventuelles recherches automatiques de formes difficiles ?

Transcrire de manière orthographique, c'est interpréter, comprendre, une situation langagière comme nous le faisons dans les interactions tous les jours. La fonction du transcripteur est donc en quelque sorte de reproduire la fonction de l'interlocuteur et de fournir une donnée sur laquelle le scientifique va travailler. Et on peut, grâce à cette réflexion sur le rapport au code écrit dans la transcription de données pour l'étude de l'acquisition, s'interroger de façon générale sur le rôle et le statut particuliers qu'occupe l'écrit dans les cultures de littératie : le fait même de faire un acte de littératie tend à prêter une certaine « légitimité » au produit final. Ainsi, lorsqu'on transcrit orthographiquement, par exemple, quelle que soit la précision avec laquelle le produit graphié reflète ce qui a été dit, on rend durable l'éphémère sous une forme ordonnée. Ceci a pour effet d'affecter l'attention portée à la qualité orale des productions en laissant s'introduire des valeurs traditionnellement associées à l'écrit et aux contraintes de l'écriture.

Inversement, l'utilisation d'un modèle, quel qu'il soit, probablement plus avancé que ne l'est le langage de l'enfant à un âge donné, permet de réaliser des comparaisons entre différents niveaux d'âge, entre l'enfant et l'adulte, et entre plusieurs enfants. Par exemple, réaliser un étiquetage morphosyntaxique classique ne doit pas laisser penser que l'enfant très jeune maitrise les catégories grammaticales. Mais cet étiquetage permet de comparer l'usage de l'enfant à l'usage de l'adulte, ou d'un enfant jeune à un enfant plus âgé, ou d'un enfant français à un enfant d'une autre langue maternelle. En dépit des limitations inhérentes au type de transcription retenu ainsi que le codage (linguistique ou autre), les données transcrites restent une source fabuleuse d'informations sur le développement du langage. Dans certains cas, elles appellent à des compléments d'étude, par exemple par des études expérimentales de compréhension ou de production grammaticale pour mieux évaluer ce que l'enfant comprend ou sait généraliser.

Ainsi, et conformément à l'idée que le corpus ne doit pas contenir d'informations inutiles, superflues à la prise en compte des données (sauf si le logiciel de transcription permet d'insérer des commentaires ou des explications – comme c'est cas dans les transcriptions CHAT[3] par exemple), on privilégie l'orthographe la plus régulière, la moins déformée possible. Et dans tous les cas, il est conseillé lors du travail de transcription de mettre en place une vérification multiple des transcriptions par plusieurs personnes. Il est également intéressant de mesurer la qualité des accords inter-juges, c'est-à-dire des accords et désaccords entre plusieurs transcriptions réalisées par des personnes différentes.

[3] Transcription CHAT dans CHILDES : http://childes.psy.cmu.edu/.

4. Corpus en didactique

En dehors des préoccupations des linguistes (ou des sociolinguistes, acquisitionnistes, etc.) descriptivistes, le terme *corpus* est employé en didactique depuis un certain nombre d'années également. Par exemple, pour Holec (1990), le corpus de l'apprenant est la collection de documents dont dispose l'apprenant en vue de découvrir ou d'observer des faits de langue (cf. Boulton, 2009). Cette acception du terme est présente dans le discours éducatif pour désigner des collections de textes utilisés dans ou pour l'apprentissage (Eluerd, 1979 : 90). Pour Duteil-Mougel (2007), le *corpus* dans l'enseignement secondaire correspond davantage à des collections de « morceaux choisis » qu'à des collections exhaustives et/ou ciblées de textes entiers pouvant être pris pour un corpus au sens linguistique. Il s'agit donc d'un terme qui sert en quelque sorte à délimiter les contours de l'unité thématique par le biais de la documentation (choisie comme un ensemble plus ou moins homogène sur le plan linguistique ou discursif) ou qui sert à limiter tout simplement le nombre d'œuvres (ou documents iconographiques, etc.) sur lesquelles les apprenants ou élèves auront à travailler – le *corpus* représente alors une liste d'œuvres correspondant à l'unité d'apprentissage (Duteil-Mougel, 2007). Mais il est possible néanmoins de voir ici un type de regard sur les données qui n'est pas en contradiction avec ce que nous avons vu plus haut : il s'agit dans tous les cas d'aborder la langue à travers un certain nombre de points d'accès plutôt qu'à partir de l'exemple unique.

Ce que l'on retient de cette utilisation du terme est à la fois le sens de « collection » (ce qui permet une certaine démarche méthodologique pouvant s'apparenter à celle inspirée de la linguistique de corpus), mais c'est aussi le caractère fortement « matériel » (et non électronique) du corpus. Par ailleurs, il peut s'agir d'une collection potentiellement très hétérogène de documents, dont des éléments peu ou pas textuels. On peut aussi souligner la question de la taille : en didactique, un corpus peut contenir quelques documents (ou extraits) seulement, le tout faisant à peine quelques centaines de mots. Néanmoins, comme le soulignent Boulton et Tyne (2014), le fait de travailler sur une quantité de textes limités, tout comme le fait d'avoir accès à un corpus matériel, « palpable », peut même être un atout en didactique des langues tant la proximité des données semble être un facteur motivant dans l'apprentissage. Et on peut même aller plus loin en insistant sur l'aspect méthodologique du travail sur corpus ou bien du travail de constitution de corpus (Tyne, 2009). C'est donc avant tout la méthodologie de la découverte qui est mise en avant plutôt que le travail linguistique sur le corpus fini, même si à travers la pratique que l'on nomme en anglais « *data-driven learning* » (Johns & King, 1991) on rencontre des approches qui s'apparentent à de la linguistique de corpus en didactique (cf. Landure, 2014).

4.1. Corpus et didactique de L1

Comme pour la didactique de la L2, la didactique de la L1 s'appuie traditionnellement sur un ensemble de textes réunis, permettant aux élèves d'observer, de manipuler et de comprendre des faits de langue envisagés comme pertinents pour

rendre compte du fonctionnement de la langue. Cependant, on ne peut avoir affaire à des corpus similaires étant donné les perspectives et les publics concernés par l'un et l'autre champ. Néanmoins, les frontières entre les champs ne sont pas infranchissables : au contraire, notamment à propos des « corpus de travail », on aurait tout à gagner à croiser les réflexions sur la didactique de la L1 et la didactique de la L2 (voir par exemple à propos du français, Cadet & Guerin, 2012).

Au terme d'une note de synthèse tout à fait éclairante sur les rapports entre pratiques langagières et scolarisation, Bautier (2001 : 161) suggère qu'après trente années à s'interroger sur les raisons des difficultés rencontrées par certains élèves, il serait pertinent que les chercheurs se centrent sur « les moyens d'y remédier ». Interroger le corpus de travail en L1 peut constituer une piste, étant donné que les motivations de sa constitution sont, pour une grande part, implicites (voir à propos du « corpus scolaire » pour l'enseignement des Lettres, continuité de l'enseignement du français L1 dans le second degré, en France, Viala, 2014).

En théorie, le choix des textes devrait être motivé par le fait que les élèves sont déjà locuteurs de la langue étudiée : l'objectif serait donc de donner à voir des productions dans le but d'enrichir et de développer des compétences plutôt que d'appréhender un nouveau système (cf. Eluerd, 1979 : 42 pour la spécificité de l'oral). Le « corpus de travail » ne devrait donc pas être perçu par les élèves comme illustrant des faits langagiers sans rapport avec les savoirs sur lesquels ils s'appuient quotidiennement pour communiquer. Cela implique premièrement que l'on pose l'hypothèse du niveau de compétence des élèves pour un niveau de classe donné, afin de déterminer sur quelle base il est raisonnable de s'appuyer pour l'enrichir/le développer ; et deuxièmement que le corpus ainsi établi soit adapté et qu'il permette aux élèves d'accéder à la compréhension des savoirs en jeu relativement à ce même niveau de compétence, c'est-à-dire que les textes sélectionnés permettent une mise en relation desdits savoirs et de ce qui est ou sera potentiellement appréhendable en dehors du cadre scolaire. La didactique de la L1 concerne un public (élèves et enseignants) pour qui la langue enseignée n'est pas exclusivement observable et pratiquée dans le cadre scolaire. Ceci ne vaut que dans une approche de la langue qui intègre sa variabilité et sa dynamique. Autrement dit, on ne peut pas penser que les élèves s'appuieraient sur leurs expériences langagières extra-scolaires pour développer et enrichir leurs compétences si on ne considère pas que leurs productions quotidiennes et spontanées illustrent, au même titre que les productions sur lesquelles reposent l'enseignement (le corpus), des aspects d'une même langue.

Ceci étant dit, pour comprendre toute la complexité de l'enseignement des savoirs relatifs à la L1 et de la sélection d'un « corpus de travail » pertinent, c'est la notion même de *langue* qu'il faut interroger. La L1 renvoie, dans le cadre terminologique de l'enseignement, à la langue dite « maternelle » (LM). L'expression, problématique à bien des égards, évoque, dans les faits, plusieurs réalités. Herlitz et al. (2007), s'appuyant notamment sur les travaux de Gagne, les distinguent ainsi :

– la « *home language* », la langue des premiers échanges, développée dès l'enfance, avant les apprentissages scolaires ;

– la langue du « *fatherland* », qui s'inscrit à un niveau politique et culturel (par opposition à la première acception qui s'inscrit à un niveau individuel, même si les deux niveaux sont nécessairement imbriqués), qui conditionne l'identité régionale ou nationale ;

– la langue en tant qu'objet de l'enseignement destiné à ses locuteurs, qui se confond avec la forme standard.

Si les deux derniers aspects de la notion peuvent, au moins dans la plupart des pays d'Europe de l'Ouest, concerner la même forme de la langue, ce que recouvre le premier aspect renvoie à des formes hétérogènes, imprévisibles, qui ont en commun un écart plus ou moins grand avec la forme standard. Pourtant, si l'on parle, dans tous les cas, de LM c'est qu'il est toujours question de la langue dans laquelle évoluent et se construisent les locuteurs dont l'identité est à la fois individuelle et collective. Penser une LM revient donc à penser l'imbrication des trois aspects de la notion. Du point de vue de l'individu, elle se joue sur le plan des représentations. Même si la « *home language* » n'est jamais la forme standard, elle n'est jamais ignorée. Elle influence plus ou moins fortement, par imitation ou par opposition, selon les cas, les pratiques familières. Du point de vue du collectif, où se situe la question de l'enseignement, les effets de l'imbrication ne relèvent pas de l'influence de la LM « individuelle » : la conceptualisation de la forme standard n'intègre les pratiques individuelles qu'au titre de contre-exemples, puisque, par définition, le principe de standardisation s'oppose au principe de variation : « *The 'standard language' interpretation of mother tongue furthermore disregards the many regional and local varieties of that standard, it disregards the multilingual construction of nowadays society.* » (Herlitz et al., 2007 : 16). Comme le souligne Halliday (2007 : 28), l'enseignement de la LM, dans sa fonction prescriptive, n'ajoute rien à la performance de l'élève mais la rend plus socialement acceptable. Ainsi, on envisage la question du « corpus de travail » comme une sélection exclusive de textes illustrant le modèle légitime, valorisé et valorisable.

Qu'en est-il de l'enseignement du français langue (dite) maternelle (FLM) ? On peut dire, d'un certain point de vue, que le corpus de textes sélectionné s'inscrit bien dans la démarche d'enrichissement des compétences des élèves puisque, en étant constitué pour l'essentiel de données illustrant la forme standard, il permet la découverte et l'acquisition de savoirs qui ne sont *a priori* pas appréhendables sans un certain encadrement. À ce titre, et quand on sait la valeur sociale attribuée à quiconque en a la maîtrise, un tel corpus semble incontournable. En revanche, on voit moins clairement quel rapport ce corpus entretient avec l'ensemble des autres productions langagières. Théoriquement, on a affaire à un ensemble d'actualisations d'une forme de langue adaptée aux contraintes relatives à un type de situations caractérisées par une grande distance symbolique et/ou physique entre les interactants, à l'image de la relation entre l'auteur d'un roman et un lecteur (Guerin, 2008). Ainsi présenté, le corpus ne tend pas à la représentativité de la langue (si tant est qu'un tel objectif soit atteignable) mais à l'illustration d'un type de productions particulières et situées. De fait, certains aspects ne sont pas pris en charge par la description grammaticale scolaire, c'est-à-dire observables dans le corpus de textes exploité, bien que massivement

observables dans les usages ordinaires (tel est le cas de la plupart des unités non standard que des corpus conçus par des linguistes ont révélées). Inversement, d'autres faits sont très bien représentés dans le corpus alors qu'ils sont effectivement absents des usages ordinaires (par exemple, l'emploi du passé simple, la présence de *ne* dans des constructions négatives, les interrogations par inversion du sujet…). C'est le cadre de l'enseignement (a fortiori de la LM), sa fonction sociale, qui impose une restriction du champ des observables au modèle de référence. Comme le soulignent Chiss et David (2012 : 138) à propos de l'étude de la langue en FLM, « l'appui sur de gros corpus permet de complexifier les données : vision pertinente sans doute de la réalité linguistique mais qui peut conduire à une antinomie didactique et pédagogique (quelle forme de stabilisation du savoir à transmettre ?) tout en reposant le problème épistémologique chomskyen – et néanmoins incontournable – de la prédictibilité ».

Bien que l'on admette que les enjeux socio-politiques de l'enseignement de la LM contraignent le corpus de travail à l'illustration de la forme standard, on peut cependant se demander comment il est introduit dans le discours scolaire pour qu'il soit perçu par les élèves comme relevant de leur langue (dite maternelle). En l'occurrence, les textes sont montrés en exemple, des exemples de la langue « correctement » employée, au point où il est en fait question d'affirmer, plus ou moins explicitement, que le corpus représente la langue. De fait, pour les élèves, ce que donne à voir le discours scolaire est bien mis en lien avec ce qu'ils sont en mesure d'observer par ailleurs, en dehors du cadre scolaire, mais exclusivement en tant que cela constitue la façon correcte et attendue d'actualiser la langue en toute situation (voir par ex. Boutet, 2002). Autrement dit, s'il y a bien un rapport entre leurs compétences de locuteurs/auditeurs du français et l'objet d'enseignement du FLM, il ne s'établit que dans une conception hiérarchique excluante, le non-standard n'apparaissant que comme un dérivé, une dérive, d'une forme absolue, illustrée par l'ensemble de textes sur lequel s'appuie la description grammaticale scolaire. Dans ces conditions, ce qui était présenté comme les spécificités de la didactique de la L1 au début de cette section n'est plus valable puisqu'à une représentation horizontale de la variation de la langue qui plaçait sur le même plan les formes de langue illustrées par le « corpus de travail » et les autres formes (parmi lesquelles celles maîtrisées par les élèves, ce qui permettait de penser l'enseignement comme un élargissement des compétences par accumulation de savoirs), on préfère une représentation verticale, où l'enseignement vise une ascension conditionnée par l'abandon de tout ce qui n'est pas standard.

Il semble que ce qui peut être un problème dans l'usage d'un corpus (tel qu'il a été défini dans cette section) dans le cadre de l'enseignement du FLM n'est pas la sélection des textes, qui est pertinente puisqu'elle permet d'aborder un code peu voire non accessible. Les savoirs en jeu peuvent ainsi constituer des éléments d'enrichissement des compétences linguistiques des élèves. Cependant, pour que ces éléments soient effectivement intégrés en tant que tels, qu'ils participent du développement de compétences déjà en place, il n'est pas possible de leur attribuer un statut exclusif. Il serait alors davantage question d'un appauvrissement, sauf à adhérer à une certaine idéologie, l'idéologie du standard. La seule maîtrise de la forme « légitime », celle

qu'illustre ce corpus, ne peut suffire à la « maîtrise de la langue », pour reprendre l'expression consacrée dans les textes officiels, dans sa dimension dynamique. Il n'est pas inutile de rappeler que le français est une langue vivante : dans les textes officiels qui encadrent l'enseignement scolaire, l'expression ne vaut jamais pour le français.

En somme, l'exemple de la didactique du FLM fait apparaître une contradiction : alors que les élèves sont par définition des locuteurs de ladite L1, la façon d'envisager le « corpus de travail » se fait sans lien avec ce qui n'y est pas représenté, donc avec les savoirs des élèves. Par le biais du corpus, l'objet d'enseignement peut être perçu comme une langue étrangère sans ainsi la nommer (si tel était le cas, une attention toute autre serait portée aux compétences des élèves, relativement aux recommandations du CECRL). La sélection de textes n'est pas à remettre en cause puisqu'elle permet l'observation, la manipulation et l'apprentissage de faits de langue nouveaux pour la majorité des élèves et que, par ailleurs, ils illustrent la forme de langue socialement valorisée, condition à l'intégration. En revanche, on peut s'interroger sur la pertinence d'en faire un modèle de référence au point où tout autre fait serait jugé incorrect. Il semble donc nécessaire d'amorcer un questionnement objectif autour de la façon d'introduire et de traiter les corpus dans le cadre de la didactique de la LM, jusqu'à présent faussé par la non-remise en question de l'idéologie d'un certain « bon usage » absolu. Déjà en 1984, Gagne et Lazure préconisaient, à propos de la place de l'oral dans l'enseignement de la LM, l'intégration positive des formes de langue non standard dans le discours scolaire, c'est-à-dire « en évitant toute intervention de nature corrective et en ayant précisé le 'modèle' linguistique qu'il est convenable et réaliste de faire acquérir » (Gagne & Lazure, 1984 : 28).

Il semble que lorsque la langue enseignée n'est pas la LM, il y ait bien moins de crispation, davantage de souplesse. On peut ainsi affirmer que, pour ce qui est de l'utilisation de la méthodologie de travail de la linguistique de corpus, la didactique des L1 n'a pas bénéficié du même apport que la didactique des L2, avec notamment le développement des approches plus ou moins directes, mettant le corpus entre les mains des apprenants (voir cependant les travaux de Sealey sur l'anglais L1 à l'école primaire – par ex. Sealey, 2011). En revanche, il existe des initiatives visant la formation des enseignants où l'utilisation de « corpus » ressemble à ce que l'on voit pour la L2 : étudier des enregistrements pour comprendre comment fonctionne la classe, pour analyser la gestion de la parole, etc. (par ex. Le Cunff, 2005 ; voir aussi Canut et al., ce volume). Quant à l'apport des observations faites sur corpus par des linguistes, la question de la fréquence des formes est relevée dès les années 1950 avec le français fondamental (voir aussi Guiraud, 1954), et se trouve mise en avant dans les applications de la linguistique pour l'enseignement de la langue dans les années 1970 (Peytard & Genouvrier, 1970 ; Eluerd, 1979). Des utilisations indirectes de corpus en didactique de la L1 ont ainsi été développées, notamment avec des listes de mots qui continuent d'être d'utilisées (par ex. Dubois Buyse – Ters, 1995), même si les données en question ne couvrent pas les usages courants de la langue. La notion de « corpus pédagogique » (au sens de Hunston, 2002) permet de tenir compte de l'ensemble des productions formant l'input de l'élève ou de l'apprenant. Par exemple, les recherches sur le développement de la lecture en FLM, se servant du corpus MANULEX (voir

Lété, 2006 entre autres) fait de manuels scolaires, permettent d'avoir une idée des types de mots auxquels sont confrontés les enfants (dans leurs activités scolaires autour de l'écrit) ainsi que de leur fréquence d'occurrence dans l'input.

4.2. Corpus et didactique de L2

Comme nous l'avons vu, le terme corpus a différentes valeurs et significations pour différents chercheurs en linguistique, pour différents acteurs. Et même si la communauté entière arrivait à partager une seule définition, il n'en reste pas moins que les usages possibles des « collections de documents » (quels qu'ils soient) sont potentiellement très nombreux. Il en est de même en didactique des L2 où la composition du corpus peut être plus ou moins importante, représentative, variée, etc., comportant des textes ou des documents autres. À la différence de la didactique de la L1, le corpus en L2 peut s'envisager comme input, c'est-à-dire comme source de données principales permettant l'acquisition. Du point de vue de l'exploitation des corpus (productions d'apprenants ou données de natifs en langue cible), on peut faire une distinction entre, d'une part, la description en amont d'un aspect d'une langue et les choix pédagogiques qui y sont liés, et d'autre part, en aval, son exploitation plus ou moins directe par les apprenants.

D'un point de vue linguistique, l'un des rôles principaux du corpus est d'approfondir nos connaissances des langues et d'améliorer les outils qui les décrivent. Différents corpus linguistiques ont été conçus dans le but d'appuyer la construction de dictionnaires, de manuels d'usage, de grammaires, et (dans une moindre mesure au départ) pour analyser différents genres, registres et variétés, notamment en contrastant l'écrit et l'oral (ou plutôt des variétés écrites et orales). Les produits finis sont souvent destinés à un public d'apprenants. Si on cite souvent des outils pour l'anglais (cf. la *General Service List* de West, 1953), ceux pour le français se développaient en parallèle. On pense en particulier au français fondamental (Gougenheim et al., 1964) et au projet ESLO (Enquêtes sociolinguistiques sur le français parlé à Orléans – Blanc & Biggs, 1971). Le premier a donné lieu à des dictionnaires et à des référentiels fondés sur la fréquence d'occurrence ainsi qu'à des critères pédagogiques destinés aux apprenants et autres non-natifs du français, tandis que le second répondait principalement à une demande émanant d'enseignants du français en Grande-Bretagne. Les corpus monolingues ont informé en priorité des dictionnaires et autres outils de référence, mais d'autres types de corpus ont aussi joué un rôle important. Ainsi, les corpus parallèles (notamment des corpus de traductions alignées mais aussi des corpus de textes comparables) ont contribué à un certain regain d'intérêt pour les études contrastives où l'on compare non seulement les formes *possibles* dans deux langues mais aussi les formes fréquemment *attestées* (par ex. Salkie, 2000). Associé à ce courant nous pouvons aussi signaler un certain regain d'intérêt pour l'analyse de l'interlangue à travers des corpus d'apprenants (cf. Granger, 2007). Ces travaux peuvent aussi informer les outils de référence pour cibler des éléments de langue sous- ou surexploités (comparés aux productions de natifs ou à d'autres apprenants) ou qui posent problème à différents niveaux pour des apprenants d'une L1 donnée face à une L2 donnée. Si certains regrettent que l'accent soit mis fermement sur les « erreurs » des

apprenants, il existe des initiatives se servant de corpus d'apprenants pour déterminer ce que les apprenants de différentes L1 sont capables de faire à différents niveaux (par ex. *English Profile* qui puise dans les scripts et autres transcriptions d'examens du centre Cambridge ESOL ; voir http://www.englishprofile.org).

Les corpus peuvent aussi aider à la conception de manuels et autres matériels. Ainsi, *Les Orléanais ont la parole* (Biggs & Dalwood, 1976) reprend des extraits authentiques des enquêtes sociolinguistiques réalisées dans la ville d'Orléans ; c'est avant tout la qualité authentique des données qui forment le corpus qui sera mise en avant (Chambers, 2009). Cette notion du corpus comme collection de documents authentiques persiste en didactique des L2 notamment via les manuels d'apprentissage. Le *Collins COBUILD English course* (Willis & Willis, 1988) présente quant à lui non seulement un langage « authentique » mais applique aussi rigoureusement des critères de sélection basés sur le corpus pour choisir les formes (mots, structures, etc.) ainsi que les usages et les sens qui y sont associés.

Mais avec la mise en ligne gratuite d'un nombre croissant de corpus, ainsi que la plus grande disponibilité de textes sous forme électronique et des concordanciers simples à utiliser pour les traiter, les enseignants peuvent eux aussi s'en servir directement. Ceci leur permet d'analyser des textes destinés à une utilisation en classe pour repérer différents phénomènes dont la fréquence peut être un indicateur du travail à effectuer ; par ailleurs, si toutes les occurrences d'un même fait de langue peuvent être facilement repérées dans un texte, cela permet de cibler plus facilement l'attention – toujours en contexte. De la même façon, l'enseignant peut réunir de nombreux textes auxquels les apprenants seront exposés sous forme d'un corpus dit « pédagogique » (Braun, 2007). Ce type de corpus peut se révéler très utile pour les étudiants de type Lansad (langues pour spécialistes d'autres disciplines) qui ont souvent à se concentrer sur du langage disciplinaire dans des genres et des registres bien spécifiques. L'enseignant peut alors établir des listes de mots ou de groupes de mots fréquents dans la variété cible et les comparer avec un corpus de langue « générale », soit à l'œil nu, soit grâce à une étude des mots-clés, possibilité offerte par divers logiciels tels qu'AntConc (http://www.laurenceanthony.net/antconc_index.html) et Wordsmith Tools (http://www.lexically.net/wordsmith/). Les résultats peuvent, encore une fois, informer (mais pas contraindre) les décisions concernant les éléments à traiter, les formes, les sens et usages, l'ordre de présentation des items, etc. Et l'enseignant peut aussi créer un corpus à partir des productions de ses apprenants afin de pointer des phénomènes récurrents ou spécifiques à certains d'entre eux, ou alors pour mieux cerner leur développement langagier. Ce processus est relativement simple lorsque les étudiants produisent des textes sous forme électronique (devoirs, examens, forums…), mais on peut aussi impliquer les apprenants dans la création d'un corpus à partir de leurs propres manuscrits (Seidlhofer, 2002) ou en enregistrant et transcrivant leurs propres productions orales (Lynch, 2001 ; Mennim, 2012). Le recours au corpus (ainsi qu'à la méthodologie de travail sur des données) semble ici aller de soi, ce qui est peut-être moins évident lorsqu'on travaille à partir de corpus existants (par ex. Braun, 2007).

Les apprenants peuvent aussi se servir de corpus à des fins d'apprentissage de façon plus ou moins directe (Johns & King, 1991). Dans une démarche encadrée ou contrôlée par l'enseignant (Kerr, 2013), celui-ci détermine les objectifs et choisit le corpus approprié, fait ses propres requêtes afin de repérer les informations pertinentes qu'il sélectionne selon les besoins et profils de ses apprenants, et les présente sur papier accompagnées d'instructions précises. Ainsi les données, les questions, le déroulement et même les résultats sont établis au préalable par l'enseignant qui reste maître de tout, ce qui peut être un avantage dans certains cas. Il peut aussi se servir d'un corpus comme simple source d'exemples ou comme « informateur » lorsqu'une question difficile se présente, ou en amont pour créer des tests ou des phrases/textes à trous, des exercices ou activités où les apprenants doivent remettre ensemble des phrases coupées, comparer deux phénomènes, repérer des tendances ou régularités, etc. – tout exercice ou activité qui peut être rendu plus aisé par un accès facile à des quantités de données. La spécificité de cette approche réside dans le travail sur un langage authentique et pertinent, généralement à travers plusieurs lignes de concordance qui mettent en relief le phénomène cible en contexte et y attire l'attention des apprenants. L'approche peut être déductive (par exemple, suite à la présentation d'une « règle » afin de la tester), mais elle est surtout inductive lorsque les apprenants ont eux-mêmes à repérer des régularités, à trier les données, à formuler des hypothèses quant aux sens ou aux usages, à les tester, et ainsi de suite (voir O'Sullivan, 2007 pour une liste d'avantages de ce genre de travail sur corpus). Cette approche s'accorde bien avec certaines théories récentes de l'acquisition (cf. Tomasello, 2000 ; Hoey, 2005) selon lesquelles lors de chaque rencontre avec un fait linguistique nous raffinons nos connaissances de ce dernier (collocations, connotations, prosodie et préférence sémantiques, ainsi que nos connaissances des discours, genres, registres, etc.) ; et une exposition systématisée grâce à l'utilisation de données attestées via des concordances peut accélérer ce processus (cf. Gaskell & Cobb, 2004). Taylor (2012), quant à lui, conçoit nos connaissances langagières sous forme d'un « corpus mental » : on n'apprendrait ainsi pas de « règles » pour ensuite les appliquer, mais on accumulerait un stock important de fragments langagiers d'où émergent des tendances, des régularités. On revient alors à une conceptualisation « exemplariste » de l'appropriation, basée sur l'usage, probabilistique et en lien avec les modèles complexes ou des systèmes dynamiques où l'acquisition du langage se fait essentiellement à partir d'exemples (par ex. Ellis & Larsen-Freeman, 2006).

Dans certains cas, l'enseignant peut donner aux apprenants une part de responsabilité lorsqu'ils ont accès à un corpus et un concordancier (ou autre collection de textes et autre type d'outil d'analyse). Dans la version la plus autonomisante, ce serait alors à chaque apprenant de choisir (ou de construire) son corpus ainsi que les outils d'analyse ; ce serait aussi à lui de formuler les requêtes en fonction de ses propres interrogations, d'en interpréter les résultats, etc. L'apprenant a la possibilité de poursuivre ce travail en dehors de la salle de classe, voire même plus tard dans sa vie à chaque fois qu'il en aura besoin – c'est donc avant tout une méthodologie de l'observation des faits linguistiques qui est censée être productive, pouvant servir à tout moment. On peut faire la comparaison avec le simple fait d'utiliser un moteur de recherche sur Internet, par exemple (cf. Boulton, 2015) ; en général, on finit bien par

trouver ce qu'on cherche (et d'autres choses encore, parfois de façon tout à fait sérendipiteuse) parce qu'on a cette habitude de chercher, de s'informer, de trouver des réponses aux questions que l'on se posait (concernant le prix d'un billet d'avion, ou l'année de naissance d'un écrivain, le nombre d'habitants d'une ville ou d'un pays…). Bien entendu, cette approche sur corpus n'est pas sans poser de problèmes, tant pour le choix des corpus et des outils que pour l'interprétation des données. Pour cette raison, le rôle le plus ouvert du chercheur ou « détective » linguiste (Johns, 1997) doit être réservé aux apprenants de niveau avancé avec des besoins très précis. Mais même auprès d'apprenants moins avancés, l'enseignant peut laisser davantage de responsabilité en mettant les apprenants en situation de travail direct sur un corpus (Boulton, 2010). Pour commencer, on peut se servir des outils plus familiers, par exemple en aidant les apprenants à mieux se servir du web comme corpus et d'un moteur de recherche comme concordancier (Boulton & Tyne, 2014). Ces premiers pas sont intéressants car ils s'appuient sur des pratiques existantes, et ils peuvent ainsi servir de point d'entrée pour un travail plus autonomisant sur corpus.

5. Conclusion

Si les définitions de ce qu'est un corpus varient considérablement d'un domaine à l'autre, ce que l'on retient ici est que la notion d'une collection de données (quelle qu'elle soit) semble être de mise dans tous les cas, qu'il s'agisse de décrire une L2 ou une L1 pour les besoins de l'apprentissage, de comprendre les productions d'enfants ou d'apprenants en cours d'acquisition, d'induire des règles d'usage, etc. De plus, si les besoins et les enjeux ne sont pas les mêmes d'un domaine à l'autre (par ex. pour la didactique : corpus comme input en L2 ; corpus comme lieu d'observation des interactions ou de la variation en L1), on retrouve certains questionnements fondamentaux : quelle représentativité des données ? Quel rapport à la norme ou aux formes prescriptives ?

Les différents articles qui forment la suite de ce numéro reprennent ces questionnements (et d'autres encore) en les développant dans le cadre de recherches ou de réflexions portant sur différents aspects de l'appropriation de L1 ou de L2. Dans tous les cas il apparaît que l'apport des corpus (et de la méthodologie sur corpus) est désormais incontournable. Et on voit difficilement comment chacun des différents domaines concernés pourraient s'en passer tant leur développement est lié à l'apport de techniques et de technologies de l'étude sur corpus.

Bibliographie

AARTS, B. (2001), « Corpus linguistics, Chomsky and fuzzy tree fragments », dans C. Mair & M. Hundt (éds.), *Corpus Linguistics and Linguistic Theory*, Amsterdam, Rodopi, p. 5-13.

ADJEMIAN, C. (1976), « On the nature of interlanguage systems », *Language Learning*, 26/2, p. 297-320.

ANDRÉ, V. & CANUT, E. (2010), « Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement des Corpus Oraux en Français) », *Pratiques*, 147-148, p. 35-51.

BARONI, M. & BERNARDINI, S. (éds.) (2006), *Wacky! Working Papers on the Web as Corpus*, Bologna, Gedit.

BAUTIER, E. (2001), « Pratiques langagières et scolarisation », *Revue française de pédagogie*, 137, p. 117-161.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S. & FINEGAN, E. (1999), *Longman Grammar of Spoken and Written English*, Londres, Pearson.

BIGGS, P. & DALWOOD, M. (1976), *Les Orléanais ont la parole*, Londres, Longman

BILGER, M. (éd.) (2000), *Linguistique sur corpus : études et réflexions*, Perpignan, Presses Universitaires de Perpignan.

BLANC, M. & BIGGS, P. (1971), « L'enquête sociolinguistique sur le français parlé à Orléans », *Le français dans le monde*, 85, p. 16-25.

BLANCHE-BENVENISTE, C. & JEANJEAN, C. (1987), *Le français parlé. Transcription et édition*, Institut National de la Langue Française (INALF-CNRS), Paris, Didier.

BLOOM, L. (1970), *Language Development: Form and Function in Emerging Grammars*, Cambridge MA, MIT Press.

BLOOM, L. (1973), *One Word at a Time: The Use of Single Word Utterances before Syntax*, La Haye, Mouton.

BRAINE, M. (1963), « The ontogeny of English phrase structure: The first phrase », *Language*, 39, p. 1-14.

BRAUN, S. (2007), « Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora », *ReCALL* 19/3, p. 307-328.

BROWN, R. (1973), *A First Language: The Early Stages*, Cambridge MA, Harvard University Press.

BOULTON, A. (2009), « Documents authentiques, oral, corpus », *Mélanges CRAPEL*, 31, p. 5-13.

BOULTON, A. (2010), « Data-driven learning: Taking the computer out of the equation », *Language Learning*, 60/3, p. 534-572.

BOULTON, A. (2015), « Applying data-driven learning to the web », dans A. Leńko-Szymańska & A. Boulton (éds.), *Multiple Affordances of Language Corpora for Data-driven Learning*, Amsterdam, John Benjamins, p. 267-295.

BOULTON, A. & TYNE, H. (2014), *Des documents authentiques aux corpus : démarches pour l'apprentissage des langues*, Paris, Didier.

BOUTET, J. (2002), « 'I parlent pas comme nous' : pratiques langagières des élèves et pratiques scolaires », *VEI Enjeux*, 130, p. 163-177.

CADET, L. & GUERIN, E. (éds.) (2012), « FLM, FLS, FLE... au-delà des catégories », *Le français aujourd'hui*, 176.

CANUT, E. & VERTALIER, M. (2008), « Des données *représentatives*... de quoi en acquisition du langage ? Constitution de données à observer et objectifs d'analyse », *Verbum*, 30/4, p. 299-312.

CAPPEAU, P. & GADET, F. (2007a), « L'exploitation sociolinguistique des grands corpus : maître-mot et pierre philosophale », *Revue française de linguistique appliquée*, 12/1, p. 99-110.

CAPPEAU, P. & GADET F. (2007b), « Où en sont les corpus sur les français parlés ? », *Revue française de linguistique appliquée*, 12/1, p. 129-133.

CAPPEAU, P. & GADET F. (2010), « Transcrire, ponctuer, découper l'oral : bien plus que de simples choix techniques », *Cahiers de linguistique*, 35/1, p. 187-202.

CHAMBERS, A. (2009), « Les corpus oraux en français langue étrangère : authenticité et pédagogie », *Mélanges CRAPEL*, 31, p. 15-33.

CHISS, J.-L. & DAVID, J. (2012), *Didactique du français et étude de la langue*, Paris, Armand Colin.

CHOMSKY, N. (1979), *Language and Responsibility (based on conversations with Mitsou Ronat)*, New York, Pantheon.

CLARK, E. (2009), « What shapes children's language? Child-directed speech, conventionality, and the process of acquisition », dans V. Mueller Gathercole (éd.), *Routes to Language: Studies in Honor of Melissa Bowerman*, Mahwah NJ, Lawrence Erlbaum, p. 233-254.

COMBETTES, B. (2014), « Diachronic linguistics and electronic corpora », dans H. Tyne, V. André, C. Benzitoun, A. Boulton & Y. Greub (éds.), *French through Corpora: Ecological and Data-driven Perspectives in French Language Studies*, Newcastle, Cambridge Scholars, p. 2-11.

Conseil de l'Europe (2000), *Un cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*, http://www.coe.int/t/dg4/linguistic/Source/Framework_FR.pdf.

CORDER, S. P. (1971), « Idiosyncratic dialects and error analysis », *IRAL*, 9/2, p. 147-60.

CORDER, S. P. (1973), « The elicitation of interlanguage », *Errata: Papers in Error Analysis*, Lund, CWK Gleerup.

DELEFOSSE, J. (2010), *Sur le langage de l'enfant. Choix de textes de 1876 à 1962*, Paris, L'Harmattan.

DUBOIS, M., KAMBER, A. & SKUPIEN DEKENS, C. (2014), « A quantitative and qualitative analysis of French L2 students' spelling problems: The case of adjective agreement », dans H. Tyne, V. André, C. Benzitoun, A. Boulton & Y. Greub (éds.), *French through Corpora: Ecological and Data-driven Perspectives in French Language Studies*, Newcastle, Cambridge Scholars, p. 312-334.

DUTEIL-MOUGEL, C. (2007), « Groupements de textes et corpus : point de vue de linguiste », dans F. Rastier & M. Ballabriga (éds.), *Corpus en Lettres et Sciences sociales : des documents*

numériques à l'interprétation, Toulouse, Presses de l'Université de Toulouse-Le-Mirail, p. 225-235.

ELLIS, N. & LARSEN-FREEMAN, D. (éds.) (2006), « Language emergence : implications for Applied Linguistics », *Applied Linguistics*, 27/4.

ELUERD, R. (1979), *L'usage de la linguistique en classe de français : critiques et perspectives* (Vol. 2), Paris, Editions ESF.

FORCHINI, P. (2012), *Movie Language Revisited: Evidence from Multi-Dimensional Analysis and Corpora*, Francfort, Peter Lang.

GADET, F., LUDWIG, R., MONDADA, L., PFÄNDER, S. & SIMON, A.-C. (2012), « Un grand corpus de français parlé : choix épistémologiques et réalisations empiriques », *Revue française de linguistique appliquée*, 17/1, p. 39-54.

GAGNE, G. & LAZURE, R. (1984), « Deux décennies de recherches américaines en pédagogie de la langue maternelle », *Revue française de pédagogie*, 66, p. 69-98.

GASKELL, D. & COBB, T. (2004), « Can learners use concordance feedback for writing errors? », *System*, 32, p. 301-319.

GILQUIN, G. & GRIES, S. (2009), « Corpora and experimental methods: A state-of-the-art review », *Corpus Linguistics and Linguistic Theory*, 5/1, p. 1-26.

GOLDBERG, A. (2006), *Constructions at Work: The Nature of Generalization in Language*, Oxford, Oxford University Press.

GOUGENHEIM, G., RIVENC, P. & HASSAN, M. (1964), « Le français fondamental », *Tendances nouvelles en matière de recherche linguistique. L'éducation en Europe*, 5/2, p. 53-72.

GRANGER, S. (2007), « The computer learner corpus: A versatile new source of data for SLA research », dans W. Teubert & R. Krishnamurthy (éds.), *Corpus Linguistics: Critical Concepts in Linguistics*, Londres, Routledge, p. 166-182.

GUERIN, E. (2008), « Le 'français standard' : une variété située ? », dans J. Durand, B. Habert & B. Laks (éds.), *Actes du Congrès mondial de linguistique française - CMLF'08*, http://www.linguistiquefrancaise.org.

GUIRAUD, P. (1954), *Les caractères statistiques du vocabulaire : essai de méthodologie*, Paris, Presses Universitaires de France.

HALLIDAY, M. (2007), *Language and Education* (Vol. 9), Londres, Continuum.

HERLITZ, W., ONGSTAD, S. & VAN DE VEN, P.-H. (2007), *Research on Mother Tongue Education in a Comparative International Perspective: Theoretical and Methodological Issues*, Utrecht, Rodopi.

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

HOLEC, H. (1990), « Des documents authentiques, pour quoi faire ? », *Mélanges pédagogiques* 1990, p. 65-74.

HUNSTRON, S. (2002), *Corpora in Applied Linguistics*, Cambridge, Cambridge University Press.

JOHNS, T. (1997), « Contexts: The background, development and trialling of a concordance-based CALL program », dans A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (éds.), *Teaching and Language Corpora*, Harlow, Addison Wesley Longman, p. 100-115.

JOHNS, T. & KING, P. (éds.) (1991), « Classroom Concordancing », *English Language Research Journal*, 4.

KERR, B. (2013), « Grammatical description and classroom application. Theory and practice in data-driven learning », *Bulletin suisse de linguistique appliquée*, 97, p. 17-39.

KOESTER, A. (2010), « Building small specialised corpora », dans A. O'Keeffe & M. McCarthy (éds.), *Routledge Handbook of Corpus Linguistics*, Londres, Routledge, p. 66-79.

KUČERA, H. & FRANCIS, W. (1967), *Computational Analysis of Present-Day American English*, Providence RI, Brown University Press.

LABOV, W. (1971), « Some principles of linguistic methodology », *Language in Society*, 1, p. 97-120.

LANDURE, C. (2014), *Étude comparative de l'exploitation directe de corpus générique et spécifique par des apprenants LANSAD*, thèse de doctorat, Paris, Université Paris 7 Diderot.

LE CUNFF, C. (2005), « De l'usage des corpus en didactique de l'oral : recherche et formation », dans G. Williams (éd.), *La linguistique de corpus*, Rennes, Presses Universitaires de Rennes, p. 397-406.

LÉTÉ, B. (2006), « L'apprentissage implicite des régularités statistiques de la langue et l'acquisition des unités morphologiques. L'exemple des homophones-hétérographes », *Langue française*, 151/3, p. 41-58.

LYNCH, T. (2001), « Seeing what they meant: Transcribing as a route to noticing », *ELT Journal*, 55/2, p. 124-132.

MAYER, M. (1969), *Frog, Where Are You?*, New York, Penguin Putnam.

MCENERY, T., XIAO, R. & TONO, Y. (2006), *Corpus-based Language Studies: An Advanced Resource Book*, Londres, Routledge.

MENNIM, P. (2012), « Learner negotiation of L2 form in transcription exercises », *ELT Journal*, 66/1, p. 52-61.

MILROY, L. (1987), *Observing and Analysing Natural Language*, Oxford, Blackwell.

MORGENSTERN, A. & PARISSE, C. (2007), « Codage et interprétation du langage spontané d'enfants de 1 à 3 ans », *Corpus*, 6, p. 55-78.

MORGENSTERN, A. & PARISSE, C. (2012), « The Paris corpus », *Journal of French Language Studies*, 22/1, p. 7-12.

MORGENSTERN, A. & PARISSE, C. (2013), « Premières formes de conditionnel chez l'enfant », *Faits de langue*, 40, p. 219-223.

NEMSER, W. (1971), « Approximative systems of foreign language learners », *IRAL*, 9/2, p. 115-123.

O'SULLIVAN, Í. (2007), « Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy », *ReCALL*, 19/3, p. 269-286.

PEYTARD, J. & GENOUVRIER, E. (1970), *Linguistique et enseignement du français*, Paris, Larousse.

ROY, D., PATEL, R., DECAMP, P., KUBAT, R., FLEISCHMAN, M., ROY, B., MAVRIDIS, N., et al. (2006), « The human speechome project », dans P. Vogt, Y. Sugita, E. Tuci &

C. Nehaniv (éds.), *Symbol Grounding and Beyond. Lecture Notes in Computer Science*, Berlin & Heidelberg, Springer, p. 192-196.

SALKIE, R. (2000), « Quelques questions méthodologiques dans l'exploitation des corpus multilingues », dans M. Bilger (éd.), *Corpus : méthodologie et applications linguistiques*, Paris, Champion & Perpignan, Presses Universitaires de Perpignan, p. 180-195.

SEALEY, A. (2011), « The use of corpus-based approaches in building children's knowledge about language », dans S. Ellis & E. McCartney (éds.), *Applied Linguistics and Primary School Teaching*, Cambridge, Cambridge University Press, p. 93-106.

SELINKER, L. (1972), « Interlanguage », *IRAL*, 10/2, p. 209-231.

SEIDLHOFER, B. (2002), « Pedagogy and local learner corpora: Working with learner-driven data », dans S. Granger, J. Hung & S. Petch-Tyson (éds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Amsterdam, John Benjamins, p. 213-234.

TAYLOR, J. (2012), *The Mental Corpus: How Language is Represented in the Mind*, Oxford, Oxford University Press.

TERS, F. (1995), *Les 100 mots fondamentaux de l'école élémentaire. Echelle Dubois Buyse. Vocabulaire actif*, Paris, MDI.

TOGNINI-BONELLI, E. (2010), « Theoretical overview of the evolution of corpus linguistics », dans A. O'Keeffe & M. McCarthy (éds.), *Routledge Handbook of Corpus Linguistics*, Londres, Routledge, p. 14-27.

TOMASELLO, M. (2000), « First steps toward a usage-based theory of language acquisition », *Cognitive Linguistics*, 11/1-2. p. 61-82.

TOMASELLO, M. & STAHL, D. (2004), « Sampling children's spontaneous speech: How much is enough? », *Journal of Child Language*, 31/1, p. 101-121.

TYNE, H. (2009), « Corpus oraux par et pour l'apprenant », *Mélanges CRAPEL*, 31, p. 91-111.

TYNE, H., ANDRÉ, A., BENZITOUN, C., BOULTON, A. & GREUB, Y. (éds.) (2014), *French through Corpora: Ecological and Data-driven Perspectives in French Language Studies*, Newcastle, Cambridge Scholars.

VIALA, A. (2014), « Corpus, savoirs et choix de textes : à quand les fondamentaux réels ? », *Le français aujourd'hui*, 185, p. 89-94.

WELLS, G. (1985), *Language Development in the Pre-school Years* (*Language at Home and at School, Vol. 2*.), Cambridge, Cambridge University Press.

WEST, M. (1953), *A General Service List of English Words*, Londres, Longman.

WILLIS, J. & WILLIS, D. (1988), *Collins COBUILD English Course*, Londres, Collins.