Petits corpus oraux bilingues et plurilingues – enjeux théoriques et méthodologiques

Small bi- and multilingual spoken corpora – a theoretical and methodological challenge

Inga Hennecke

Édition électronique

URL : <http://journals.openedition.org/corpus/3451>

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Référence électronique

Inga Hennecke, « Petits corpus oraux bilingues et plurilingues – enjeux théoriques et méthodologiques », *Corpus* [En ligne], 18 | 2018, mis en ligne le 09 juillet 2018, consulté le 08 septembre 2020. URL : <http://journals.openedition.org/corpus/3451>

Ce document a été généré automatiquement le 8 septembre 2020.

© Tous droits réservés

Petits corpus oraux bilingues et plurilingues – enjeux théoriques et méthodologiques

Small bi- and multilingual spoken corpora – a theoretical and methodological challenge

Inga Hennecke

Introduction

- 1 Dans les dernières décennies, la linguistique de corpus a beaucoup gagné en importance. Cela s'explique bien entendu par les progrès techniques réalisés dans le domaine informatique, qui nous offre de meilleures possibilités en ce qui concerne la collecte, le traitement et l'analyse de données. Néanmoins, la linguistique de corpus est toujours confrontée à une terminologie et une méthodologie hétérogènes. Selon Pusch (2014), un corpus « peut être défini de façon très générale, comme collection de données langagières structurée selon certains critères » (Pusch 2014 : 174). D'ailleurs, il n'existe toujours pas d'accord concernant la définition et la démarcation d'un corpus d'autres types de collections de données, par exemple des collections de textes (journaux, collections de lettres, enregistrements du langage parlé...), des livres ou même des pages web. À cet égard, il ne ressort pas toujours clairement quelles caractéristiques un texte ou une collection de données linguistiques doivent posséder afin d'obtenir la dénomination « corpus ». Ce problème se pose particulièrement par rapport aux petits corpus, car il n'est pas clairement défini à partir de quelle longueur de texte ou de quel nombre de textes, nous pouvons parler de « corpus ».
- 2 En plus de la définition imprécise du terme « corpus », nous faisons aussi face à une polysémie du terme « linguistique de corpus ». D'une part, la linguistique de corpus désigne le processus de création de corpus, plus précisément des processus de collecte de données, de transcription de données ainsi que d'annotation et de standardisation de ces dernières. C'est surtout ce dernier point qui peut être considéré comme un

champ de recherche spécifique dans le domaine du traitement automatique des langues (TAL). D'autre part, le terme « linguistique de corpus » peut se référer à l'exploitation de corpus en tant qu'outil empirique à des fins de recherche. Dans ce contexte, la linguistique de corpus peut être combinée avec diverses disciplines théoriques de la linguistique afin d'analyser des hypothèses de travail ou des questions de recherche à l'aide de données linguistiques naturelles.

- 3 Cela soulève une question importante, que Pusch (2014 : 174) formule ainsi : « [...] la linguistique de corpus : est-ce une méthode, un nouveau paradigme d'analyse ou une sous-discipline de la linguistique ? ». De ce qui précède, il ressort qu'il n'existe pas de réponse claire à cette question. Leech trouve une définition assez satisfaisante et décrit la linguistique de corpus comme « *a methodologically-oriented branch of linguistics* » (Leech 2011 : 158).
- 4 Toute cette discussion montre que la linguistique de corpus occupe aujourd'hui une large place dans la recherche linguistique et que la linguistique de corpus en soi est un champ de recherche très vaste. Cette contribution tentera de décrire les enjeux théoriques ainsi que méthodologiques d'un type de corpus très précis, à savoir les petits corpus oraux bilingues et plurilingues.
- 5 Dans une première partie de cette contribution sera reprise la discussion au sujet des petits et grands corpus pour ensuite évoquer le cas des corpus bi- et plurilingues. Il sera d'abord question d'illustrer les enjeux des petits corpus bilingues oraux ainsi que la comparabilité et la généralisation de leurs résultats d'analyse pour ensuite présenter l'exemple concret d'un petit corpus bilingue – le Corpus Franco-Manitobain (Corpus FM). Ce corpus contient des données linguistiques françaises et anglaises d'une variété du français canadien parlée dans la province du Manitoba et qui a servi de base de données pour une recherche sur le changement linguistique induit par le contact linguistique.

1. Petits corpus issus du langage oral

- 6 Dans la discussion scientifique sur la linguistique de corpus, les chercheurs font généralement la différence entre les corpus de petite taille et ceux de grande taille. Dans les dernières décennies, le but de la linguistique de corpus a été de compiler des corpus de plus en plus grands afin de pouvoir effectuer des recherches et des analyses quantitatives et qualitatives sur un échantillon le plus représentatif possible. Cette évolution va de pair avec les progrès dans les domaines de l'informatique et de la linguistique quantitative. C'est alors seulement ces dernières années que les corpus de petite taille ont été de nouveau pris en considération par la communauté scientifique. Dans la présente section, je reprendrai la discussion au sujet des grands et des petits corpus en soulignant les avantages et les inconvénients de ces deux types de bases de données.
- 7 Koester (2010) cite dans son article sur les petits corpus Sinclair (2004), qui postule que « *small is not beautiful ; it is simply a limitation* » (Sinclair 2004 : 189 dans Koester 2010 : 66). Dans cette ligne d'argumentation, les petits corpus sont moins utiles et moins efficaces que les grands corpus parce qu'ils ne donnent que des résultats limités et non généralisables. Quelques années plus tôt, Sinclair exprime un point de vue différent :

'small' and 'large' corpora are seen in contrast with each other. So the difference must be methodological, because it cannot be just size, whether relative or absolute size. (Sinclair 2001 : xi)

- 8 Dans cette citation, Sinclair évoque une différence méthodologique entre les petits et les grands corpus, qu'il décrit avec la terminologie de « *early human intervention (EHI)* » pour les petits corpus et de « *delayed human intervention (DHI)* » (*ibid.*) pour les grands corpus. Dans le cas d'une méthodologie EHI, les chercheurs constituent un petit corpus avec un but précis et clair. Ce petit corpus est ensuite consultable manuellement et sans recours aux méthodes informatiques (*ibid.*). Au contraire, les grands corpus demandent une méthodologie DHI, c'est-à-dire qu'ils ne peuvent pas être analysés manuellement et que les chercheurs doivent s'appuyer sur des méthodes et processus automatiques pour obtenir des résultats interprétables. Dans ce même sens méthodologique, Koester (2010) compare les petits et les grands corpus en fonction de leurs avantages et de leurs inconvénients (Koester 2010 : 66-67). Les grands corpus offrent un vaste réseau de sources linguistiques et peuvent indiquer des informations sur des occurrences d'unités lexicales peu fréquentes. Cela implique que les résultats d'une recherche à l'aide de grands corpus sont plus généralisables que les résultats de recherche issus de petits corpus. En outre, dans les petits corpus, il existe le risque de distorsion des résultats à cause de l'usage idiosyncratique dominant d'un locuteur spécifique (Cameron & Deignan 2003 : 151). En revanche, les grands corpus peuvent contenir des informations non vérifiées et donc fausses, qui ne reflètent pas l'usage réel d'un phénomène linguistique. Dans ce cas, les petits corpus sont une source pertinente et fiable (Sinclair 2001 : xi), qui permet une investigation plus précise des contextes d'occurrence. Par conséquent, les petits corpus offrent la possibilité d'analyser toutes les occurrences et pas seulement les occurrences de haute fréquence (Koester 2010 : 66).
- 9 En général, les petits corpus ne visent pas à être des corpus de référence. Ils contiennent plutôt un discours spécialisé sur une thématique précise. De Beaugrande (2001) distingue deux types de petits corpus. D'un côté, il fait référence aux corpus spécialisés, qui sont limités par un registre, un domaine de discours ou un sujet spécifique. De l'autre, il évoque les corpus d'apprentissage, qui sont limités par le niveau de compétence linguistique des apprenants (de Beaugrande 2001 : 3).

2. Petits corpus oraux bilingues et plurilingues

- 10 Les corpus bilingues et plurilingues sont d'une importance majeure pour une multitude de disciplines linguistiques. Les chercheurs s'en servent dans la recherche sur le contact et le changement linguistique, l'acquisition d'une langue seconde ou étrangère, la variation linguistique, l'alternance codique et langues mixtes, la linguistique de migration ainsi que le traitement psycholinguistique de plusieurs langues. Pourtant, les définitions de corpus bilingues et plurilingues sont nombreuses et la terminologie reste souvent hétérogène. Cela produit forcément des effets inévitables sur les chercheurs en train de compiler ou de trouver un corpus bi- ou plurilingues apte à satisfaire leurs besoins. C'est pour cela que cette partie est consacrée à une présentation et une définition de corpus bi- et plurilingues pour ensuite présenter le type de corpus sous-jacent à la présente contribution.
- 11 Backus (2008) donne un aperçu de la différence entre corpus monolingues et bilingues :

Tableau 1. Différences entre corpus monolingues et bilingues

Corpus monolingues	Corpus bilingues
grands	petits
facilement accessibles	difficilement accessibles
annotés	en général non annotés
apte à usage général	apte à usage spécifique

(Backus 2008 : 237)

- 12 Ce tableau montre que selon Backus (2008) la taille est une caractéristique inévitable d'un corpus bilingue ou plurilingue. En plus, il décrit des corpus bi- et plurilingues que les chercheurs emploient dans la recherche sur le bilinguisme ainsi :

They tend to be small, hard for other researchers to get access to, and without much annotation, all characteristics that are generally held to be the opposite of what a good corpus has to offer. (Backus 2008 : 233)

- 13 Dans cette citation, Backus évoque déjà plusieurs problèmes de corpus bi- et plurilingues, qui seront discutés plus tard dans cette partie. Dans ce qui suit, il sera d'abord question de donner un aperçu terminologique sur les définitions de bi- et plurilinguisme. Ensuite, il est important de différencier précisément entre différents types de corpus bi- et plurilingues et de comparer leurs caractéristiques. Dans une partie suivante, il s'agit d'appliquer les enjeux théoriques (2.1) aux enjeux méthodologiques, c'est-à-dire aux problèmes de compilation et de création de corpus bi- et plurilingues (2.2).

2.1. Enjeux théoriques

- 14 Les enjeux théoriques dans la recherche sur les corpus bi- et plurilingues commencent déjà avec la définition vague de ce que sont le bi- et le plurilinguisme. En général, le bilinguisme implique deux langues tandis que le plurilinguisme fait référence à plus de deux langues (Backus 2008 : 237). Quand un corpus contient des données linguistiques en deux langues, il s'agit donc d'un corpus bilingue, dans le cas de plus de deux langues, nous parlons d'un corpus plurilingue. Une approche alternative serait de parler de corpus plurilingue quand le corpus contient plus d'une langue et de traiter un corpus comme monolingue quand il ne contient qu'une langue. Pourtant, les termes bi- et plurilingues font face à d'autres incertitudes terminologiques. Selon Oksaar, une personne bilingue peut sans problème changer de langue si la situation l'exige (Oksaar 1984 : 852). Cela implique qu'une personne bilingue, ou plurilingue, dispose de compétences similaires ou du moins comparables dans les deux (ou dans les différentes) langues. Bien évidemment, cela n'est pas toujours le cas quand nous faisons référence au bi- au plurilinguisme. Par conséquent, Lüdi (1996) différencie plusieurs facteurs qui influencent le plurilinguisme individuel. L'aspect le plus important est sans doute le moment de l'acquisition des langues. Lüdi distingue entre l'acquisition simultanée et l'acquisition successive de plusieurs langues. Dans le premier cas, il s'agit généralement d'un apprentissage de deux langues maternelles (2L1), qu'un enfant

apprend dès la naissance. Dans le second cas, il est question d'une première langue maternelle (L1), qui est apprise dès la naissance, et d'une seconde langue maternelle (L2), que l'enfant apprend plus tard (Lüdi 1996 : 335). À ce sujet, le contexte d'apprentissage est d'une importance majeure. L'apprentissage L1-L2 ou 2L1 implique un contexte naturel, c'est-à-dire que les langues sont apprises à la maison ou dans l'environnement social des enfants (apprentissage incontrôlé). Si une ou plusieurs langues sont enseignées dans un contexte scolaire et sans contact naturel avec la langue, nous évoquerons plutôt la terminologie de langue étrangère (apprentissage contrôlé) (*ibid.*). Hormis ces facteurs, qui influencent le bi- et plurilinguisme d'un individu, Lüdi souligne le rôle du degré de compétence linguistique (symétrique ou asymétrique) et de maîtrise de la langue écrite (*ibid.*). Tout cela suggère que la notion de bi- et plurilinguisme inclut une large variété de facteurs individuels et sociaux qui influencent la compétence individuelle de chaque locuteur plurilingue. À cela s'ajoute un autre aspect non moins important pour la compilation de corpus bi- et plurilingues, celui du choix individuel de la langue. Comme indiqué précédemment, chaque locuteur plurilingue possède la compétence de choisir une langue spécifique en fonction de la situation (Oksaar 1984). Ce choix de la langue dépend de plusieurs critères qui peuvent être subdivisés entre ceux qui dépendent du domaine linguistique et ceux qui dépendent de la situation discursive spécifique. Les critères du domaine linguistique impliquent l'usage spécifique d'une langue dans le contexte religieux, familial ou éducationnel. Les critères de situation comprennent la compétence linguistique de l'interlocuteur, ses idées normatives ainsi que le degré de formalité de la situation, sans oublier des facteurs sociolinguistiques comme l'âge, le sexe et le niveau de formation. En ce qui concerne les critères de situation, Grosjean (2001) emploie le terme de *language modes*. Selon lui, « the state of activation of the bilingual's languages and language processing mechanisms has been called the language mode » (Grosjean 2001 : 3). Dans un mode monolingue, l'individu bilingue essaye d'opprimer les autres langues autant que possible afin de communiquer avec un interlocuteur monolingue. Dans un mode bilingue, la personne bilingue communique avec d'autres personnes bilingues et active alors les différentes langues (Grosjean 2001). Dans une communication bi- ou plurilingue, les interlocuteurs choisissent d'abord une langue de base et insèrent ensuite dans la conversation la ou les autres langue(s) à l'aide de mécanismes spécifiques, comme par exemple l'alternance codique.

- 15 La discussion terminologique évoquée constitue une base importante pour la discussion sur les différents types de corpus bi- et plurilingues. Comme déjà indiqué, il existe une certaine ambiguïté par rapport à la définition de ce qu'est un corpus bi- ou plurilingue. McEnery et Xiao (2008) distinguent entre trois types de corpus qui incluent plus d'une langue :

1. Type A : Textes de source avec traductions
2. Type B : Sous-corpus monolingues compilés selon le même cadre d'échantillon
3. Type C : Une combinaison de A et B
(McEnery & Xiao 2008 : 19)

- 16 Les auteurs soulignent l'insécurité terminologique en indiquant que Aijmer et Altenberg (1996) considèrent le type A comme corpus de traduction et le type B comme corpus parallèle, tandis que McEnery et Wilson (1996 : 57) envisagent le type A comme corpus parallèle et le type B comme corpus comparable (McEnery et Xiao 2008 : 20). McEnery et Xiao optent pour cette dernière définition, qui semble être la définition

courante dans la littérature récente. Hansen-Schirra *et al.* (2017) définissent un corpus parallèle comme un recueil de textes en version originale avec leurs traductions, qui peuvent s'avérer utiles dans les domaines de la traductologie, de la traduction automatique, de la TAL ainsi que pour le traducteur humain (Hansen-Schirra *et al.* 2017 : 1). Les domaines du TAL notamment se servent des corpus parallèles comme de base de données pour l'extraction automatique de terminologie, de désambiguïsation, de sens des mots ainsi que pour la lexicographie automatique et l'induction de grammaires multilingues (*ibid.*). Selon Sinclair (2001), les corpus parallèles sont généralement alignés de façon coordonnée pour que les chercheurs puissent y retrouver les équivalences de traduction (Sinclair 2001 : xii). Un exemple très important d'un corpus parallèle constitue la base de données EUR-LEX, qui contient des textes « du droit européen » dans les langues officielles de l'Union européenne¹ :

Tableau 2. Le corpus parallèle EUR-LEX²

Diese Schwierigkeiten werden mit der Zahl der an der Bewirtschaftung von Gemeinschaftsgeldern beteiligten Mitgliedstaaten, Unternehmen und Verwaltungen erheblich zunehmen.	Ces difficultés vont s'accroître avec l'augmentation du nombre d'États membres et du nombre d'opérateurs et administrations impliqués dans la gestion des fonds communautaires.
---	---

- 17 Au contraire, un corpus comparable est un corpus qui contient des données linguistiques comparables, c'est-à-dire des données collectées à l'aide d'un même type d'échantillon et d'une représentativité similaire. Cela peut concerner entre autres le genre, la proportion des textes ainsi que la période de publication (en question) (McEnery et Xiao 2008 : 20). Selon McEnery et Xiao, les corpus parallèles et comparables ne servent pas seulement différents buts d'analyse, comme la traduction ou l'étude contrastive, mais ils se basent aussi sur différents aspects dans la création de corpus. Pour un corpus comparable, le choix de l'échantillon est d'une importance majeure tandis qu'un corpus parallèle ne doit pas forcément être adapté à ces critères, vu qu'il contient déjà un échantillon assez précis, à savoir les traductions des textes de source.
- 18 Dans cette contribution, je n'inclus pas les corpus parallèles ou comparables dans ma définition de corpus bi- ou plurilingue. Selon ma définition, un corpus bi- ou plurilingue ne consiste pas en un assemblage de textes monolingues, comme c'est le cas pour les corpus parallèles et comparables. J'évoquerais ici une définition plus restreinte de ce qu'est un corpus bi- ou plurilingue comme collecte de données bi- ou plurilingues. Ainsi, dans cette définition, je parle d'un corpus bi- ou plurilingue seulement si les données linguistiques en soi contiennent plusieurs langues. Je m'inscris dans le choix terminologique de Backus (2008 : 234-35), qui précise que chaque corpus de données linguistiques bilingues inclut, entre autres :
- une certaine quantité d'alternance codique entre les phrases (alternation)
 - un mélange de deux langues en une seule phrase (insertion)
 - d'autres types d'alternance ou de mélange codique
- 19 Dans ce qui suit, il sera d'abord question de discuter les enjeux méthodologiques de corpus oraux bi- et plurilingues. Dans une partie suivante, plusieurs corpus bi- et plurilingues seront présentés et analysés par rapport à la question de la comparabilité des résultats.

2.2. Enjeux méthodologiques

- 20 Les corpus bi- et plurilingues sont, comme ils ont été définis précédemment, généralement des corpus oraux de petite taille. Ils se distinguent en cela des corpus parallèles et comparables, qui généralement sont des corpus écrits de taille variable. Cette partie de la contribution est consacrée aux questions et choix méthodologiques à considérer lorsque l'on souhaite compiler un corpus bi- ou plurilingue oral. Avant de collecter des données linguistiques pour un corpus bi- ou plurilingue, le chercheur doit prendre plusieurs décisions importantes. D'abord, il doit parfaitement connaître le but de sa recherche, c'est-à-dire qu'il doit savoir précisément ce qu'il veut analyser. Cela est un aspect très important qui distingue les petits corpus des grands corpus dans leur démarche méthodologique (cf. ci-dessus).
- 21 Premièrement, il faut déterminer les participants des enregistrements. Plus particulièrement, il est indispensable de décider quelles langues et quel type et degré de bi- ou plurilinguisme seront analysés (cf. 2.1.). Généralement, en élaborant un corpus bi- ou plurilingue, les chercheurs visent à obtenir des données linguistiques très spécifiques sur une situation langagière bien déterminée. Cela implique que les chercheurs désirent obtenir des données linguistiques homogènes sur un groupe de locuteurs homogènes. Pour y parvenir, il peut être indispensable d'intégrer des méthodes additionnelles afin de recevoir toutes les informations nécessaires au sujet des locuteurs en question. Ces méthodes additionnelles peuvent consister en des questionnaires ou des entretiens sociolinguistiques, qui offrent des métadonnées sur le contexte sociolinguistique familial du locuteur ainsi que des informations sur ses choix et positions personnels au sujet des langues en question. Il est également possible d'employer des tests ou questionnaires langagiers, qui aident à déterminer le niveau et la compétence linguistique des locuteurs. Toutes ces méthodes préliminaires aident non seulement à garantir un groupe de locuteurs homogène, mais aussi à offrir la possibilité d'obtenir des informations importantes pour l'analyse linguistique ultérieure. Ces aspects peuvent même dominer la collecte des données pour un corpus bi- ou plurilingue, comme le remarque dal Negro (2013) au sujet de son corpus bilingue de locuteurs allemand-italien dans le Tyrol du Sud :

On the whole, the methodological approach to build the corpus has been more ethnographic in nature than experimental, a fact that has to be kept in mind when discussing the data and their representativeness in relation to the speech community. (dal Negro 2013 : 17)

- 22 Dal Negro souligne le fait que dans le cas des corpus bi- et plurilingues, les chercheurs ne visent généralement pas à investiguer des locuteurs individuels. Ils cherchent bien plus à obtenir des informations sur les dynamiques linguistiques au sein d'une communauté linguistique afin d'analyser les processus de contact ou de changement et leurs impacts. Il faut en plus être attentif à la situation d'enregistrement. Comme déjà mentionné, les locuteurs bi- et plurilingues choisissent leur langue selon différents aspects individuels et sociolinguistiques. Cela veut dire que les chercheurs doivent contrôler et choisir très précisément les situations d'enregistrement, qui incluent entre autres la composition du groupe de locuteurs, le choix du lieu d'enregistrement et la formalité de la situation. Il faut aussi considérer soigneusement que le statut et le prestige d'une des langues ou de l'usage simultané de plusieurs langues qui peut varier selon la communauté linguistique en question. Pour cela, il semble être utile de choisir plutôt des contextes familiers pour les enregistrements du langage bi- ou plurilingue et

de les combiner avec des méthodes additionnelles, comme l'évoque dal Negro (2013 : 17) :

Most recordings involve native speakers interacting in naturally occurring situations, such as family or friends reunions at home, in bars, in the canteen, etc. In a minority of cases the conversation takes the form of a semi-structured interview focused on languages, bilingualism and similar topics.

- 23 Dans ce contexte, les chercheurs doivent être très attentifs aux *languages modes* des locuteurs en question (Grosjean 2001). Dans le cas de locuteurs bi- ou plurilingues, qui agissent dans un environnement bi- ou plurilingue, il est indispensable de prendre en compte l'aspect du choix linguistique individuel pour l'analyse des données de corpus :

Differentiating bilingual and monolingual modes can methodologically be very important to classify and interpret correctly the data. For instance, it may help to distinguish interference and transfer from code-switching as well as inference from transfer. (dal Negro 2013 : 18)

- 24 Il va de soi que des locuteurs bi- et plurilingues n'utilisent que rarement les différentes langues exactement dans la même ampleur. Pour cela, il faut être conscient du fait qu'un corpus bi/plurilingue ne contient jamais de données linguistiques équilibrées sur les différentes langues. La question de la répartition des langues dans le corpus est d'une importance majeure pour l'analysabilité des données. Vu que, dans la plupart des cas, les corpus bi- et plurilingues sont des petits corpus, les chercheurs doivent s'assurer qu'ils contiennent une quantité d'unités linguistiques suffisante dans chaque langue pour être analysables. Cela dépend bien évidemment en grande partie de la problématique sous-jacente. Quelques phénomènes linguistiques apparaissent en effet fréquemment dans le langage oral, d'autres caractéristiques ne se manifestent que rarement.

- 25 Selon Backus, dans la conception d'un corpus bi- et plurilingues, il faut également respecter d'autres points importants, qui peuvent influencer l'étude complète. La représentativité d'un corpus est une question épineuse, qui se pose aussi dans la création de grands corpus monolingues (Backus 2008 : 238). La représentativité concerne la transférabilité des résultats à la réalité langagière. Les corpus de référence, comme par exemple le DEREKO (*Deutsches Referenzkorpus*) pour l'allemand ou le BNC (*British National Corpus*) pour l'anglais britannique, doivent respecter soigneusement la distribution d'échantillonnage des données linguistiques pour fournir un corpus le plus équilibré possible. Pour les corpus bi- et plurilingues, la représentativité est un aspect moins important parce qu'ils visent généralement à analyser un phénomène spécifique et qu'ils n'ont pas forcément pour but de représenter une communauté linguistique entière. Il est malgré tout, possible d'intégrer la représentativité dans la conception d'un corpus bi- et plurilingue :

A concern for representativeness can nevertheless be built into the corpus plan by creating the right circumstances for later elaboration of the corpus, by you or others. Slowly but surely, by filling in the gaps in informant selection, the corpus can become more and more representative of the community of whose speech it is a sample. (Backus 2008 : 238)

- 26 Pour ce faire, il est indispensable de collecter le plus grand nombre de métadonnées aussi diverses que possible.
- 27 Un autre problème méthodologique se pose après la collecte des données. Il concerne la transcription et l'annotation de ces dernières. Dans le cas des corpus bi- et plurilingues oraux, il faut passer par le processus de transcription et d'annotation pour obtenir un corpus exploitable. Dans un premier temps, il faut donc trouver des logiciels aptes à inclure plusieurs langues dans la transcription. Il existe plusieurs logiciels qui

contiennent des conventions de transcription pour des données bi- ou plurilingues. Je vais en proposer ci-après deux exemples qui seront placés dans le contexte des corpus bi- et plurilingues dans la partie suivante. Le premier exemple est le logiciel CLAN³, qui est gratuitement mis à disposition par l'équipe de *Talkbank* (cf. 3. ci-dessous). Le fondateur de *Talkbank* et de CLAN propose un format de transcription sous forme de conventions de transcriptions spécifiques – le format CHAT (MacWhinney 2000). L'alternance codique et d'autres interférences linguistiques entre plusieurs langues sont envisagées dans le manuel de transcription (MacWhinney 2000 : 97ff.). Dans la transcription en CHAT, le chercheur identifie la langue du locuteur à l'aide d'une ligne spécifique pour le langage (*language header tier*). Généralement, les chercheurs marquent les alternances entre les langues au début d'un changement de langue, comme dans l'exemple suivant :

*MOT : can you see?

*CHI : [-spa] no puedo. (MacWhinney 2000 : 97)

- 28 Par ailleurs, il est possible d'intégrer une ligne additionnelle de transcription pour les traductions et d'aligner chaque mot avec un tag de la langue correspondante (MacWhinney 2000 : 98).
- 29 Un autre logiciel qui prend en compte les conventions de transcriptions bi- et plurilingues est EXMARaLDA⁴. À l'aide d'un alignement en partition, EXMARaLDA permet d'inclure des informations spécifiques, par exemple la langue employée, une traduction ainsi que des informations non-langagières :

Figure 1. Transcription en partition selon EXMARaLDA⁵

Gön	((räuspert sich))	Ja.	Köstebeği
Gön [de]			Den Maulwurf hat er
Tan	Köstebeği mi?	Hmm	
Tan [de]	Den Maulwurf?		
Mut	Kimi korkuttu?		
Mut [de]	Wen hat er erschrocken?		

- 30 Cet exemple montre clairement les possibilités de transcription de données bilingues avec le logiciel EXMARaLDA et les conventions de transcription de HIAT (*Halbinterpretative Arbeitstranskription*).
- 31 Dans ce qui suit, il sera question de traiter la comparabilité des petits corpus bi- et plurilingues ainsi que le rôle et l'importance de banques de données et de conventions de transcription comparables.

3. Analyse de corpus bilingues et plurilingues – comparabilité des résultats ?

- 32 Un problème majeur des petits corpus bi- et plurilingues concerne leur comparabilité, c'est-à-dire la possibilité de généralisation des résultats de recherche. Ce problème ne provient pas seulement de la taille des corpus, mais aussi d'un manque de préparation et d'alignement des données ainsi que d'une absence de conventions de transcription

universelles. De plus, les petits corpus bi- et plurilingues sont généralement compilés par un chercheur spécifique dans le cadre d'un projet très spécifique. Backus le formule ainsi :

The individual researcher does not have the time, or the expertise usually, to get the corpus into such a state that it can be released to the wider academic community. (Backus 2008 : 233)

- 33 C'est sur ce point particulier que les grandes banques de données peuvent intervenir et accomplir différentes tâches. D'un côté, les banques de données peuvent proposer une plateforme de collecte et de stockage de données. En effet, des banques de données comme *Talkbank* proposent un site thématique sur lequel les chercheurs peuvent entreposer leurs données de corpus. Ce site est disponible et visible pour un grand public. Dans ce cas, les chercheurs évitent la tâche coûteuse de création et d'entretien d'un site web approprié ainsi que celle de commercialisation de leur corpus. En outre, les banques de données jouent aussi un rôle méthodologique pour les petits corpus bi- et plurilingues. Comme déjà mentionné, le site *Talkbank* propose non seulement un grand nombre de différents corpus thématiques, mais aussi une quantité de différentes ressources méthodologiques, comme le logiciel de transcription CLAN et les conventions de transcription CHAT. Ces ressources permettent de garantir une certaine homogénéité de la base de données parce que chaque corpus est transcrit et préparé selon ces règles. En outre, *Talkbank* offre diverses ressources supplémentaires, comme par exemple des informations sur l'enregistrement audio et vidéo ou des groupes Google pour permettre un échange scientifique entre les chercheurs participant au projet. L'une des principales caractéristiques de *Talkbank* est d'ailleurs la répartition et la subdivision des corpus selon certains sujets spécifiques. Les corpus bi- et plurilingues se trouvent surtout dans la partie *Multilingualism Banks* et ils incluent la *BilingBank*, une base de corpus bilingues, ainsi que la *SLABank*, une base de corpus concernant l'acquisition d'une langue seconde. D'autres corpus spécifiques bilingues se trouvent entre autres dans les bases de données CHILDES (banque de données sur l'acquisition du langage infantile) ainsi que *AphasiaBank* (banque de données sur l'aphasie).
- 34 *Sketchengine*⁶ est une autre banque de données certes payante, mais permettant le stockage et l'accès à différents corpus. *Sketchengine* offre à ses membres la possibilité de créer des propres corpus à partir des corpus disponibles (actuellement plus que 400 corpus différents dans une grande quantité de langues) ou de créer des propres corpus à partir de ses propres données. Dans ce dernier cas, *Sketchengine* permet d'abriter ses propres corpus et de les intégrer à l'interface de recherche de corpus afin de les rendre accessibles à un large public⁷. En général, *Sketchengine* contient des corpus très hétérogènes, qui varient beaucoup dans leur taille (petits et grands corpus) ainsi qu'en ce qui concerne les langues en question (corpus monolingues, dialectaux, bilingues, parallèles, etc.). Toutefois, *Sketchengine* abrite en grande partie des corpus qui viennent aussi d'ailleurs, comme par exemple quelques corpus de CHILDES ou les corpus parallèles d'Eur-Lex. Outre les corpus, *Sketchengine* permet un échange scientifique entre ses membres à l'aide de la diffusion de publications et de recherches pertinentes ainsi que d'un large choix d'ateliers en ligne et de conférences scientifiques. Ces ateliers peuvent offrir aux créateurs/instigateurs de petits corpus bi- et plurilingues une formation sur des questions importantes concernant la création et la comparaison de différents corpus.

- 35 Le HZSK (*Hamburger Zentrum für Sprachkorpora*) constitue un autre exemple de banque de données contenant des petits corpus bi- et plurilingues. Cette banque de donnée est étroitement liée à l'Université de Hambourg et au SFB *Mehrsprachigkeit*, un grand projet de recherche de la *Deutschen Forschungsgemeinschaft* (DFG) qui a été situé à l'Université de Hambourg de 1999 à 2011. C'est aussi dans ce contexte que le logiciel EXMARaLDA (cf. ci-dessus) a été créé. Aujourd'hui, le HZSK « accepts corpora and other linguistic resources from research projects and other contexts in order to make these available to the academic community for research and teaching purposes. The focus of the HZSK is oral multilingual corpora and oral corpora in other languages than German »⁸.
- 36 Cette citation montre que le HZSK représente une plateforme spécifique pour les corpus oraux bi- et plurilingues. Le HZSK contient 27 corpus, qui sont en grande partie des petits corpus bi- et plurilingues, en 22 langues différentes. Néanmoins, il est possible d'y trouver différents types de corpus, par exemple des corpus d'interprétation simultanée (CoSi et DiK), des corpus d'acquisition du langage bi- et plurilingue (p. ex. PAIDUS, PHONBLA, ALCEBLA, ENDFAS), des corpus monolingues avec traductions (NSLC) ou des corpus d'apprenants adultes d'une langue L2 (HaMoTiC). Outre les corpus, HZSK met également à disposition des ressources supplémentaires et méthodologiques, comme des logiciels de transcription (EXMARaLDA, CATMA) et des ateliers de transcription sur place ou en ligne⁹. Tout cela assure une certaine homogénéité des corpus de la banque de données et fournit aux chercheurs travaillant sur de petits corpus plurilingues une importante plateforme d'échange scientifique.
- 37 Pour conclure, il s'avère indispensable pour la comparabilité des petits corpus bi- et plurilingues de disposer d'outils et de ressources universels afin de créer une certaine équivalence dans la transcription et l'annotation des données. En outre, il est d'une importance majeure de regrouper les différents petits corpus bi- et plurilingues selon des sujets plus spécifiques et selon les types de corpus. Bien évidemment, les corpus bilingues oraux avec des locuteurs bilingues équilibrés ne sont pas comparables avec des corpus parallèles écrits ou des corpus comparables dans un langage scientifique, c'est pourquoi il faut toujours soigneusement considérer le type de corpus ainsi que le type de locuteurs avant de comparer des corpus bi- et plurilingues. Pour cela, il peut être très utile de regrouper des différents corpus par thème, comme on l'a fait dans le cas de la collection de petits corpus CLAPOTY¹⁰. Le projet CLAPOTY se centre sur des petits corpus des langues en contact et regroupe ainsi 94 corpus « des échanges plurilingues entre 40 langues et 290 locuteurs »¹¹. Cette collection de corpus vise alors à une analyse de phénomènes de contact, comme l'emprunt, l'alternance codique etc. (*ibid.*) (pour plus d'informations concernant les choix méthodologiques du projet CLAPOTY voir entre autres Vaillant et Légise 2014).
- 38 Les banques de données comme celles mentionnées ci-dessus occupent un rôle central dans le stockage et la commercialisation des petits corpus bi- et plurilingues ainsi que dans la distribution des ressources et des outils de linguistique de corpus. Pour les chercheurs, elles incarnent la meilleure façon de rendre accessible et visible à un grand public académique ses propres données de corpus.

4. Présentation d'un petit corpus bilingue : Le Corpus Franco-Manitobain (Corpus FM)

- 39 Dans cette partie finale de la contribution, un petit corpus oral bilingue sera présenté, le corpus Franco-Manitobain (Corpus FM) (Hennecke 2014). Le Corpus FM a pour but de donner un aperçu de la variété du français canadien parlé dans la province de Manitoba. Manitoba est la première province de l'Ouest canadien et forme avec la province de Saskatchewan les « provinces des prairies ». Selon le Censur de 2012, seulement 3,5 % des 1 267 000 habitants indiquent le français comme leur langue maternelle, bien que les francophones soient fortement enracinés au Manitoba depuis presque deux siècles. C'est pourquoi la variété du français manitobain est très influencée par la langue anglaise et présente beaucoup de phénomènes linguistiques d'interférence et de contact. Le Corpus FM fut construit au cours d'un projet de doctorat portant sur le changement linguistique induit par le contact linguistique. L'accent porta particulièrement sur les marqueurs pragmatiques (cf. Hennecke 2014). Le corpus a donc pour but de présenter les données linguistiques d'une communauté linguistique spécifique, qui vit depuis deux siècles en contact intense avec l'anglais. Dans ce cas, le choix d'établir un corpus bilingue s'imposait naturellement, parce que le langage oral des locuteurs franco-manitobains est fortement marqué par l'anglais : il contient un grand nombre d'alternance codique et d'emprunts. Toutefois, la distribution des langues dans le corpus n'est pas équilibrée, le français constitue généralement la langue de base et représente les deux tiers des expressions linguistiques :

Figure 2. Distribution des tokens dans le FM Corpus

(Hennecke 2014 : 67)

- 40 Le corpus contient 35 660 tokens en total, qui sont repartis sur 15 communications et un total de 20 locuteurs. Le grand nombre de locuteurs fut choisi intentionnellement pour éviter que l'usage idiosyncratique d'un locuteur spécifique domine dans le petit corpus. Cet usage idiosyncratique peut être particulièrement perturbateur dans l'analyse de phénomènes pragmatiques tels que les marqueurs pragmatiques, qui sont caractérisés par une forte variation individuelle.

- 41 Pour garantir un groupe de participants homogènes, tous les locuteurs dans le corpus ont entre 17 et 30 ans, sont nés et ont grandi dans l'environnement de la communauté Franco-Manitobaine de St. Boniface, le quartier francophone de la capitale du Manitoba, Winnipeg. Chaque participant a rempli un questionnaire linguistique sur son degré de bilinguisme à la fin des enregistrements et a signé une déclaration de consentement à l'utilisation des données enregistrées. Ainsi, nous disposons de certaines métadonnées importantes quant à l'usage des deux langues par les participants. Chaque participant a indiqué le français comme langue maternelle et a suivi, au moins jusqu'à l'âge de 17 ans, une éducation purement française. En raison de la situation sociolinguistique au Manitoba, tous les participants ont été régulièrement exposés à la langue anglaise dès leur petite enfance et peuvent ainsi être considérés comme bilingues équilibrés (Hennecke 2014 : 66). De plus, tous les locuteurs sont bien conscients du fait que la langue anglaise a un impact important sur leur variété de français, comme l'indique la citation suivante, tirée du corpus FM :

DM : c'est/ c'est différent (à cause) que t'as le franco-manitobain qui est très précis • t'sais • ah t'sais • very • ah proper • disons • puis ensuite t'as le parler de tous les jours right? • puis je sais que c'est la même chose au Québec • •
 puis/ puis partout right? • • mais • j' pas il y a quelque chose qui me dit que ((1.9s)) I don't know (()) ((1.1s)) t'sais j'avais cette discussion avec Maxime Bélanger puis lui il disait comme on devrait • arrêter de/ d'être tellement • •ehm ((1.5s)) c'est d'avoir ce complexe là qu'on parle mal • qu'on parle mal right? on parle la façon qu'on parle ici et il y a rien de mal avec ça. t'sais... Ya c'est ça c'est ça so on devrait juste ((2s)) t'sais • eh comment tu dis eeh ((1.5s)) embrace it • • ya • ya. (Hennecke 2014 : 66)

- 42 Cette citation montre très bien que les locuteurs franco-manitobains ressentent une certaine insécurité linguistique à l'égard de leurs compétences en français, un aspect qu'il convient de prendre en compte lors des enregistrements en vue de créer un corpus bilingue oral. C'est pourquoi il était donc indispensable d'effectuer les enregistrements dans des situations informelles entre franco-manitobains, afin que les locuteurs se sentent à l'aise et recourent à un registre langagier informel. Pour cette raison, les enregistrements eurent lieu dans des situations quotidiennes réelles, informelles (amis buvant un café ou un verre ensemble) comme partiellement informelles (rencontres entre collègues lors d'une pause de travail ou la réunion d'un groupe musical avant d'entrer en scène) (Hennecke 2014 : 65). Il était ainsi possible d'enregistrer des données authentiques de la variété franco-manitobaine et de l'influence de la langue anglaise sur cette variété. Cela peut être démontré dans l'exemple suivant, dans lequel un locuteur parle de la dernière tournée de son groupe musical et qui montre une certaine quantité d'alternances codiques entre les deux langues :

PJ : ya • c'est eh • • some gigs were like • there wasn't many people like we brought more people • than the local bands • puis c'était comme "what the hell?" • • eeh • mais il y avait d'autres gigs où est-ce que it was just like • there were a ton of people • puis • like we sold a/ • • li/ • • relatively good amount of stuff puis on est sorti comme • • n/ not losing money. so • c'était quand même • pas mal bon. (Hennecke 2014 : FM Corpus)

- 43 Le Corpus FM fournit des informations importantes quant au changement linguistique induit par le contact dans une communauté linguistique spécifique. Il convient ensuite de le comparer avec des données monolingues de corpus oraux du français contemporain pour explorer les cas où le changement linguistique s'opère. D'autre

part, il était possible d'effectuer des comparaisons avec des données bilingues d'autres communautés bilingues en contact avec l'anglais, par exemple une communauté croate-anglais (Hlavac 2006) ainsi que des communautés espagnol-anglais (p. ex. Lipski 2005, Torres et Potowski 2008). La comparabilité des résultats est assurée par le fait que les différents chercheurs détectent les mêmes types de phénomènes de contact avec l'anglais dans les différentes communautés linguistiques (cf. Hennecke 2014 : 130ff. pour une discussion de comparabilité des résultats). Il sera même possible de comparer les données avec les corpus du HZSK, parce que le Corpus FM suit également les conventions de transcription de HIAT et fut transcrit à l'aide du logiciel EXMARaLDA.

- 44 D'autre part, les données du corpus ont servi de base pour établir des hypothèses spécifiques, qui ont été plus tard vérifiées expérimentalement (cf. Hennecke 2014). Les investigations expérimentales avaient un double objectif : d'un côté, il était d'un intérêt majeur pour le projet de vérifier le traitement du langage bilingue (dans ce cas précis, des marqueurs pragmatiques) dans le cerveau bilingue et de le comparer avec le traitement du langage monolingue. De l'autre, les expériences psycholinguistiques permettaient de contrôler et de vérifier les résultats de l'analyse du corpus, c'est-à-dire la réalité psycholinguistique du changement linguistique induit par le contact (cf. Hennecke 2014 : 160ff.).

Conclusion

- 45 Le présent article visait à présenter un type de petits corpus très spécifique et hétérogène, les petits corpus oraux bi- et plurilingues. Ce type de corpus fait toujours face à une grande insécurité terminologique et méthodologique. Il était donc indispensable de souligner dans la présente contribution les avantages méthodologiques des petits corpus, c'est-à-dire leur méthodologie EHI ('early human intervention') (Sinclair 2010 : xi). Par ailleurs, les petits corpus sont consultables et vérifiables manuellement et constituent ainsi un outil approprié pour la recherche qualitative. En outre, les petits corpus ont l'avantage de traiter généralement un sujet très spécifique et, partant, de donner un aperçu important d'aspects linguistiques qui ne se trouvent que rarement dans les grands corpus, comme le plurilinguisme, l'acquisition du langage, les troubles du langage ou la variation dialectale.
- 46 Dans le cas de petits corpus bi- et plurilingues, il est essentiel de distinguer adéquatement entre corpus parallèles, corpus comparables et corpus bi- et plurilingues. Lorsque l'on considère les petits corpus oraux bi- et plurilingues, comme définis ci-dessus, il est important de considérer quelques aspects spécifiques dans la collecte des données et la création de corpus. Premièrement, un petit corpus oral bilingue ne traite généralement pas des locuteurs individuels, mais se concentre plutôt sur des communautés bilingues spécifiques. Cela implique que les chercheurs doivent impérativement considérer des aspects sociolinguistiques complexes de dynamiques langagières dans la communauté bilingue en question. Pour ce faire, ils peuvent recourir à des méthodes supplémentaires, comme des questionnaires ou des expériences psycholinguistiques.

Afin d'assurer la comparabilité des résultats de la recherche, les grandes banques de données jouent un rôle central. Elles peuvent influencer la visibilité et la diffusion des petits corpus et mettre à disposition des outils et ressources universels. De plus, elles aident à la formation des chercheurs en matière de linguistique de corpus et offrent

une plateforme d'échange scientifique. La comparabilité des petits corpus oraux bilingues est possible et même nécessaire si on se concentre sur un groupe de locuteurs comparables (p. ex. une communauté linguistique en contact avec l'anglais) et un phénomène spécifique (p. ex. l'impact du contact linguistique sur les marqueurs pragmatiques). Pour conclure, les petits corpus oraux bi- et plurilingues jouent un rôle primordial dans différents domaines de la recherche linguistique sur le plurilinguisme. Dans le futur, il faudra donc veiller à ce qu'ils soient accessibles à et visibles par un grand public académique afin d'intensifier et d'améliorer la diffusion et la comparaison des résultats.

BIBLIOGRAPHIE

- Aijimer K. et Altenberg B. (1996). « Introduction », in K. Aijimer, B. Altenberg et M. Johansson (éd.) *Languages in contrast. Papers from a symposium on text-based and cross-linguistic studies in Lund, 4-5 March 1994*. Lund : Lund University, 53-62.
- Backus A. (2008). « Data banks and corpora », in L. Wei et M. G. Moyer (éd.) *The Blackwell Guide to Research Methods in Bilingualism and Monolingualism*. Malden. Mass. : Blackwell Publ., 232-249.
- Cameron L. et Deignan A. (2003). « Combining Large and Small Corpora to Investigate Tuning Devices Around Metaphor in Spoken Discourse », *Metaphor and Symbol* 18(3) : 149-160.
- Dal Negro S. (2013). « Dealing with bilingual corpora : parts of speech distribution and bilingual patterns », *Revue française de linguistique appliquée* 18(2) : 15-28.
- de Beaugrande R. (2010). « Large corpora, small corpora, and the learning of 'language' », in M. Ghadessy, A. Henry et R. L. Roseberry (éd.) *Small corpus studies and ELT. Theory and practice*. Amsterdam/Philadelphia : John Benjamin, 3-31.
- Grosjean F. (2001). « The bilingual's language modes », in J. L. Nicol (éd.) *One mind, two languages : Bilingual Language Processing*. Oxford : Blackwell, 1-22.
- Hansen-Schirra S., Neumann S. et Čulo O. (éd.) (2017). *Annotation, exploitation and evaluation of parallel corpora*. Berlin : Language Science Press.
- Hennecke I. (2014). *Pragmatic markers in Manitoban French - a corpuslinguistic and psycholinguistic investigation of language change*. Thèse de doctorat : Ruhr-Universität Bochum.
- Hlavac J. (2006). « Bilingual discourse markers : Evidence from Croatian-English code switching », *Journal of Pragmatics* 38(11) : 1870-1900.
- Koester A. (2010). « Building small specialized corpora », in A. O'Keeffe et M. J. McCarthy (éd.) *The Routledge Handbook of Corpus Linguistics*. London : Routledge Publ., 66-79.
- Leech G. (2011). « Principles and applications of Corpus Linguistics », in V. Vander, S. Zyngier et G. Barnbrook (éd.) *Perspectives on Corpus Linguistics*. Amsterdam/Philadelphia : Benjamins, 155-170.

- Lipski J. M. (2005). « Code-switching or Borrowing? No sé so no puedo decir, you know », in L. Sayahi et M. Westmoreland (éd.) *Selected Proceedings of the Second Workshop on Spanish Sociolinguistics*. Somerville MA : Cascadia Proceedings Project, 1-15.
- Lüdi G. (1996). « Mehrsprachigkeit (Multilingualism/Plurilingualism) », in H. Goebel, P. H. Nelde, Z. Stary et W. Wölck (éd.) *HSK. Kontaktlinguistik / Contact Linguistics / Linguistique de Contact XII.1*. Berlin / New York : Walter de Gruyter, 233-245.
- McEnergy A. et Xiao Z. (2008). « Parallel and comparable corpora : What is happening ? », in G. Anderman et M. Rogers (éd.) *Incorporating corpora : Translation and the linguist*. Clevedon UK : Multilingual Matters, 18-32.
- McEnergy T. et Wilson A. (1996). *Corpus Linguistics*. Edinburgh : University Press.
- MacWhinney B. (2000). *The CHILDES Project. Tools for analyzing talk. The Database II* (3^e éd.). Mahwah/ New Jersey : Lawrence Erlbaum.
- Oksaar E. (1984). « Terminologie und Gegenstand der Sprachkontaktforschung », in Besch. Reichmann et Sonderegger (éd.) *HSK. Sprachgeschichte-Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung II.1*. Berlin : De Gruyter, 845-854.
- Pusch H. D. (2014). « Les corpus romans contemporains », in A. Klump, J. Kramer et A. Wilems (éd.) *Manuel des langues romanes*. Berlin/Boston : De Gruyter, 173-195.
- Sinclair J. M. (2001). « Preface », in M. Ghadessy, A. Henry et R. L. Roseberry (éd.) *Small corpus studies and ELT. Theory and practice*. Amsterdam/Philadelphia : John Benjamin, vii-xvii.
- Sinclair J. M. (2004). *Trust the Text : Language, Corpus and Discourse*. London : Routledge.
- Torres L. et Potowski K. (2008). « A comparative study of bilingual discourse markers in Chicago Mexican, Puerto Rican and MexiRican Spanish », *International Journal of Bilingualism* 12(4) : 263-279.
- Vaillant P. et Légise I. (2014). « À la croisée des langues. Annotation et fouille de corpus plurilingues », *Revue des Nouvelles Technologies de l'Information*, 81-100.

NOTES

1. <http://eur-lex.europa.eu/homepage.html?locale=fr>
2. Exemple tiré de <http://eur-lex.europa.eu/legal-content/DE-FR/TXT/?uri=CELEX:52000DC0608&from=EN>.
3. <http://talkbank.org/clang/index.php>
4. <http://exmaralda.org/de/>
5. http://www.exmaralda.org/hiat/Kapitel7_4.html
6. <https://www.sketchengine.co.uk>
7. <https://www.sketchengine.co.uk/user-guide/user manual/ corpora/ corpus-hosting/>
8. <https://corpora.uni-hamburg.de/hzsk/en/corpus-hosting>
9. <http://exmaralda.org/de/>
10. <http://clapoty.vjf.cnrs.fr/index.html>
11. <http://clapoty.vjf.cnrs.fr/donnees.html>

RÉSUMÉS

Cette contribution présente des types de petits corpus très hétérogènes, qui ne sont toujours pas clairement définis : les petits corpus oraux bilingues et plurilingues. Elle envisage d'abord de donner un aperçu de petits corpus et discute l'insécurité terminologique du domaine des corpus bi- et plurilingues. Ensuite, les bases de données de petits corpus bilingues *Talkbank*, *HZSK*, *Sketchengine* et *CLAPOTY* sont brièvement décrites pour discuter les possibilités de visualisation ainsi que de mise en disposition des données de corpus. Finalement, un exemple concret est présenté : le corpus FM, un petit corpus bilingue français-anglais d'une variété du français canadien, parlée au Manitoba.

This contribution presents a very heterogeneous type of small corpora, which still lacks a clear definition: small bi- and plurilingual corpora. First, it gives an overview of small corpora and discusses the terminological insecurity of research on and with bi- and plurilingual corpora. Then, the databases *Talkbank*, *HZSK*, *Sketchengine* and *CLAPOTY* are described in order to discuss the possibilities of visualization and availability of the corpus data. Finally, a concrete example is presented: The FM Corpus, a small French-English bilingual corpus of a variety of Canadian French, spoken in Manitoba.

INDEX

Mots-clés : corpus bilingue, corpus parallèles, corpus comparables, bases de données

Keywords : bilingual corpora, parallel corpora, comparable corpora, databases

AUTEUR

INGA HENNECKE

Institut de Philologie Romane
Universität Tübingen