Aspects théoriques et méthodologiques de la représentation des corpus

Najib Arbach, Saandia Ali

Corela

HS-13 (2013)
Statut et utilisation des corpus en linguistique

Najib Arbach et Saandia Ali

Aspects théoriques et méthodologiques de la représentativité des corpus

Référence électronique
Najib Arbach et Saandia Ali, « Aspects théoriques et méthodologiques de la représentativité des corpus », *Corela* [En ligne], HS-13 | 2013, mis en ligne le 15 mai 2014, consulté le 16 octobre 2014. URL : http://corela.revues.org/3029

Éditeur : Université de Poitiers
http://corela.revues.org
http://www.revues.org

Document accessible en ligne sur :
http://corela.revues.org/3029
Document généré automatiquement le 16 octobre 2014.

Najib Arbach et Saandia Ali

Aspects théoriques et méthodologiques de la représentativité des corpus

1. Introduction

1 Les raisons de la systématicité de la notion de représentativité dans les manuels tiennent dans ses objectifs ; Leech (1991 : 9) ou les statisticiens Manning & Schütze (1999 : 119) suggèrent qu'un corpus est représentatif si les conclusions basées sur l'analyse du corpus peuvent être généralisées à l'ensemble du langage étudié ; cette idée se retrouve chez Sinclair (2004) qui évoque les corpus de référence :

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

2 La représentativité est donc, avant tout, liée à un souci de scientificité des corpus : résultats fiables, pouvant être exploités et généralisés à l'ensemble du langage. Leech (2006 : 135) résume l'importance de la notion de la représentativité de la sorte :

> 'Without representativeness, whatever is found to be true of a corpus, is simply true of that corpus – and cannot be extended to anything else.'

3 Nous discuterons dans cet article des méthodologies de constitution d'un corpus représentatif, mais il convient auparavant d'évoquer la faisabilité de l'objectif. La question de la représentativité peut en effet être rattachée aux critiques de Chomsky (1965, 2004) qui voyait en tout corpus, fussent-ils aussi grands que le *Bank of English*, le *American National Corpus* ou le *British National Corpus* (*BNC*), une portion infime du langage et donc une représentation biaisée de celui-ci. La réponse à ce type de critique représenté par Chomsky[1] tient en un seul mot : l'échantillonnage, procédé inhérent à la création de corpus, tel que le formule Sinclair (2004) :

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

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

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

5 Ainsi la représentativité ne dépend pas uniquement du nombre de mots, mais également des catégories choisies, du nombre d'échantillons au sein de chaque catégorie et de la taille de chaque échantillon. Un manque de représentativité sur l'axe horizontal induit ce que Biber (1993b : 219) nomme « bias error », et que Habert (2000) traduit par « déformation ». Une déformation survient quand les caractéristiques linguistiques du corpus ne correspondent pas à celles de la population visée. Si le corpus n'est pas suffisamment représenté verticalement,

Biber parle alors de « random error », « incertitude » selon Habert. L'incertitude est due au fait que le corpus est trop petit pour que les conclusions qui en sont tirées soient généralisables.

6 Biber (1993a : 243) souligne que l'idée première des chercheurs qui abordent la notion de représentativité concerne la représentation verticale, à savoir le volume des données. Or il rapporte que la représentation verticale n'est pas la considération la plus importante dans le processus de sélection des échantillons : les questions de définitions de la nature des textes en ce qui concerne les corpus écrits, et des populations cibles en ce qui concerne les corpus oraux, sont des considérations non seulement plus importantes, mais également plus difficiles à prendre en compte.

7 La représentation horizontale peut être conduite selon deux méthodologies différentes : soit les catégories sont prédéfinies a priori, soit il est considéré que ces catégories ne peuvent être déduites que grâce à un corpus. Ce sont ces deux méthodologies que nous allons détailler ; la première est celle décrite dans un article de Biber abondamment cité, « Representativeness in corpus design » (1993a), la seconde est la méthodologie décrite généralement dans les travaux de Sinclair (1996, 2004).

8 Nous poursuivrons par la question du deuxième axe, la représentativité liée à la taille du corpus, et nous terminerons en dressant l'état des lieux actuel.

2. Stratification en amont

9 Comme nous l'avons évoqué, Biber (1993a) propose une constitution de corpus se basant sur une catégorisation en amont, puis la constitution du corpus selon la catégorisation obtenue. Or la stratification d'un corpus ne repose pas sur un type de paramètres unique, mais sur une liste plus ou moins modifiable de paramètres qui peuvent s'entrecroiser. Le schéma de catégorisation requiert donc l'inventaire de ces paramètres, ainsi qu'une hiérarchie entre eux.

10 La population étudiée est donc divisée en catégories (strates), et les catégories choisies sont pourvues. Les problématiques concernant le nombre de catégories et la quantité de données à pourvoir au sein de chacune d'entre elles trouvent leurs réponses avant la constitution du corpus. Biber propose (1993a : 245) comme exemple une hiérarchie d'échantillonnage détaillée, qu'il qualifie comme suit :

> '(…) a reduced set of sampling strata, balancing operational feasibility with the desire to define the target population as completely as possible.'

11 Biber réfute donc le principe de « la représentativité proportionnelle » qui aboutirait à un corpus à l'image de la langue étudiée. Selon lui, la proportionnalité n'est qu'un indicateur numéraire des fréquences des registres et ne pourvoit pas de représentation des registres « importants » ou « influents », tels les livres ou les journaux. Dans la sélection des textes, Biber se base donc sur un ensemble de critères principalement liés aux textes eux-mêmes, indépendamment des auteurs et récepteurs ; cette démarche dans la sélection implique un jugement que Váradi qualifie de subjectif et qu'il critique de la sorte (Váradi, 2001 : 591-592) :

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

12 Biber se justifie (Biber, 1993a : 247) en avançant qu'un corpus proportionnel représentera les registres linguistiques selon leur utilisation réelle, et suppose que ce type de corpus devrait contenir, selon ses estimations, 90 % de langue orale, 3 % de lettres et de notes et 7 % de registres recouvrant les reportages de presse, d'écrits de magazines populaires, de prose académique, de fiction, de conférences, de communiqués et d'écrits non publiés. Un tel corpus, quand bien même il serait constitué, ne susciterait que peu d'intérêt aux yeux de Biber (1993a : 247) :

> 'These kinds of generalizations, however, are typically not of interest for linguistic research. Rather, researchers require language samples that are representative in the sense that they include the full range of linguistic variation existing in a language.'

2.1. Critiques de la stratification non proportionnelle

13 Nous illustrerons les critiques de la représentativité non proportionnelle prônée par Biber avec deux exemples de corpus constitués selon sa méthode : le *Brown Corpus* en ce qui concerne l'écrit, et la partie orale du *BNC*.

14 Le *Brown Corpus* est un corpus censé représenter l'anglais américain écrit du début des années 1960, et il contient pour ce faire 500 échantillons de 2000 mots chacun, pour un total d'un million de mots ; nous ne pouvons détailler ici ses catégories[2], nous soulignerons uniquement que les critères ayant permis ces catégorisations qualitatives et quantitatives n'apparaissent ni dans le manuel du Brown, ni ailleurs ; Biber (1993a), en citant le *Brown* et le *London/Oslo/ Bergen Corpus* qui fut constitué sur le même modèle, prône leur modèle de constitution sans offrir un protocole de stratification rigoureux. Il en résulte que des corpus constitués selon cette méthodologie peuvent être soumis à la critique du point de vue de la représentativité. Citant le *Brown Corpus*, Váradi (2001 : 590) affirme que pour être représentatif de la population étudiée, un échantillon doit suivre le principe de la proportionnalité : les différentes catégories du corpus doivent être échantillonnées selon leur ratio au sein de la langue en général. Váradi (2001: 590) donne pour exemple:

> 'For the BROWN corpus to qualify as a representative sample of the totality of written American English for 1963 for humorous writing, it would have to be established that humorous writings did make up 1.8% of all written texts created within that year in the US. This single requirement serves to illustrate the enormous difficulty if not impossibility of the task.'

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

16 Là encore, l'échantillon démographique du *BNC* ne peut être considéré comme représentatif au sens proportionnel du terme. La stratification reposa sur une répartition en catégories égales et non selon les réalités démographiques de la société anglaise. D'autre part, une distribution proportionnelle aurait nécessité la consultation des données démographiques de la société anglaise ; le fait de ne pas avoir consulté ces données fut qualifié de « laxisme méthodologique » par Váradi. Burnard (1995 : 20) rapporte que l'enregistrement d'un nombre égal de locuteurs au sein de chaque catégorie était bien un objectif revendiqué, en reprenant pour justification des raisons similaires à celles avancées par Biber (voir supra), à savoir que la représentation proportionnelle n'est pas justifiée en raison de situations d'énonciation où un nombre restreint de locuteurs produit un nombre restreint d'énoncés mais qui seront destinés à un grand nombre de récepteurs.

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

2.2. Équilibre d'un corpus

18 Avant de poursuivre, il nous apparaît nécessaire de dire quelques mots d'une notion étroitement liée à celle de la représentativité dans la littérature : la notion de corpus équilibré. Généralement, un corpus est dit équilibré quand la taille de ses sous-catégories (genres, registres etc.) est proportionnelle à leurs fréquences d'occurrence au sein du langage général. En d'autres termes et selon Leech (2006 : 137), l'équilibre d'un corpus est synonyme de la proportionnalité que nous avons discutée supra. Nous rappelons que Biber rejeta cette proportionnalité pour prôner l'équilibre entre les catégories elles-mêmes. Nous abondons dans le sens de Leech qui considère qu'un corpus équilibré, au sens que Biber donne à ce terme (un corpus dont les différentes sous-catégories seraient identiques en volume), est un corpus qui n'a pas pour objectif d'être représentatif (Leech, 2006 : 139) :

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

3. Monitor corpus

19 La seconde grande approche de la représentativité des corpus, par rapport à la stratification en amont, est principalement représentée par John Sinclair. Sinclair propose l’échantillonnage en tant que solution à la problématique de la représentation du langage. En cela, sa démarche ne diffère pas de celle employée dans la représentativité proportionnelle puisque cette dernière repose sur l’échantillonnage également. Néanmoins c’est dans leurs conceptions de l’échantillonnage que les deux méthodologies divergent ; il convient de rappeler que Sinclair appartient au courant dit « corpus-driven » pour lequel – entre autres – « le sens des textes » est l’objectif primordial de la linguistique de corpus (voir Teubert, 2001). L’approche contextualiste de Sinclair est une approche probabiliste, que Léon (2008 : 19) décrit en ces termes :

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

20 En d’autres termes, la finalité des corpus serait de dresser une liste du ou des sens d’un « item lexical » (Teubert, 2010 : 7) à partir des collocations de cet item. Et ce sont les impératifs de cette dernière opération qui impliquent la notion d’un corpus « toujours augmentable et jamais fini » que Sinclair nomme « monitor corpus », au sein duquel la probabilité de cerner le sens d’un item lexical est proportionnelle au volume des données. Les « monitor corpus », ou corpus de référence dont nous allons détailler la constitution sont définis par Sinclair de la sorte (1991 : 17) :

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

Ce type d’approche est résumé par Habert (2000) de la sorte :

> "La conviction sous-jacente est que l’élargissement mécanique des données mémorisables (les centaines de millions de mots actuelles deviendront à terme des milliards) produit inévitablement un échantillon de plus en plus représentatif de la langue traitée. Si l’on n’arrive pas à cerner précisément les caractéristiques de l’ensemble des productions langagières, il ne reste qu’à englober le maximum d’énoncés possibles. À terme, la nécessité de choisir finirait par s’estomper."

21 La méthodologie d’échantillonnage de Sinclair (détaillée par exemple dans Sinclair, 2004) peut être résumée de la sorte : l’échantillonnage est un procédé devant prendre en considération les trois points suivants :

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

3.1. Orientation des textes

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

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

23 Sinclair préconise donc, au lieu des sélections élitistes ou normatives, de sélectionner les textes selon un ensemble de critères que nous discutons ci-dessous.

3.2. Critères de sélection

24 En ce qui concerne ces critères de sélection, Jeremy Clear (1992) distingue deux types de critères en jeu dans le choix des textes d'un corpus.

25 Le premier regroupe des considérations essentiellement linguistiques et représente « les critères internes » : la catégorisation d'un texte sur la base de critères syntaxiques et lexicaux sera ainsi une catégorisation basée sur des critères internes.

26 Le second type de critères, « les critères externes », concernent les informations métalinguistiques du texte, comme l'âge ou le sexe de l'auteur ou du locuteur.

27 C'est uniquement sur la base de critères externes que Sinclair préconise l'échantillonnage, sur les recommandations de Clear (1992 : 29) :

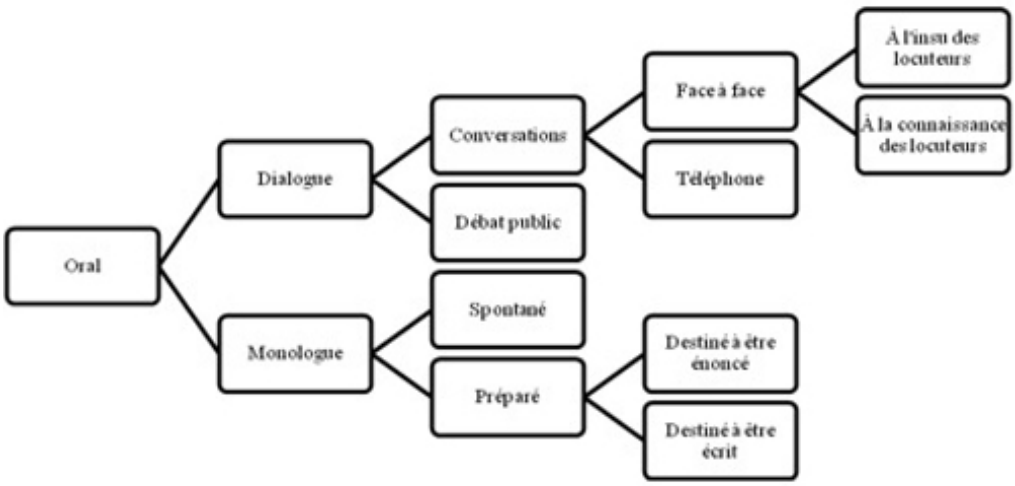
> 'A corpus selected entirely on internal criteria would yield no information about the relation between language and its context of situation. A corpus selected entirely on external criteria would be liable to miss significant variation among texts.'

28 Outre les exemples ci-dessus, Sinclair préconise l'emploi d'un nombre restreint de critères, clairement définis et établis de manière à parachever ou parfaire la représentativité du corpus.

3.3. Échantillonnage

29 La question qui se pose est ici l'échantillonnage, soit, en d'autres termes, le nombre de composants d'un corpus, le nombre de cellules au sein d'un corpus, ainsi que le volume des données au sein de chaque cellule.

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



31 Sinclair impose ensuite une condition : non pas des catégories strictement égales à l'image des catégories du *Brown Corpus*, mais des cellules qui posséderaient un nombre minimum de données. À chaque niveau supplémentaire, le nombre de mots du corpus se voit théoriquement au minimum doubler. C'est pourquoi Sinclair préconise l'emploi d'un nombre restreint de critères pour des raisons pratiques.

32 En ce qui concerne le nombre minimum de mots au sein de chacune des cellules, Sinclair indique que c'est une décision qui dépend principalement du type de recherche que l'on mènera sur le corpus, mais stipule que le volume des données se doit d'être « substantiel » afin que le chercheur puisse en tirer des conclusions scientifiques viables. Ainsi Sinclair donne, à titre

d'exemple, le nombre minimum d'un million de mots par cellule. C'est ici que l'on comprend les raisons de sa préconisation d'un nombre restreint de critères : dans l'exemple du *London Lund Corpus* ci-dessus, le nombre d'un million de mots par cellule implique un corpus oral de 16 millions de mots. L'ajout d'un seul niveau de critères aux cellules inférieures ferait de ce corpus un corpus de 32 millions de mots au minimum. Nous rappelons que la partie orale du *BNC* ne contient « que » 10 millions de mots et que le plus grand corpus oral constitué en France est le corpus du GARS, qui ne dépasse pas le million.

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

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

4. Taille des corpus

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

1. tous les théoriciens de la linguistique de corpus n'ont pas différencié la représentation horizontale de la représentation verticale, et certains traitent le corpus en tant qu'entité ;
2. l'existence de corpus dit spécialisés, tels les corpus de personnes âgées, corpus d'acquisition, corpus historiques ou des corpus d'apprenants.

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

37 Avant de rapporter les éléments de réponse que nous avons retrouvés dans la littérature, il est nécessaire de préciser que la question ne peut être résolue uniquement sur des critères statistiques. Deux points sont à prendre en compte, en amont et indépendamment des principes linguistiques et mathématiques :

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

38 Par exemple, les premiers corpus numérisés tel le *Brown Corpus* et le *LOB* durent faire face à certaines difficultés logistiques induites par la numérisation de données sur papier. Ces difficultés furent surmontées avec l'utilisation de scanners à reconnaissance optique des caractères, mais le problème persiste en ce qui concerne la collecte des données orales, la reconnaissance vocale n'étant pas aussi performante que la reconnaissance optique. Les corpus oraux impliquent donc toujours l'enregistrement de données et la transcription de celles-ci, deux étapes lourdes en temps et en ressources humaines en raison du travail de terrain nécessaire à la première, et de la transcription manuelle des données. Meyer (2002 : 32) rapporte ainsi que le *Santa Barbara Corpus of Spoken American English* nécessita six heures de travail pour la transcription et l'annotation d'une minute de parole, et que « ces faits logistiques expliquent pourquoi 90 % du BNC est écrit et que seulement 10 % est de la langue parlée ». Ces difficultés matérielles impliquent, toujours selon Meyer (2002 : 32), des considérations triviales quant à la taille désirée du corpus :

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

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

40 La question des objectifs linguistiques du corpus est même primordiale pour Kennedy qui doute de la nécessité des méga-corpus – nous reviendrons sur ce point infra – et propose de privilégier la qualité des données plutôt que leur quantité (Kennedy, 1998 : 68) :

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

41 La qualité des données signifie précisément pour Kennedy la prise en considération des objectifs linguistiques du corpus : Kennedy poursuit en mentionnant quelques exemples, selon lesquels la quantité de données nécessaire à une analyse prosodique serait de 100 000 mots, sous la condition que ces données soient du type « parole spontanée » ; ou qu'une étude sur la morphologie des formes verbales nécessiterait entre 500 000 et un millions de mots.

42 Considérons maintenant les autres avis concernant la taille des corpus. En 1992, Geoffrey Leech (Leech, 1991) prédisait pour 2021 un ensemble de corpus de mille milliards de mots. Leech n'effectue ces calculs que pour souligner la relative importance de la taille d'un corpus, jugée comme « critère naïf », cela pour quatre raisons :

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

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

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

45 Biber propose dans son étude (1993a : 248) un calcul strictement mathématique de la taille d'un corpus pour que celui-ci soit verticalement représentatif. L'équation sur laquelle Biber repose son étude permet de calculer l'erreur type (σ_m) d'un échantillon, qui est égal à l'écart type (σ) divisé par la racine carrée de la taille de l'échantillon (n), soit :

$$\sigma_m = \sigma / n1/2$$

46 Or Biber discute lui-même du principal obstacle à de tels calculs : calculer le potentiel de représentativité d'un échantillon suppose avoir à disponibilité a priori d'un échantillon représentatif. Or il n'existe pas, en ce qui concerne par exemple le français, un corpus oral

scientifiquement reconnu comme représentatif de la langue. L'application des calculs de Biber nécessitait pourtant de tels corpus (Biber, 1993a : 256) :

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

4.1. Les méga-corpus

47 La nécessité de corpus de très grande taille est défendue par Sinclair, qui considère que « grand » est la valeur par défaut de la quantité de données (Sinclair, 1996) et préconise un corpus « aussi grand que possible », se basant en cela sur la loi de Zipf (1991 : 18) :

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

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

49 Il y a moins deux obstacles à ce type de méthodologie : en premier lieu, que le linguiste prévoie le taux de fréquence minimal d'un phénomène suppose que tout compilateur de corpus sache, dès la constitution du corpus, ce à quoi il servira, ce qui n'est pas le cas et ce qui ne devrait de toute façon pas être le cas. En second lieu, dans le cadre d'un corpus de référence, la quantité des données nécessaire à l'obtention de la double occurrence de certains phénomènes relève de l'inaccessible, a fortiori quand il s'agit de données orales ; à titre d'exemple, Geyken (2008 : 82) rapporte les résultats d'une étude selon laquelle « il faudrait disposer d'un corpus représentant cinquante années du journal *Le Monde* si l'on voulait « voir apparaître au moins une fois tous les mots composés recensés selon les critères définitoires du lexique-grammaire ». En nous fondant sur les chiffres d'une autre étude menée sur un corpus de presse compilé à partir du *Monde*[3], nous avons estimé qu'un tel corpus représenterait environ un milliard de mots. Sachant que si une occurrence apparaît dans un corpus de n mots, elle n'apparaîtra pas obligatoirement une seconde fois dans le même corpus doublé à 2 x n mots ; sachant également qu'il y a sans doute des faits linguistiques encore plus rares que les mots-composés cités dans l'exemple ci-dessus, nous pouvons en conclure qu'un corpus horizontalement représentatif au sens entendu par Sinclair[4] serait un corpus de x milliards de mots.

50 La théorie de Sinclair est toutefois en phase avec ses applications pratiques, puisque le corpus *Bank of English* sur lequel travailla Sinclair, comporte environ 650 millions de mots. D'autre part, non seulement les projets de Sinclair ont permis la création du premier dictionnaire entièrement constitué sur corpus (projet COBUILD[5]), mais Kennedy (1998 : 70) rapporte que l'analyse des collocations au sein de grands corpus a permis l'identification de nouvelles variantes, de nouveaux modèles de construction, voire de nouvelles caractéristiques grammaticales. Kennedy illustre son propos en citant une étude de Sinclair (1989) sur la préposition « of », dans laquelle il démontre que la description linguistique de l'une des plus fréquentes prépositions de la langue anglaise dans les grammaires antérieures ne correspond pas totalement à ce qu'il a pu constater dans les corpus.

4.2. La nécessité de corpus moins grands, plus spécifiques

51 Qu'un corpus soit plus grand qu'un autre, ou enrichi avec le temps, peut paraître indiscutablement positif. Mais les auteurs sont nombreux à plaider pour des corpus plus spécifiques, mieux construits, plus accessibles et surtout plus adaptés aux besoins du linguiste. Nous citerons notamment Cappeau & Gadet (2007 : 101), qui rappellent judicieusement que si l'évolution de l'informatique a permis la constitution et l'exploitation de corpus de grande

taille, cette évolution doit rester pour les linguistes « une condition nécessaire, mais non suffisante, pour espérer disposer d'un recueil exploitable. » Les auteurs poursuivent en mettant en garde contre les expansions injustifiées d'un corpus :

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

52 Il est nécessaire de préciser que ces réflexions concernent l'état actuel des choses, à savoir la relative abondance des données textuelles, mais que la difficulté à obtenir des données orales transcrites reste d'actualité[6].

53 Pour résumer la question de la taille adéquate des corpus, nous dirons que bien qu'il faille des corpus de très grande taille pour des études lexicographiques, le calcul des collocations ou une description exhaustive de la langue qui reposerait uniquement sur des données attestées, la qualité des résultats de dépend pas du volume des données, et pourrait même en pâtir. En effet, il est de notoriété que plus de 90 % des phénomènes langagiers apparaissent dans des corpus restreints ; d'une part le gain obtenu grâce aux méga-corpus ne compensera pas les efforts matériels et humains fournis pour la constitution de tels corpus, et d'autre part la manipulation d'un volume de données largement inaccessibles à l'esprit humain risque grandement d'être erratique.

54 Déterminer la taille d'un corpus amènera donc le chercheur à définir ses besoins, prendre en compte les possibilités humaines et matérielles entrant en jeu ainsi que les outils d'exploitation dont il dispose. Dans une configuration logistique idéale, la linguistique de corpus n'est pas – encore ou ne le sera-t-elle jamais, nous l'ignorons – une science exacte. À ce propos, nous conclurons par cette phrase de Kennedy (1998 : 68) :

> 'At this stage we simply do not know how big a corpus needs to be for general or particular purposes.'

5. La représentativité, état des lieux et conclusion

55 La représentativité ne saurait dépendre des critères linguistiques des données mais est définie par les métadonnées du corpus, soit la documentation du corpus ; autrement dit la représentativité d'un corpus ne saurait être appréciée sans que les critères de sélection des auteurs des textes, ainsi que les textes eux-mêmes soient clairement définis et disponibles.

56 Par ailleurs, un article de Leech (2006) montre bien qu'à l'heure actuelle, aucun corpus n'est unanimement reconnu comme représentatif du langage. Cela vient du fait que les méthodologies employées pour représenter le langage dans son ensemble ne font pas encore consensus. D'ailleurs, Cappeau & Gadet (2007 : 108-109) remettent en question l'idée même d'un grand corpus qui serait représentatif de tous les aspects du langage.

57 Il est également à noter que la problématique de la représentativité diffère selon le type de corpus ; elle est d'ordre horizontal quant aux données écrites en raison de la relative facilité à regrouper des textes écrits, tandis que la représentativité verticale concerne davantage les données orales du fait des moyens matériels et humains considérables à engager pour l'obtention de données orales retranscrites et constituées selon des protocoles scientifiques.

58 En raison de ces trois points (la représentativité dépend de la documentation des corpus, un seul corpus ne saurait être représentatif du langage, le manque de données concerne les corpus oraux), il s'est développé une certaine idée que nous pourrions appeler la « représentativité participative » : il ne s'agit plus de prétendre représenter le langage en un seul corpus, mais de participer à la représentation du langage par le corpus constitué, les différents corpus constitués indépendamment pourraient alors être regroupés ou consultés simultanément. Cela suppose l'utilisation du corpus par des chercheurs qui ne l'ont pas constitué. Bien que cette idée paraisse évidente, force est de constater que la plupart des corpus constitués en France n'ont servi que leurs propriétaires[7]. À propos de la représentativité participative des corpus, nous laissons la parole à Cappeau & Gadet (2007 :108-109) :

> "Une autre possibilité envisagée récemment serait d'œuvrer pour un regroupement des corpus d'équipes différentes. Ce rapprochement de disparates permettrait de disposer rapidement de « gros » corpus oraux. L'état des besoins, la demande de chercheurs qui souhaitent étendre leurs

> investigations aux corpus oraux rend cette solution attrayante. Elle pose néanmoins des questions qui ont été indiquées dans ces pages, et elle repose sur des mythes - qui arrivent trop tardivement pour être considérés comme fondateurs : l'illusion d'une transcription minimale fructueuse pour toutes les disciplines intéressées à l'oral, l'illusion que toutes les situations présentent le même intérêt pour toutes les études, bref que tout corpus serait bon pour tous et pour tout !"

59 Ainsi la possibilité de la réexploitation des corpus par des chercheurs qui ne les ont pas constitués rencontre des obstacles que nous résumons ainsi :

1. droits d'exploitation et de diffusion ;
2. la transcription minimaliste ou standardisée des données en ce qui concerne les corpus oraux ;
3. la documentation du corpus.

60 La question d'une transcription minimaliste et standardisée implique l'imposition du protocole de transcription aux chercheurs ; cela ne serait effectivement pas scientifiquement rentable, et n'est donc pas souhaitable. En revanche, 1) une documentation selon les standards proposés[8], 2) des protocoles rigoureux en ce qui concerne les droits d'exploitation des corpus et 3) la mise à disposition et la diffusion des corpus sont des démarches aisées en comparaison aux efforts fournis pour la compilation des données du corpus.

61 Il semble donc nécessaire, pour la représentativité des corpus, que ces démarches soient davantage respectées en France, via l'institutionnalisation de la linguistique de corpus au sein d'initiatives telle celle de Corpus (Infrastructure de Recherche)[9], afin de limiter le phénomène de « corpus fantômes » selon l'expression de Baude & Abouda (2006 :3), qui considèrent qu'un corpus non disponible (pour des questions de droits d'exploitation ou pour d'autres raisons) n'existe pas, et ne représente – de fait – rien.

Bibliographie

BAUDE, O. & ABOUDA, L. (2006). "Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO". Consulté sur : http://icar.univ-lyon2.fr/ecole_thematique/idocora/documents/Abouda-Baude-ESLO.pdf

BIBER, D. (1993a). "Representativeness in corpus design". *Literary and linguistic computing*, 8 (4), 243–257.

BIBER, D. (1993b). "Using register-diversified corpora for general language studies". *Computational linguistics*, 19(2), 219–241.

BURNARD, L. (1995). *Users Reference Guide British National Corpus*. Oxford : University Computing Service.

BURNARD, L. (2007). "Une introduction au British National Corpus dans son édition XML". *Texte et Corpus*, 3, 17-34.

CAPPEAU, P. & GADET, F. (2007). "L'exploitation sociolinguistique des grands corpus". *Revue française de linguistique appliquée*, 12 (1), 99–110.

CHOMSKY, N. (1965). *Aspects of the Theory of Syntax* (Vol. 119). Cambridge (Mass.) : The MIT Press.

CHOMSKY, N. (2004). "The master and his performance. Interview by Jozsef Andor". *Intercultural Pragmatics*, 1(1), 93–111.

CLEAR, J. (1992). "Corpus sampling". *New directions in English language corpora,* 21–31.

Francis, W. N. (1991). "Language corpora BC". *Directions in Corpus Linguistics : Proceedings of Nobel Symposium 82*, Stockholm, 4—8 August 1991, 17–31.

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

GRANGER, S. (2007). "Corpus d'apprenants, annotation d'erreurs et ALAO : une synergie prometteuse". *Cahiers de lexicologie,* 91, 2, 117-132.

HABERT, B. (2000). "Des corpus représentatifs : de quoi, pour quoi, comment". In Bilger, M. (ed.) : *Linguistique sur corpus. Etudes et réflexions*, (31), 11–58, Perpignan : Presses Universitaires de Perpignan.

KENNEDY, G. (1998). *An introduction to corpus linguistics*. Londres : Longman.

LEECH, G. (1991). "The state of the art in corpus linguistics". *English corpus linguistics : studies in honour of Jan Svartvik* (p. 8-29). Londres : Longman Publishing Group.

LEECH, G. (2006). "New resources, or just better old ones ? The Holy Grail of representativeness". *Language and Computers*, 59 (1), 133–149.

LÉON, J. (2005). "Claimed and unclaimed sources of corpus linguistics". *Henry Sweet Society Bulletin*, 44, 36–50.

LEON, J. (2008). "Aux sources de la « Corpus Linguistics » : Firth et la London School". *Langages*, 3, 12–33.

MANNING, C. & SCHÜTZE, H. (1999). *Foundations of statistical natural language processing* (Vol. 59). Cambridge (Mass.) : The MIT Press.

MEYER, C. F. (2002). *English corpus linguistics : An introduction*. Cambridge : Cambridge University Press.

SINCLAIR, J. (1989). "Uncommonly common words". in Tickoo, M. (ed.) *Learner's Dictionaries : State of the Ar*t, Anthology Series 23, RELC, Singapore, 135-52.

SINCLAIR, J. (1991*). Corpus, concordance, collocation*. Oxford : Oxford University Press.

SINCLAIR, J. (1996). "Preliminary recommendations on corpus typology". EAGLES. Consulté sur : http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html

SINCLAIR, J. (2004). "Corpus and Text - Basic Principles". *Developing Linguistic Corpora : a Guide to Good Practice*. Consulté sur : http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

TEUBERT, W. (2001). "Corpus linguistics and lexicography". *International Journal of Corpus Linguistics*, 125–153.

TEUBERT, W. (2010). "Corpus Linguistics : An Alternative". *Semen, Critical Discourse Analysis I. Les notions de contexte et d'acteurs sociaux*, 27. http://semen.revues.org/8912

VÁRADI, T. (2001). "The linguistic relevance of corpus linguistics". *Proceedings of the Corpus Linguistics 2001 Conference*. UCREL Technical Papers, 13, 587–593.

Notes

1 Nous ne nous positionnons pas dans le courant qui considère l'avant ou l'après Chomsky. Ses critiques ne stoppèrent ni ne ralentirent les travaux sur corpus. Voir à ce sujet Léon (2005).

2 Voir le manuel du corpus : http://khnt.aksis.uib.no/icame/manuals/brown/

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

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

5 *Collins Birmingham University International Language Database*, pour la création du *Collins COBUILD English Language Dictionary*.

6 Comme exemple des « dérives » dans la surabondance des données textuelles, voir Burnard (2007 : 21).

7 Nous utilisons le terme de « propriétaire » à dessein, car la question des droits d'exploitation est un obstacle majeur à la diffusion des corpus, sans en être le seul.

8 Conventions OLAC pour les métadonnées ou TEI.

9 http://www.corpus-ir.fr/

Pour citer cet article

Référence électronique

Najib Arbach et Saandia Ali, « Aspects théoriques et méthodologiques de la représentativité des corpus », *Corela* [En ligne], HS-13 | 2013, mis en ligne le 15 mai 2014, consulté le 16 octobre 2014. URL : http://corela.revues.org/3029

À propos de l'auteur

Najib Arbach
LIDILE EA3874, Université Rennes 2
najibarbach@hotmail.com, saandia.ali@uhb.fr

Droits d'auteur

Résumés

En 1982, Francis (1991 :17) définit un corpus comme suit :
'A collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis.'
Le critère de la représentativité allait ensuite être évoqué par la quasi-totalité des ouvrages et articles de référence sur la linguistique de corpus. Cet article tentera de définir la représentativité en illustrant ses axes, et d'expliciter les méthodologies de la représentativité qui incluent les notions de catégorisations, d'échantillonnage et de volume des données.
Pour ce faire, nous tenterons de comprendre l'importance de cette notion et de sa récurrence au sein de la littérature traitant de la linguistique de corpus. Nous distinguerons ensuite les différentes méthodologies employées dans le but d'atteindre la représentativité dans la constitution de corpus. Les deux principaux courants méthodologiques que nous examinerons sont ceux de la « stratification en amont » représenté par Biber (1993a, 1993b) pour le premier, et celui des « monitor corpus » représenté par Sinclair (1991, 1996, 2004) pour le second. Nous nous intéresserons en détail à la question de la taille des corpus, et nous conclurons par une revue rapide de la situation actuelle, accompagnée de quelques recommandations destinées aux compilateurs de corpus constitués ou futurs.

In 1982, Francis (1991: 17) defines a corpus as:
'A collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis.'
The representativeness of a corpus would then be taken into account by most of the main publications which dealt with corpus linguistics. This paper aims at defining the concept of representativeness in corpus design and at illustrating its main features as well as the various methods used to achieve it, which will include a discussion on the issues of categorization, sampling or the required size of a corpus.
We will try to achieve a better understanding of the concept of representativeness through a review of the related literature on corpus linguistics. The various methods that are proposed and implemented in order to achieve representativeness in corpus design will be discussed and contrasted. The two main methods that will be examined are Biber's stratification techniques (1993a, 1993b) on the one hand, and the methods represented by Sinclair's "monitor corpus" (1991, 1996, 2004) on the other hand. Finally, we will address the issue of the required size of a corpus and provide a brief review of the current situation regarding corpus design along with some recommendations for corpus building.

Entrées d'index

Mots-clés : constitution de corpus, méthodologies en linguistique de corpus, représentativité, structure des corpus, taille des corpus
Keywords : design corpora, methodology in corpus linguistics, representativeness, corpora structure, corpora size