CORPUS

Corpus

18 | 2018
Les petits corpus

De la représentativité à la spécialisation : exemple d'un petit corpus sur la synonymie

From representativity to specialisation : the case of a small corpus about synonymy

Gaëlle Doualan

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/3331
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Référence électronique
Gaëlle Doualan, « De la représentativité à la spécialisation : exemple d'un petit corpus sur la synonymie », *Corpus* [En ligne], 18 | 2018, mis en ligne le 09 juillet 2018, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/3331

Ce document a été généré automatiquement le 8 septembre 2020.

De la représentativité à la spécialisation : exemple d'un petit corpus sur la synonymie

From representativity to specialisation : the case of a small corpus about synonymy

Gaëlle Doualan

Introduction

1 La question du corpus constitue un enjeu majeur en sciences du langage, puisque le corpus est devenu le point central de la méthodologie en linguistique. Le choix du corpus, sa méthode de constitution et les objectifs de recherche qui président à sa constitution façonnent les études sur le langage et donc leur validité scientifique, d'où une attention particulière à porter à la notion de corpus. Reprenant et précisant la définition du corpus donné par Sinclair (1996 : 4), Habert (2000 : 13) écrit :

> Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue.

2 Le corpus échantillonné[1] vise donc à saisir un fait linguistique au travers d'échantillons représentatifs, car la réalité langagière est trop vaste pour être saisie en globalité[2]. « De sorte que l'on est conduit à faire l'hypothèse (le pari) que les régularités susceptibles d'être découvertes par l'analyste sont potentiellement récursives et donc qu'une analyse limitée à un sous-ensemble de faits peut être de nature à rendre compte de l'ensemble » (Dalbera 2002 : § 7). De ce fait, le corpus constitue l'observatoire que se donne le chercheur pour appréhender le fait linguistique, d'où la nécessaire attention portée à la constitution du corpus : du point de vue épistémologique, il est l'intermédiaire entre le fait empirique et les hypothèses théoriques. Un corpus échantillonné peut être établi pour divers objectifs :

> appréhender et donner à voir cette réalité trop vaste pour être embrassée dans sa totalité (par exemple, décrire le français oral du XX^e^ siècle, comparer l'anglais britannique et l'anglais américain, etc.) ; se donner les bases empiriques nécessaires pour répondre à un questionnement théorique ou étayer une hypothèse structurale (par exemple décrire, comprendre et unifier les emplois du conditionnel dans le système verbal français contemporain) ; constituer enfin les bases de connaissances indispensables au développement des nouveaux outils réclamés par l'essor de l'ingénierie linguistique.
> (Mellet 2002 : § 4)

3 Si le corpus échantillonné est construit dans un but donné, il est assujetti à une hypothèse de travail, mais les biais présidant à sa constitution doivent être contrôlés par cette hypothèse de travail. Le corpus devient alors un lieu d'expérimentation : il permet « i) de déterminer à volonté la production du phénomène ; ii) d'isoler les effets des différents paramètres constituant le phénomène. À l'aide de ce montage, on doit pouvoir corroborer ou infirmer une hypothèse précise » (Auroux 1998 : 166). Enfin, il importe de considérer à quelle étape de la recherche intervient le corpus : lors de l'analyse des données ou lors de la confrontation de l'hypothèse aux résultats. « Le recours explicite au corpus peut intervenir dans une phase liminaire de la recherche au moment où l'on tente de cerner les faits pertinents ou en fin de recherche au moment de valider les hypothèses émises » (Dalbera 2002 : § 10)[3]. En somme, le corpus est un objet construit et en tant que tel il est indispensable d'exposer les objectifs qui ont présidé à sa constitution ainsi que la manière dont il a été construit, ces deux aspects faisant partie du cercle herméneutique appliqué aux données linguistiques.

4 Le développement des nouvelles technologies, par l'accroissement des possibilités de stockage, n'a eu de cesse de modeler la notion de corpus[4]. Ces changements technologiques ont suscité des questionnements méthodologiques et donc épistémologiques sur la notion de corpus.

> *What started as a methodological enhancement but included a quantitative explosion (I am referring here to the quantity of data processed thanks to the aid of the computer) has turned out to be a theoretical and qualitative revolution in that it has offered insights into the language that have shaken the underlying assumptions behind many well-established theoretical positions in the field.*
> (Tognini Bonelli 2010 : 17)
> *Halliday foresaw the signs of a qualitative change in the results of the quantitative studies opened up by corpus research. He warned that not only language but semiotic systems in general would be affected by this new proximity of theory and data.*
> (Halliday & James 1993 : 1-25).
> *This is clearly a stage beyond methodology*.
> (*Ibid.* : 18)

5 La possibilité de créer des corpus de plus en plus grands – plusieurs millions de mots – n'a eu de cesse de reposer la question de la représentativité des corpus :

> *This view of the need for large corpora was summed up by Sinclair when he said that 'The only guidance I would give is that a corpus should be as large as possible and keep on growing' (1991 : 18). Sinclair based this need for large corpora on the fact that words are unevenly distributed in texts and that most words occur only once. Thus 'In order to study the behaviour of words in texts, we need to have available quite a large number of occurrences' (Sinclair 1991 : 18). While this view of corpora was the prevailing one, it did not go unchallenged.*
> (Nelson 2010 : 55)

6 et leur capacité à rendre compte de phénomènes linguistiques récurrents.

> *What do you get from a large corpus that you do not get from a small one ? Essentially you get repetitions of multi-word choices in combination. The large number of words in a language, and their characteristically uneven distribution (Zipf 1935) mean that despite the clear tendency of languages to practice coselection, that coselection is subject to so much variation that if one wants to study collocation or phraseology by automatic methods then even the 9-figure corpora are pitifully small.*
> (Sinclair 2001 : X)

7 Les grands corpus sont donc très utilisés par toutes les recherches nécessitant des statistiques sur les langues, comme notamment la lexicographie.

> *In this regard, lexicographers led the way. Their aim has always been to collect the maximum amount of data possible, so as to capture even the rare events in a language.*
> (McCarthy & O'Keeffe 2010 : 6)

8 À l'ère des grands corpus[5], avec les apports des technologiques informatiques et les injonctions à créer des corpus de plus en plus grands, comment est-il possible que les petits corpus continuent d'exister ? Quelles sont leurs conditions de possibilité pratiques et scientifiques qui justifient la coexistence de petits corpus à côté des grands corpus ? Comment se définissent les petits corpus par rapport aux grands corpus ? Que permettent-ils en termes de recherches que ne permettent pas de grands corpus ?

9 Dans une première partie, la notion de petit corpus est longuement discutée en s'arrêtant sur les questions de taille, de représentativité et de spécialisation du corpus. La deuxième partie expose le cadre de recherche qui a nécessité un petit corpus : il s'agit d'une étude de la synonymie en corpus. Enfin, la troisième partie présente le petit corpus constitué pour cette étude en détaillant les contraintes méthodologiques rencontrées lors de cette élaboration.

1. Réflexions sur les petits corpus

10 Puisque l'objectif de recherche détermine le corpus à construire, divers questionnements émergent quant aux paramètres à établir lors de la constitution du corpus :

> *there then come crucial decisions to make regarding the size of the corpus, of how it should be balanced, the sampling methods to use, the kinds of texts that should be used, the use of full, or samples of, text and how representativeness could be achieved. Further issues concern whether you want to create a corpus for a specialist purpose or for more general purposes.*
> (Nelson 2010 : 53)

11 Selon les réponses apportées à ces questions méthodologiques, le meilleur choix peut être le petit corpus. S'il est souvent fait cas des grands corpus et de leur représentativité, il faut pouvoir définir le petit corpus et ses spécificités. Si la taille, c'est-à-dire la quantité de textes ou de mots, est un critère de définition, le corpus se définit surtout par sa représentativité. La spécialisation et la possibilité du retour au contexte sont deux particularités des petits corpus. Enfin, il ne faut pas négliger l'impact des contraintes méthodologiques et institutionnelles lors de la constitution d'un corpus.

1.1. Taille du corpus

12 Les corpus sont souvent définis par leur taille, c'est-à-dire le volume de mots ou de phrases qu'ils contiennent. Cela amène à penser que le nombre de mots est un gage de qualité du corpus, alors que d'autres critères, comme la représentativité, sont en jeu pour déterminer la qualité d'un corpus. Certes, une indication de taille permet de donner un ordre d'idées quant à la nature du corpus. Face à l'injonction de création de corpus de plus en plus grands (cf. Sinclair 1991 *supra*), les grands corpus d'hier sont de petits corpus à comparaison des grands corpus d'aujourd'hui. Ainsi, les qualificatifs, « petit corpus » ou « grand corpus », ont peu de sens, puisqu'ils sont relatifs. Mais sans chercher à donner une claire idée de la taille du corpus, ces qualificatifs peuvent être réinvestis pour signifier l'intention sous-jacente aux corpus constitués. Un grand corpus est la recherche d'une maximalité, nécessaire pour étudier la langue comme un tout, pour étudier des phénomènes statistiques qui ne deviennent pertinents qu'avec de grands nombres ; typiquement, il s'agit d'études lexicographiques, sur les collocations, ou la phraséologie. À l'inverse, un petit corpus ne témoigne pas d'une quête de quantité, mais s'oriente vers le qualitatif : il s'agit de circonscrire un pan donné de la langue, ce qui lui confère un statut spécialisé. Ainsi, se dessine une ligne de fracture entre les types de corpus ; la taille n'est qu'une conséquence de l'objectif de recherche qui préside à la constitution du corpus : la dichotomie petit/grand peut être remplacée par une dichotomie général/spécialisé. En d'autres termes, construire un petit corpus revient à construire un corpus spécialisé et donc à circonscrire le domaine linguistique qui servira de champ d'investigation.

13 Même s'il y a consensus sur le caractère spécialisé[6] des petits corpus, certains auteurs ont tenté d'en donner une quantification afin de les situer par rapport aux grands corpus dont la quantification ne cesse d'augmenter.

> *According to O'Keeffe et al. (2007 : 4), any spoken corpus containing over a million words of speech is considered large, whereas with written corpora anything under five million words of text is quite small. But many small corpora, even written ones, are a great deal smaller than that, and Flowerdew (2004 : 19) notes that there is general agreement that small corpora contain up to 250,000 words.*
> (Koester 2010 : 67)

14 Bien sûr, aucun consensus quant à la taille moyenne des petits corpus n'est possible ou atteint, puisque du quart de million à cinq millions, il y a une large fourchette. Cette approximation se retrouve également lorsqu'il s'agit de définir la taille idéale des grands corpus et le problème est d'autant plus épineux pour les grands corpus que leur taille ne cesse de croître au cours des décennies. Fondamentalement, l'approximation qui entoure la définition de la taille des corpus s'explique par les différents objectifs de recherche auxquels ils obéissent. Cela vaut aussi pour les petits corpus : la taille du corpus peut varier en fonction de la spécialisation de la recherche menée. Ainsi, la question de la taille demeure toute relative quel que soit le type de corpus ; ce qui importe davantage est l'objectif de recherche auquel il est subordonné. Mais la question de la taille du corpus amène deux points :

> *the question of size is resolved by two factors : representativeness (have I collected enough texts (words) to accurately represent the type of language under investigation ?) and practicality (time constraints).*
> (Reppen 2010 : 32)

1.2. Représentativité et spécialisation

1.2.1. Discussion de la notion de représentativité

15 Cruciale pour la constitution d'un corpus échantillonné, la représentativité garantit sa validité scientifique. Un corpus représentatif doit rendre compte d'une large variabilité au sein d'une population (Biber 1993), cette variabilité étant d'ordre linguistique ou situationnel (*ibid.*) se manifeste notamment par les genres du discours ; mais si on se focalise sur un seul genre, cela n'exclut pas qu'il y ait variabilité à l'intérieur de ce genre[7]. Cela rend la phase d'échantillonnage du corpus complexe et déterminante, mais montre également l'illusion que la quête d'exhaustivité représente. Afin de pallier ce type de problème, la représentativité peut se transformer en typicalité :

> *There are, of course, practical limitations to sampling, and it will never be possible, particularly for a small corpus, to collect samples from all the situations in which a fairly widespread genre is used. What is important is to ensure that the samples are collected from a range of fairly typical situations.*
> (Koester 2010 : 69)

16 Au-delà de la complexité pratique de la notion de représentativité, c'est sa fragilité qui transparaît, car il est impossible d'évaluer la représentativité d'un corpus de manière totalement objective (Tognini Bonelli 2001 : 57). D'ailleurs, certains auteurs tendent à rejeter cette notion comme non opérante :

> *In reality, there are so many variables that the notion of 'representativeness' can almost be seen as a 'non-concept'. Kennedy notes that 'it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even of a particular genre or subject field or topic' (Kennedy 1998 : 62). Any attempt at corpus creation is therefore a compromise between the hoped for and the achievable.*
> (Nelson 2010 : 60)

17 Si la notion d'exhaustivité n'est pas tenable et est remplacée par la représentativité, cette dernière montre également ses limites. Cela est d'autant plus vrai pour les petits corpus qui ne recherchent pas la représentativité du point de vue quantitatif comme les grands corpus. Les petits corpus sont tournés vers une représentativité d'ordre qualitatif, assise sur la typicalité et la spécialisation.

1.2.2. Spécialisation du petit corpus

18 En comparaison des grands corpus, on peut s'interroger sur la représentativité d'un corpus de petite taille : en quoi peut-il constituer un bon échantillon d'un phénomène linguistique au regard des corpus toujours plus grands ?

19 Pour les petits corpus, l'échantillonnage est d'autant plus important que le corpus est de taille réduite. À l'inverse des grands corpus, cet échantillonnage est plus aisé à mettre en place dans la mesure où il est circonscrit à un domaine réduit des possibles du discours. Une fois déterminé le champ d'investigation – par exemple, la critique journalistique à propos du dernier prix Goncourt – il est possible de collecter des articles de presse s'y référant et d'opérer un choix parmi les textes les plus représentatifs du genre. La spécialisation du corpus laisse donc peu de latitude dans le choix des textes composant le corpus. D'ailleurs, plus le domaine d'investigation sera restreint, plus le corpus sera spécialisé et donc pourra être représentatif[8] ; cela permet

de tendre vers l'idéal de l'exhaustivité. Flowerdew (2004 : 21) propose des critères pour établir la spécialisation d'un corpus :

- *Specific purpose for compilation, e.g. to investigate a particular grammatical or lexical item.*
- *Contextualisation : particular setting, participants and communicative purpose.*
- *Genre, e.g. promotional (grant proposals, sales letters).*
- *Type of text/discourse, e.g. biology textbooks, casual conversation.*
- *Subject matter/topic, e.g. economics.*
- *Variety of English, e.g. Learner English.*

20 La restriction du domaine d'investigation et donc la spécialisation s'établissent en neutralisant un certain nombre de paramètres textuels (genres, type de discours ou de textes, etc.). Ces critères spécialisent le corpus et le rendent homogène, l'homogénéité venant compléter la représentativité du corpus. Un petit corpus spécialisé permet d'étudier un phénomène linguistique donné dans des conditions très précises en raison de la neutralisation des paramètres textuels. Cette recherche ciblée a l'avantage d'éviter les biais dus à l'hétérogénéité du corpus. Mais son inconvénient est de n'être valable que pour ce corpus, aussi faut-il pouvoir la déployer ensuite sur d'autres corpus, plus larges, moins homogènes et ainsi tester la validité de l'hypothèse dans d'autres conditions.

1.3. Retour au contexte

21 Si les grands corpus ont le plus souvent des visées statistiques et lexicographiques, les petits corpus permettent d'observer des phénomènes plus rares et plus précis. Ils permettent d'entrer dans le détail de la langue et de son contexte, sans rester dans une approche superficielle à la manière des grands corpus.

> *Where very large corpora, through their de-contextualisation, give insights into lexico-grammatical patterns in the language as a whole, smaller specialised corpora give insights into patterns of language use in particular settings.*
> (Koester 2010 : 67)

22 Les petits corpus se distinguent donc par leur maniabilité, puisqu'ils rendent aisé le retour au contexte.

> *Small, carefully targeted corpora (by which we commonly mean corpora of fewer than a million words of running text) have proved to be a powerful tool for the investigation of special uses of language, where the linguist can 'drill down' into the data in immense detail using a full armoury of software and shed light on particular uses of language.*
> (McCarthy & O'Keeffe 2010 : 6)

23 La constitution d'un petit corpus étant plus aisée à mettre en place, le chercheur qui constitue le corpus est aussi bien souvent l'observateur qui interprète les données. Aussi, un retour au contexte est un outil indispensable de l'interprétation des données.

> *With a small corpus, the corpus compiler is often also the analyst, and therefore usually has a high degree of familiarity with the context. This means that the quantitative findings revealed by corpus analysis can be balanced and complemented with qualitative findings (Flowerdew 2004 ; O'Keeffe 2007). As we shall see, specialised corpora are also usually carefully targeted and set up to reflect contextual features, such as information about the setting, the participants and the purpose of communication.*
> (Koester 2010 : 67)

24 Selon O'Keeffe (2007), les patterns révélés dans le corpus peuvent être reliés et expliqués par le contexte de la manière suivante :

> *The patterns can first of all be linked to a particular context, because the corpus analysis shows that they are concentrated within that context. We can see that these patterns are localised, as they are traceable to local situational conditions, such as gender, power or discourse goal. As a result, the patterns can be linked to pragmatically specialised uses within that particular context of situation.*
> (Koester 2010 : 74)

25 Selon, Flowerdew (2008), le corpus peut se montrer pertinent de deux manières pour l'interprétation des données :

> *1) The context can inform the corpus-based analysis, for example when the compiler-cum-analyst of a small specialised corpus has access to background information to aid in the interpretation of the data.*
> *2) The linguistic patterns identified through corpus analysis can tell us something about the social and cultural context from which the data were taken.*
> (Koester 2010 : 74)

26 Le retour au contexte renforce la part qualitative des petits corpus : l'interprétation des données n'est plus désincarnée car quantitative comme pour les grands corpus. Au contraire, le chercheur étant à la fois créateur et observateur, il a une connaissance approfondie du corpus qui lui permet une investigation fine avec des buts précis, contrairement à la navigation parfois aveugle dont sont l'objet les grands corpus.

1.4. Contraintes institutionnelles et méthodologiques

27 Les aspects pratiques de la recherche influent également sur la constitution du corpus. La recherche est une activité socialement ancrée dans des institutions. Dans le meilleur des cas, une recherche doctorale peut se trouver financée mais ce pour une durée déterminée, ce qui impose des contraintes au déploiement de cette recherche. Dans le cadre d'un contrat doctoral, le financement étant d'une durée de trois ans, durée qui correspond à celle d'un doctorat selon les directives européennes, il faut donc présenter un projet de recherche qui puisse être poursuivi et achevé dans le temps imparti par les contraintes institutionnelles. Un petit corpus semble donc un choix judicieux face à ces impératifs, lorsque l'on souhaite mener une recherche sur corpus qui ne soit pas complétement automatisée et qui suppose une connaissance approfondie des textes du corpus et rende possible l'examen fouillé du corpus.

28 Par ailleurs, la recherche est aussi l'occasion d'émettre des hypothèses, qu'il faut confronter à la réalité de la langue. Il est nécessaire de tester l'hypothèse formulée sur un premier échantillon de données afin de l'ajuster selon un cheminement inductif. Intervient donc le corpus de test, corpus nécessairement restreint. Si l'hypothèse et la méthode d'analyse qui l'accompagne donnent des résultats probants sur le corpus de test, alors il est possible de les appliquer à un corpus de travail. Dans le cas inverse, soit elles doivent être révisées, soit il est nécessaire de modifier le protocole expérimental, qu'il s'agisse de la méthode d'analyse ou le corpus lui-même. En bref, du point de vue méthodologique, le petit corpus constitue une bonne entrée en matière. D'ailleurs, ce serait une perte de temps considérable de tester une hypothèse sur un large corpus, en sachant que l'établissement de celui-ci prend du temps et que l'on n'a aucune certitude quant à la pertinence des résultats obtenus. Le développement de la recherche par projet, assujettie à des durées et des financements déterminés, a un impact sur la méthodologie de recherche : un petit corpus semble plus efficace pour tester un protocole expérimental.

2. Étude de la synonymie en corpus

29 Afin de justifier pleinement la constitution du corpus qui est présenté en troisième partie, un détour par la perspective théorique et méthodologique qui le fonde s'impose. Ce corpus ayant pour objectif de participer à un renouvellement de l'étude de la synonymie, il est nécessaire de préciser le contexte de recherche qui entoure la notion de synonymie et les ouvertures proposées par ce renouvellement.

30 En tant que relation sémantique, la synonymie est étudiée en contexte, condition nécessaire pour faire émerger le sens des unités lexicales. Les études descriptives qui se consacrent à cette notion établissent des corpus et les explorent pour étudier le sens d'une ou deux unités lexicales et tester en contexte les synonymes de cette unité. Or ces études conduites selon une approche différentialiste de la synonymie en viennent le plus souvent à rejeter les relations de synonymie entre les unités lexicales étudiées (Honeste 2007). En effet, un item A considéré comme le synonyme d'un item B voit sa relation entérinée en langue par des dictionnaires de synonymes. Ces études cherchent donc à confronter ces relations en langue au discours en examinant en contexte le sens des items en jeu, ces études aboutissant donc à rejeter la relation de synonymie. La recherche menée cherche à se départir de cette optique différentialiste pour recentrer la synonymie sur les équivalences de sens, c'est-à-dire sur la communauté de sens qui relie des synonymes (Doualan 2015). Cette nouvelle approche suppose de renouveler la méthode d'analyse de la synonymie et recourt à un corpus construit selon un tout autre objectif de recherche.

2.1. Contexte de la recherche sur la synonymie

31 Afin de situer la nouvelle approche de la synonymie, il est nécessaire de revenir sur le différentialisme qui gouverne actuellement les études sur la synonymie et sur sa méthodologie qui laisse entrevoir les critiques pouvant lui être adressées.

2.1.1. Le différentialisme

32 Appliqué à l'étude du sens, le différentialisme entend rechercher les différences de sens entre les unités linguistiques en s'appuyant sur le principe d'économie formulé de la manière suivante : à toute différence de forme peut être attribuée une différence de sens. Ce principe d'économie linguistique amène à rejeter la notion de synonymie, qui, par définition, y contrevient, puisqu'elle suppose une redondance sémantique : des items linguistiques de forme différente pourraient avoir des sens identiques. Si le différentialisme n'a pas éradiqué la synonymie, il a converti cette notion centrée sur la communauté de sens et les équivalences sémantiques, en une synonymie distinctive, c'est-à-dire, une synonymie qui met en avant les différences de sens entre les items lexicaux.

33 Le différentialisme a très tôt pris part à l'étude de la synonymie, puisqu'il est la position théorique adoptée par les synonymistes, les premiers auteurs de dictionnaires de synonymes aux XVIII^{e} et XIX^{e} siècles. Pour ces auteurs, il s'agit d'étudier le sens des mots synonymes en le décomposant en idées principales et accessoires, les idées principales permettant de rassembler plusieurs synonymes et les idées accessoires permettant de

les différencier (Adamo 1999 ; Aruta Stampacchia 2006). Si cette approche ne peut encore être qualifiée de scientifique, elle inspire Saussure pour sa théorie de la valeur (Auroux 1985 : 295). Il reprend le terme *valeur*, déjà courant dans l'étude de la langue, et contribue à en modifier l'acception, modifications qui peuvent se résumer en trois points :

> i) la valeur est la véritable réalité des éléments linguistiques ; ii) la valeur est déterminée par la position du terme dans le système (donc par des différences) ; iii) rien ne préexiste à la détermination de la valeur par le système.
> (Auroux 1985 : 295)

34 Saussure distingue la signification de la valeur des mots : la première met en lien un mot et une idée, l'idée pouvant être échangée contre le mot, et la seconde compare les mots entre eux, ce qui suppose de faire ressortir leurs différences. Selon la théorie de la valeur, les mots se définissent en creux, les uns par rapport aux autres, étant donné la vision systémique de la langue propre à Saussure. Cette approche du sens renvoie clairement à l'étude de la synonymie distinctive. D'ailleurs, Saussure prend l'exemple de plusieurs synonymes pour définir et exemplifier la théorie de la valeur :

> Dans l'intérieur d'une même langue, tous les mots qui expriment des idées voisines se limitent respectivement : des synonymes comme redouter, craindre, avoir peur n'ont de valeur propre que par leur opposition ; si redouter n'existait pas, tout son contenu irait à ses concurrents.
> (Saussure 1916/1986 : 160)

35 Les synonymes sont donc nécessairement des mots de valeurs différentes. Bien que Saussure et les synonymistes s'opposent quant à la synonymie, « la théorie classique de la synonymie, non seulement est intégrée à la théorie linguistique saussurienne, mais encore [qu'] elle en est probablement un élément générateur » (Auroux 1984 : 105).

36 Du point de vue théorique, le différentialisme s'appuie sur le principe d'économie linguistique, qui relève du bon sens linguistique, puisqu'il avance l'idée selon laquelle la langue ne peut s'embarrasser de mots redondants. Ce principe est aussi le garant des limites de la cognition humaine : comment un cerveau humain pourrait-il emmagasiner une grande quantité d'unités lexicales, redondantes qui plus est ? Pris dans sa version extrême, il revient à dire que toute modification de forme entraîne une modification de sens, autrement dit, la synonymie contrevient nécessairement à ce principe.

37 Concernant l'étude de la synonymie, l'approche différentialiste a été adoptée par les sémanticiens et les lexicologues, et continue d'avoir cours aujourd'hui dans les études lexicales. Cela suppose de rechercher les différences de sens entre les synonymes et s'apparente à une démarche sémasiologique. Si toute étude de la synonymie commence par être onomasiologique – il faut rassembler plusieurs mots ayant un signifié commun – elle se meut en démarche sémasiologique, puisqu'elle consiste à détailler les sens de ces items pour en cerner les différences. De ce fait, l'approche différentialiste ne s'intéresse qu'aux signifiés des synonymes et omet le plan de l'expression au profit du plan du contenu (García-Hernández 1997), ce qui l'amène à rejeter le plus souvent les relations de synonymie entre les items étudiés.

2.1.2. Méthodologie : le test de la substitution

38 Si, historiquement, le différentialisme a pour méthode d'analyse la décomposition sémantique des synonymes, il s'est pourvu d'une méthode en contexte : le test de la substitution. Ce test consiste à faire commuter deux items supposés synonymes dans un

contexte afin de confronter leurs sens. Il a pour prérequis l'identité de catégorie grammaticale des items substitués : s'il est nécessaire de modifier le contexte pour effectuer la substitution, alors il y aura nécessairement changement de sens, ce qui biaisera le test. Par ailleurs, même s'il suppose la constitution d'un corpus d'énoncés[9], le test est effectué dans un contexte phrastique et ainsi, ne prend guère en compte la dimension textuelle de la langue qu'un corpus permettrait d'exploiter. Incidemment, le différentialisme s'en tient à une conception phrastique du sens et laisse de côté les phénomènes propres à la textualisation, qui ont pourtant un impact sur le choix lexical et ne laissent pas indifférent un test comme la substitution. Pour finir, ce test suppose une conception paradigmatique de la synonymie : les items substitués doivent appartenir au même paradigme. Mais cela pose la question de la constitution du paradigme : le plus souvent, il s'agit de s'appuyer sur les dictionnaires de synonymes. Les synonymes sont d'abord considérés en langue pour ensuite être confrontés au discours via le test, ce qui constitue déjà un biais étant donné le déphasage entre les relations répertoriées en langue et celles qui apparaissent en discours.

39 Concrètement, si le test conduit à un changement de sens, à une agrammaticalité ou à une incorrection dans le contexte, on conclut à l'absence de synonymie entre les items substitués dans ce contexte donné. Pour illustrer ce test, voici un exemple de substitution mettant en jeu les noms *part* et *portion* (Stein-Zintz 2009) :

(1) On m'attribue une part de responsabilité dans la situation actuelle de la Corse.
(1') On m'attribue une portion de responsabilité dans la situation actuelle de la Corse.

40 Il est nécessaire d'étudier un grand nombre de contextes avant de statuer sur la synonymie entre deux items. Cette relation est mise en doute dès lors qu'il existe un contexte dans lequel la substitution n'est pas possible. Ainsi, le but du test est essentiellement de montrer les différences de sens entre les synonymes ; aussi sert-il expressément le différentialisme, à défaut de décrire la synonymie pour ce qu'elle est.

2.1.3. Critique du différentialisme

41 Même si le différentialisme est largement dominant dans les études sur la synonymie, il demeure critiquable à plusieurs égards. Du point de vue théorique, une conception exclusivement négative du sens n'est pas tenable : il n'est pas possible de considérer le sens des unités lexicales en creux sans les avoir rapprochées au préalable en raison de leurs significations proches. Il faut pouvoir situer les mots dans le système avant d'examiner leurs différences de sens, d'où la nécessité d'allier signification et valeur, ce qui réinvestit la conception positive du sens au cœur de l'étude de la synonymie. D'ailleurs, la conception négative du sens s'en tient au seul plan du signifié et omet le plan du signifiant pourtant essentiel dans la définition de la synonymie. À l'inverse, la signification met en relation des idées – des signifiés – avec des items lexicaux, c'est-à-dire des signes et donc leurs signifiants. Cela mène à une critique onomasiologique du différentialisme qui tend à ne considérer que le plan du contenu, écartant celui de l'expression. Or la synonymie est avant tout une relation onomasiologique, puisqu'elle met en relation des unités de formes différentes mais de sens proche. Ainsi, le signifiant a toute son importance dans la définition de la notion. La critique onomasiologique consiste à détacher la synonymie de la conception sémasiologique du sens propre au différentialisme pour en faire une équivalence sémantique approchée, ce qui suppose de ne plus se focaliser sur les différences de sens.

42 Le différentialisme témoigne également d'une faille épistémologique en rejetant la synonymie. En effet, il prend cette notion comme point de départ de ses études mais s'attache à la déconstruire. Le différentialisme part de la synonymie en langue – celle qui est répertoriée dans les dictionnaires – pour montrer qu'elle ne tient pas face au discours et il détaille l'impact que le contexte peut avoir sur cette notion. Or la synonymie en langue n'est qu'une abstraction construite à partir du consensus entre locuteurs au niveau de l'interdiscours[10]. Par ailleurs, le différentialisme postule qu'il y a des différences de sens lorsqu'il y a des différences de forme, puis il cherche à montrer les différences de sens entre les synonymes, en s'appuyant notamment sur le test de la substitution. En posant d'emblée l'impossibilité de la synonymie au travers du principe d'économie de la langue, le différentialisme se heurte à une pétition de principe. Cependant, il arrive que le test n'échoue pas, ce qui contrevient au principe d'économie et suscite la réserve, voire le désarroi, des linguistes. Ces deux exemples issus de Sikora (2009) étudiant *venir* et *arriver* montrent la réussite et l'échec du test en fonction des types de contexte :

43 - Dissimilateur fort

(2) Je te donne mon disque mais tu viens à mon concert.
(2') ? Je te donne mon disque mais tu arrives à mon concert.

44 - Assimilateur fort

(3) Franchement, quand je viens à Lyon, je suis surpris par l'absence de bateaux-mouches entre Rhône et Saône pour visiter le centre de Lyon jusqu'à l'Île Barbe.
(3') Franchement, quand j'arrive à Lyon, je suis surpris par l'absence de bateaux-mouches entre Rhône et Saône pour visiter le centre de Lyon jusqu'à l'Île Barbe.

45 La synonymie semble donc possible dans les rares cas où le test n'échoue pas, mais elle reste une notion avec un faible rendement explicatif et donc une faible valeur heuristique. Aussi peut-on se demander s'il y a intérêt à conserver cette notion. Pourtant, le différentialisme continue d'y recourir, et les études lexicales continuent de discréditer la synonymie telle qu'elle est répertoriée en lexicographie en montrant le déphasage qui existe avec la réalité du discours. Même si le différentialisme exploite à juste titre le déphasage entre langue et discours, il n'explique en rien le fonctionnement de la synonymie, qui est foncièrement une équivalence sémantique approchée, d'où la nécessité d'une nouvelle approche pour sortir des paradoxes du différentialisme.

2.2. Nouvelle approche de la synonymie

2.2.1. Une synonymie onomasiologique et syntagmatique

46 La synonymie étant par définition une notion onomasiologique, il est préférable d'adopter cette démarche pour l'étudier. La démarche onomasiologique consiste à partir d'un concept ou d'un faisceau de signifiés pour rechercher les signifiants qui instancient ce concept. Le plan de l'expression et le plan du contenu sont tous deux pris en compte, ce qui équilibre la recherche menée. Pour renouveler l'étude de la synonymie à l'aide de cette perspective, les équivalences de sens entre les synonymes sont mises en avant, au travers de la communauté de sens et les signifiants sont réhabilités (García-Hernández 1997). Si, à première vue, la perspective onomasiologique peut paraître grossière, puisqu'elle semble faire des amalgames entre les unités lexicales en raison de leur communauté de sens, elle cherche à dépasser les

postulats du différentialisme qui bloquent toute étude de la synonymie, en rejetant d'emblée cette notion au nom du principe d'économie. Cette perspective a pour objectif de revenir au fonctionnement même de la synonymie, en tant que notion fondée sur un noyau de sens commun[11].

47 Le différentialisme se caractérise par sa méthodologie appuyée sur le test de la substitution. Or ce test suppose de prendre la phrase, voire la proposition comme unité de base de la langue, d'où une conception réduite de l'élaboration du sens en contexte. Ainsi, l'approche adoptée ne considère plus la phrase mais le texte comme unité de base de la langue. Il n'est donc plus possible de recourir au test de la substitution dans la mesure où celui-ci échouerait systématiquement, les contraintes textuelles venant s'ajouter aux contraintes phrastiques. Par ailleurs, ce test suppose implicitement de rechercher des différences de sens. Ainsi, afin d'étudier la synonymie à l'échelle textuelle et selon une optique onomasiologique, il faut adopter une conception syntagmatique de la synonymie, qui permet de rompre avec la conception paradigmatique propre au test de la substitution. Dans une approche syntagmatique, les synonymes sont à rechercher dans la linéarité du texte, c'est-à-dire, selon leur répartition dans les textes. Bien sûr, cela suppose d'avoir une conception assez large du terme *synonyme* : il faut pouvoir rassembler assez d'unités lexicales pour en trouver réparties à l'échelle du texte. Cela s'apparente à l'étude des isotopies propres à un concept ou un faisceau de signifiés, mais en y ajoutant l'aspect syntagmatique, qui suppose un ancrage lexical dans le texte, il s'agit en réalité d'examiner les connexions entre les instanciations d'un même signifié, c'est-à-dire, à étudier les réseaux que forment les unités lexicales à l'intérieur du texte. Ainsi, faut-il partir d'un noyau de sens et examiner ses instanciations lexicales et la manière dont elles se déploient.

2.2.2. L'analyse thématique

48 Pour mettre en place une méthode correspondant aux exigences de cette analyse de la synonymie, le choix a été fait de s'appuyer sur l'analyse thématique telle qu'elle a été définie par Rastier (1987). Certes, cette analyse n'a pas vocation à étudier la synonymie, mais l'optique qu'elle adopte sur le sens est résolument onomasiologique et inscrit clairement dans une approche textuelle du sens. L'analyse thématique consiste en la description des thèmes abordés dans un texte et l'étude des instanciations lexicales de ces thèmes. Ainsi, les deux plans de l'expression et du contenu sont pris en compte à parts égales, ce qui a son importance pour l'approche onomasiologique.

49 Le thème se définit comme « une structure stable de traits sémantiques (ou *sèmes*), récurrente dans un corpus, et susceptible de lexicalisations diverses » (Rastier 2001 : 197). Autrement dit, il est composé d'un signifié ou d'un ensemble de sèmes. Il est à rechercher en discours, d'où une analyse nécessairement ancrée dans les textes. De ce fait, l'analyse thématique suppose de constituer un corpus et exploite celui-ci dans sa dimension textuelle. Enfin, il se manifeste selon divers signifiants, que l'on nomme lexicalisations. Plusieurs types de signifiants sont possibles puisque « le thème est représenté par une séquence linguistique (une phrase, un groupe nominal, un nom propre ou commun) » (Erlich 1995 : 85). Mais le thème reste une construction ; il est ce que le chercheur se donne. La définition d'un thème en vue d'une analyse thématique est donc tributaire de trois paramètres : la définition des traits sémantiques qui le composent, le choix des lexicalisations qui seront étudiées comme instanciations de ce thème et enfin le corpus dans lequel l'étude sera menée. Le choix des lexicalisations ne

peut se faire qu'une fois le thème clairement défini. Quant au corpus, il peut être choisi en lien étroit avec le thème, mais cela n'est pas une condition sine qua non. À titre d'exemple, Bourion (1995) étudie le thème de la peur sur un large corpus littéraire constitué de romans sans restreindre son choix à des romans qui évoquent expressément cette thématique.

50 Les lexicalisations, en tant qu'instanciations lexicales du thème, comportent tout ou partie des sèmes constituant ce thème. Cela dépend bien sûr de la complexité sémique du thème et de la quantité de lexicalisations en jeu. L'inventaire des lexicalisations ne pouvant être exhaustif, il est nécessaire de se fixer des limites quant au choix des lexicalisations. Ce choix peut s'effectuer *a priori* c'est-à-dire à l'aide de ressources lexicales préexistantes, telles que des dictionnaires, ou il peut s'effectuer par une fouille lexicale du corpus, les deux méthodes pouvant bien sûr être combinées. Du point de vue formel, les lexicalisations sont souvent des substantifs, mais les autres catégories grammaticales fournissent également des lexicalisations pertinentes du thème. Cela suppose donc de recourir à la famille dérivationnelle de la lexicalisation privilégiée du thème. Les synonymes sont également sollicités lors de l'inventaire des lexicalisations, puisqu'ils permettent d'augmenter la couverture lexicale du thème. Ainsi, la synonymie n'est pas étrangère à la mise en place d'une analyse thématique. De ce fait, cette méthode s'avère pertinente pour une analyse de la synonymie selon l'approche onomasiologique et syntagmatique.

2.3. Contraintes méthodologiques pour la constitution du corpus

51 Même si l'analyse thématique semble convenir pour l'approche de la synonymie adoptée ici, il est nécessaire d'opérer quelques modifications dans la méthodologie et de justifier certains choix dans la mise en place du protocole expérimental. L'analyse thématique s'intéresse en priorité aux thèmes qui structurent des textes, les lexicalisations n'étant que des manifestations de ces thèmes. Or pour une étude de la synonymie, l'accent est mis sur les lexicalisations[12] et les réseaux lexicaux qu'elles forment à l'intérieur des textes, le thème n'est finalement qu'un prétexte pour fédérer des unités lexicales autour d'un signifié commun.

52 Il est nécessaire de choisir un thème sémantiquement large, dont le signifié est composé de sèmes génériques afin que ce thème donne lieu à de nombreuses lexicalisations. Il y a peu d'intérêt à mener une étude onomasiologique sur un petit nombre de lexicalisations, car cela laisse peu de chances de rencontrer ces lexicalisations en contexte et ainsi de pouvoir étudier leur répartition et les réseaux qu'elles forment. Autrement dit, le choix du thème est conditionné par l'impératif onomasiologique du nombre de lexicalisations. En ce qui concerne le corpus, celui-ci doit être choisi en lien étroit avec le thème. En effet, pour s'assurer de la présence des lexicalisations du thème dans le corpus, il doit manifester une certaine saillance du thème. Un corpus dans lequel le thème ne serait pas prégnant donnerait lieu à peu d'instanciations du thème, or ceci est un prérequis nécessaire à la méthodologie onomasiologique. L'étude des lexicalisations fait peser des contraintes sur le choix du thème et du corpus et oblige à un choix interconnecté de ceux-ci, ce qui n'apparaissait pas comme indispensable en analyse thématique.

53 Enfin, puisqu'il s'agit d'une étude exploratoire voulant confronter l'hypothèse d'une répartition des synonymes en réseaux lexicaux à la réalité linguistique, le choix d'un

petit corpus semble le plus adéquat. Cela fait donc de ce corpus un corpus de test, et en tant que tel, il est nécessairement de taille réduite. Par ailleurs, un petit corpus permet de canaliser le grand nombre de lexicalisations qu'un thème saillant en corpus sera susceptible d'engendrer. Et par là même, cela entraînera un nombre raisonnable de réseaux lexicaux à examiner. Un grand corpus donnerait une abondance de lexicalisations qui ne sauraient coïncider avec le caractère exploratoire de la recherche menée et avec les contraintes institutionnelles du doctorat[13]. Il est donc nécessaire d'adapter la taille du corpus pour compenser la saillance du thème dans les textes. Par contraste, Bourion (1995) fait le choix inverse : elle choisit le thème de la peur qui demeure assez restreint, puisqu'il donne lieu à peu de lexicalisations : huit lemmes de forme substantive auquel il faut ajouter les dérivés adjectivaux et verbaux. À l'inverse, la présente étude se focalise sur deux thèmes conjoints et généraux, ce qui donne lieu à une centaine de lexicalisations, accompagnées de leurs dérivés nominaux, adjectivaux et adverbiaux. Par ailleurs, Bourion (1995) fait le choix d'un large corpus – le fonds roman du corpus FRANTEXT pour la période 1830-1970 – qui ne manifeste pas la saillance du thème. La taille du corpus compense donc la faible saillance du thème et le petit nombre de lexicalisations qui l'instancient. Pour la présente étude, les proportions sont inversées : le petit corpus répond à un thème saillant et à un grand nombre de lexicalisations.

3. Constitution du corpus de test

3.1. Description du corpus

54 Les contraintes méthodologiques inhérentes à l'analyse de la synonymie onomasiologique et syntagmatique supposent de choisir conjointement le thème étudié et le corpus. Ainsi, le corpus est construit en lien étroit avec le thème, afin d'assurer la saillance du thème dans le corpus. Pour la présente recherche, deux thèmes associés ont été choisis : le vice et la vertu. Ces deux thèmes sont apparus comme inséparables bien qu'antagonistes. En effet, ils sont souvent traités ensemble afin de définir l'un par rapport à l'autre ou de les opposer plus aisément. Ainsi, du point de vue d'une étude onomasiologique s'apparentant à l'analyse thématique, il est nécessaire de traiter conjointement ces deux thèmes, leur intrication s'avérant particulièrement saisissante dans le corpus établi. Puisque les thèmes choisis sont des thèmes moraux, le corpus a été constitué à partir de textes traitant de questions de morale, d'où le choix des textes suivants :

- *Caractères*, J. de La Bruyère
- *Maximes et Réflexions diverses*, F. de La Rochefoucauld
- *Pensées*, B. Pascal

55 Ce corpus de trois textes de littérature française morale du XVII^e^ siècle ne dépasse pas les 300 000 mots, ce qui le place dans les limites évoquées par O'Keeffe *et al.* (2007) et Flowerdew (2004) pour définir la taille d'un petit corpus. Mais il est délicat de situer ce corpus au sein des petits corpus étant donné la latitude de ces limites.

56 Les textes du corpus ont été récupérés en version numérique, dans le but d'opérer un traitement semi-automatique sur le corpus et ainsi de faire émerger les réseaux lexicaux des thèmes du vice et de la vertu plus aisément. Les textes ont en premier lieu été récupérés au format brut sur le site du Projet Gutenberg[14]. Ils ont été convertis au

format XML car ce format a permis de les enrichir en métadonnées. Certes, pour la présente recherche, l'enrichissement s'est limité à la segmentation et au balisage en paragraphes. Le paragraphe s'est avéré un bon compromis pour saisir l'unité textuelle sur laquelle sont basés les textes du corpus, à savoir la forme brève[15] : en effet, nombreuses sont les formes brèves constituées d'un seul paragraphe. Les textes du corpus ayant été choisis pour leur forme textuelle particulière, il était nécessaire de pouvoir en rendre compte au travers du balisage. Ce balisage en paragraphes permet d'effectuer des requêtes au niveau de la forme brève[16] et ainsi d'y détecter les réseaux lexicaux instanciant les thèmes du vice et de la vertu. Par ailleurs, la conversion au format XML est indispensable pour l'interrogation du corpus. À cet effet, le langage XQuery permet, grâce à la fonction « thesaurus », d'effectuer des requêtes sur des documents textuels à l'aide d'un lexique structuré, autrement dit, les textes du corpus peuvent être interrogés à l'aide de la ressource lexicale qui répertorie les lexicalisations pour faire ressortir les réseaux lexicaux qui structurent les thèmes.

3.2. Spécialisation du corpus

57 La petite taille du corpus ne doit pas l'empêcher d'être représentatif. Cette représentativité est à rechercher dans la spécialisation : entrent en jeu la saillance thématique et l'homogénéité du corpus via la neutralisation des paramètres textuels.

3.2.1. Paramètre thématique

58 La connaissance préalable des textes laisse supposer une certaine saillance des thèmes moraux. La lecture attentive des textes et de la bibliographique critique qui entoure ces textes confirment cette saillance. Ces textes ont été écrits par des auteurs qualifiés de moralistes[17], comme La Rochefoucauld et La Bruyère (Lafond 1992), ou des philosophes (Tourrette 2008) s'intéressant à des questions de morale et de religion. La connaissance du contexte de production des textes informe sur les thèmes qui les traversent. Ce retour au contexte situationnel n'est possible qu'avec un petit corpus et est indispensable à une connaissance approfondie des textes, surtout avec un angle d'attaque thématique et donc sémantique.

59 Au-delà des lexicalisations, la définition des thèmes touche le corpus dans le choix des textes. En effet, avec des thèmes aussi généraux que le vice et la vertu, il est nécessaire d'avoir une connaissance précise et détaillée des différents signifiés dans lesquels les thèmes peuvent se décliner. Relevant de la morale, les thèmes du vice et de la vertu sont souvent considérés comme religieux ; pourtant, ces thèmes revêtent également une valeur plus sociale et psychologique, puisque les vices et les vertus font partie des traits de caractère de tout un chacun. Il a fallu choisir des textes pouvant couvrir ces deux aspects de manière équilibrée. La Bruyère traite des mœurs de ses contemporains dans une optique mondaine et donc sociale alors que Pascal est davantage tourné vers la religion. Quant à La Rochefoucauld, il se place dans l'optique mondaine à la manière de La Bruyère mais ses réflexions sont imprégnées de jansénisme, doctrine religieuse à laquelle Pascal est souvent rattaché. Le texte de La Rochefoucauld constitue donc un moyen terme entre l'approche religieuse et l'approche sociale des thèmes du vice et de la vertu, même si l'aspect religieux est loin d'être aussi prégnant chez La Rochefoucauld que chez Pascal. Ces trois auteurs s'avèrent donc complémentaires pour une étude des thèmes du vice et de la vertu.

60 Si la saillance des thèmes étudiés ne fait pas de doute, des statistiques lexicales ont permis de le confirmer en précisant quelles lexicalisations devaient être répertoriées pour chacun d'eux. Seule une fouille du corpus peut déterminer les lexicalisations à choisir pour procéder à une détection féconde des réseaux lexicaux. Si les lexicalisations sont inventoriées au préalable à l'aide de dictionnaires de synonymes et de thesaurus, la confrontation au corpus a permis d'éliminer les lexicalisations qui n'apparaissent pas ou peu fréquemment dans les textes ; en effet, ces lexicalisations ne sont pas aptes à rendre compte des thèmes dans le corpus. Cela n'exclut pas, en revanche, de conserver ces lexicalisations en vue de détecter des réseaux lexicaux dans d'autres corpus.

3.2.2. Paramètres textuels

61 Si le paramètre thématique est propre à l'approche adoptée, les paramètres textuels sont les paramètres récurrents à prendre en compte pour établir un corpus. L'échantillonnage que constitue le corpus ne prend son sens qu'en tant qu'il est représentatif d'un phénomène linguistique. Pour un petit corpus, la représentativité a pour condition la spécialisation du corpus, cette spécialisation passant par l'homogénéité du corpus concernant les paramètres textuels qui ont présidé à sa constitution. Ainsi, une grande attention a été portée aux caractéristiques des textes du corpus, afin de neutraliser au mieux leurs paramètres textuels. L'homogénéité du corpus repose sur cinq paramètres neutralisés :

62 - l'unité de l'état de langue : les textes du corpus datent tous de la seconde moitié du XVII^{e} siècle, ils appartiennent donc à l'état de langue du français classique. Ce critère permet d'assurer une certaine cohérence en ce qui concerne l'usage des unités lexicales : les unités lexicales sont susceptibles d'être employées dans les mêmes acceptions, puisque la neutralisation de l'état de langue permet d'éviter les cas de changement sémantique. Demeure malgré tout la variabilité inhérente à l'idiolecte de chaque auteur ;

63 - l'unité du discours : les textes du corpus appartiennent au discours constituant littéraire (Maingueneau & Cossutta 1995), cela permet d'assurer une certaine harmonie dans le traitement des thèmes, car un thème n'est pas traité de manière identique d'un discours constituant à l'autre. Seuls les discours philosophiques, religieux et littéraires sont susceptibles d'évoquer les thèmes du vice et de la vertu. À ce titre, la tradition place le texte de Pascal dans le discours constituant littéraire, mais la situation s'avère plus complexe, puisqu'on peut considérer qu'il se situe au carrefour de ces trois discours constituants. Pascal avait l'intention de rédiger une apologie de la religion chrétienne (Adam 1997) ; les *Pensées* en constituent les notes préparatoires. Par ailleurs, il aborde la question de la morale d'un point de vue plutôt philosophique. Ainsi, la connaissance du contexte de production des œuvres est indispensable lors du choix des textes afin d'y apporter les nuances nécessaires. Cela justifie d'ailleurs le penchant religieux pris par les thèmes du vice et de la vertu dans ce texte. Enfin, l'appartenance à un même discours constituant est un préalable nécessaire pour que les textes soient susceptibles d'appartenir au même genre ;

64 - l'unité du type de texte : les trois textes du corpus sont de type argumentatif ; un type de texte a vocation à caractériser le texte dans sa globalité, ce qui n'exclut pas la présence d'autres types de texte dans certains passages. À cet égard, on rencontre des

passages narratifs et descriptifs chez La Bruyère, puisque ses *Caractères* sont connus pour dresser les portraits caricaturaux des individus évoluant dans les cercles mondains, ces portraits ne sont pas seulement descriptifs mais prennent également un tour narratif, lorsque les personnages sont mis en scène pour accroître le réalisme de leurs portraits. De même pour le type de texte, la connaissance approfondie des textes permet de nuancer la description des textes ;

65 - l'unité de genre[18] : les trois textes peuvent être classés dans le genre essai, mais il est nécessaire d'opérer une restriction à l'intérieur de ce genre, puisque les textes appartiennent à un sous-genre particulier : les sentences. L'appartenance au même sous-genre permet de définir un corpus très homogène et très spécialisé. Toutefois, l'hétérogénéité transparaît dans les différentes manifestations de ce sous-genre en fonction des textes : chez La Bruyère, il est question de caractères, de maximes chez La Rochefoucauld et de pensées chez Pascal. Cette déclinaison du sous-genre de la sentence a trait au contenu des sentences aussi bien qu'à leur forme. Les caractères comportent nombre de portraits, c'est d'ailleurs ce qui a fait la notoriété du texte de La Bruyère ; les maximes sont aux épigraphes proférées dans les salons mondains et sont reconnues pour leur caractère ramassé et incisif, elles sont donc clairement dépendantes de leur contexte social de production : le salon mondain et les activités liées au beau discours qui s'y pratiquaient ; enfin, les pensées s'apparentent aux réflexions menées par un auteur en vue de la rédaction d'un ouvrage complet ;

66 - l'unité de la forme textuelle : le sous-genre de la sentence tire sa singularité de la forme qu'il prend, ce qui vient ajouter un paramètre supplémentaire quant à l'homogénéité du corpus. Les textes du corpus arborent une forme textuelle discontinue : ils sont composés de formes brèves (Lafond 1984) qui se déclinent selon les différentes manifestations du sous-genre de la sentence : maximes ou caractères. En revanche, il est question de fragments chez Pascal et non de formes brèves. La forme brève suppose un texte achevé, dont les sous-unités textuelles (maximes, caractères, etc.) ont été dûment établies. Or le texte de Pascal étant inachevé, il est constitué d'unités textuelles disparates et fractionnées qui n'avaient pas vocation à être publiées. Ainsi, les *Pensées* de Pascal se manifestent sous la forme d'une succession de fragments textuels de longueurs différentes. Malgré cette différence entre fragments et formes brèves, ils restent fédérés par la discontinuité du discours : en effet, les textes du corpus se caractérisent par une forme atomisée en sous-unités textuelles. En tant que manifestation du discours discontinu, les formes brèves se définissent par leur autonomie, leur concision et leur condensation linguistique (Van Delft 2006). Le fait que les fragments des *Pensées* soient en réalité des notes en vue de la rédaction d'un essai leur confère les mêmes caractéristiques. Certes, cette différence entre fragment et forme brève contribue à l'hétérogénéité du corpus, tout comme les différents types de formes brèves présents dans le corpus, mais dans l'ensemble, le corpus demeure homogène si l'on tient compte de tous les paramètres textuels.

67 Si les paramètres textuels tels que le type de texte ou le genre sont fréquemment mentionnés dans la constitution d'un corpus, il est moins souvent fait cas du paramètre de la forme. Les corpus établis à partir de textes de forme plus conventionnelle, comme des romans, des articles de presse, etc., ne pose pas la question de la forme à leurs concepteurs. Or le genre des textes du corpus est justement conditionné par leur forme. Mais pourquoi s'attacher à des textes de forme si spécifique ? Postuler une synonymie onomasiologique et syntagmatique constitue une hypothèse à tester sur un corpus.

L'étude des réseaux lexicaux formés par les mots de sens proche nécessite la constitution d'un corpus de textes. Il a fallu résoudre l'équation d'un petit corpus, autrement dit contenant peu de textes avec l'étude des synonymes à l'échelle textuelle. Le compromis a été trouvé grâce aux formes brèves, dont chacune constitue en elle-même un texte de par son autonomie. Cette condition formelle restreint drastiquement le choix des textes, auquel il faut ajouter les contraintes exercées par les autres paramètres textuels. Ces conditions réunies ont abouti à la constitution d'un petit corpus très homogène et très spécialisé, autant par sa thématique que par sa forme et son genre.

68 Pour finir, la représentativité du corpus passe par la spécialisation des textes qui le composent mais aussi par la typicalité de ceux-ci. Les textes choisis doivent être les plus typiques, c'est-à-dire les exemplaires les plus prototypiques, du discours, du genre, de la forme textuelle choisis. À ce titre, les textes du corpus sont typiques du sous-genre de la sentence et du discours discontinu. « La Rochefoucauld est universellement reconnu comme auteur de maximes » (Meleuc 1969 : 70). Cette typicalité provient de la notoriété dont jouissent ces textes dans la critique littéraire. L'ouvrage de La Rochefoucauld « n'est pas n'importe quelle suite d'énoncés, mais un livre, apparemment au même titre que, par exemple, *Les Caractères* de La Bruyère » (*ibid.*). Ainsi, la spécialisation du corpus ne dispense pas de choisir avec soin les textes les plus typiques pour constituer un petit corpus.

3.3. Contexte linguistique et situationnel

69 Le petit corpus se caractérise par le retour au contexte qu'il rend aisé. Le contexte situationnel intervient lors de la constitution du corpus afin de choisir les textes les plus représentatifs, mais aussi lors de l'interprétation des données. Ainsi, pour illustrer l'importance de la connaissance du contexte tant linguistique que situationnel dans l'interprétation des données, voici l'exemple des deux maximes les plus polémiques des *Maximes* de La Rochefoucauld :

> Les vertus ne sont, le plus souvent, que des vices déguisés. (épigraphe)
> Les vices entrent dans la composition des vertus comme les poisons entrent dans la composition des remèdes. [...] (Max. 182)

70 Ces maximes donnent à voir des réseaux lexicaux paradoxaux par les relations sémantiques qu'ils sous-tendent : les lexicalisations *vice* et *vertu* ne sont pas considérées comme antagonistes, au contraire, un rapprochement sémantique est opéré pour montrer que les thèmes et donc les notions qu'ils manifestent sont étroitement reliés. Ainsi, la relation d'antonymie établie en lexicographie n'a pas cours ici ; sans aller jusqu'à la synonymie, une certaine proximité sémantique s'instaure. L'explication de ces réseaux lexicaux pour le moins paradoxaux, qui brouillent les relations sémantiques, réside dans le contexte du texte. Le contexte situationnel rappelle la vision pessimiste de société propre à La Rochefoucauld (Adam 1997). Pour lui, ses contemporains sont mus par l'amour-propre et l'intérêt, si bien que leurs vertus ne sont jamais authentiques, mais le fruit de cet intérêt et de cet amour-propre ; aussi, sont-elles guidées par leurs vices. Le contexte linguistique du texte explique également ces réseaux reposant sur des relations sémantiques paradoxales. La forme aphoristique des maximes suppose une certaine concision de l'écriture et une condensation du sens : « la maxime-définition est surprenante et suffisante » (Beaujot 1984 : 98-99). Elle fait également intervenir des figures de l'expression – le paradoxe, l'antithèse, l'oxymore,

etc. – utilisées de la manière la plus obscure chez La Rochefoucauld (Van Delft 2006). Sans le contexte qui permet de les justifier, ces énoncés paradoxaux n'auraient pu être interprétés de façon satisfaisante. Ainsi, le petit corpus offre la possibilité de connaissance approfondie des textes qui rend possible l'interprétation des données paradoxales à première vue.

Conclusion

71 Si l'attention des sciences du langage, et plus particulièrement de la linguistique de corpus, pour les grands corpus demeure étant donné les possibilités offertes par les avancées technologiques, les petits corpus continuent d'exister et d'être utilisés :

> *Small corpora, it was held, can be very useful, providing they can offer a 'balanced' and 'representative' picture of a specific area of the language. This recognition of a need for smaller, more specialised corpora increased.*
> (Nelson 2010 : 55)

72 Les petits corpus trouvent leurs justifications dans les possibilités d'analyse qu'ils permettent comparativement aux grands corpus. Malgré leur petite taille, ils satisfont l'exigence de représentativité. Si elle passe par la typicalité des textes, elle est également rendue possible par la spécialisation du corpus. Circonscrire un champ d'investigation restreint permet de constituer un corpus à partir de quelques textes seulement. Cette spécialisation passe par la définition de critères précis concernant le contexte linguistique et situationnel des textes. Pour le présent corpus, les paramètres textuels et la saillance thématique constituent les prérequis de cette spécialisation, qui amène à établir un corpus homogène et permet d'obtenir des résultats cohérents quant aux phénomènes linguistiques étudiés. Le petit nombre des textes rend possible une connaissance approfondie de ceux-ci, notamment par la lecture des textes eux-mêmes ou de bibliographies les concernant. Cette connaissance du contexte est indispensable pour interpréter les données de manière adéquate.

73 Ainsi, les petits corpus trouvent tout à fait leur place dans le paysage de la recherche linguistique. Si le quantitatif est souvent mis en avant pour ce qui est des corpus, le qualitatif a toute sa place dans la définition d'un corpus. D'ailleurs, le petit corpus montre à quel point le qualitatif prime sur le quantitatif, tout comme le particulier prime sur le général. Le petit corpus est donc autant gage de scientificité que le grand corpus, sa représentativité se situant sur un autre plan. D'ailleurs, le petit corpus permet des études plus fines sur la langue et le discours car circonscrites à un champ d'investigation précis. Il suppose une pensée du discours et du genre, comme ayant une influence sur les phénomènes linguistiques, à l'inverse des études sur grands corpus qui semblent plus aveugles à cet égard et considèrent que la langue est la même quel que soit le discours ou le genre.

74 La constitution d'un petit corpus n'empêche pas à plus long terme d'établir un corpus plus large pour confronter les observations menées à partir du petit corpus à celles obtenues sur un plus grand corpus. Cela permet de discuter les choix effectués quant à la spécialisation du petit corpus : un corpus dont les paramètres textuels sont plus souples donne-t-il les mêmes résultats ? L'élargissement du corpus amène-t-il à nuancer l'hypothèse de départ ?

BIBLIOGRAPHIE

Références du corpus

La Bruyère J. (1696). *Les Caractères ou les mœurs de ce siècle*. Paris : E. Michallet.

La Rochefoucauld F. (1665). *Réflexions ou sentences et maximes morales*. Paris : Barbin.

Pascal B. (1998). « Pensées », *Œuvres complètes*, M. Le Guern (éd.). Paris : Gallimard.

Références bibliographiques

Adam A. (1997). *Histoire de la littérature française au XVII^e siècle*, t. 3. Paris : Albin Michel.

Adamo M. G. (1999). « Introduction », *La Justesse de la langue française ou les différentes significations des mots qui passent pour synonimes* (Abbé Gabriel Girard), texte établi, présenté et annoté par M.-G. Adamo. Fasano/Paris : Schena / Didier Erudition.

Aruta Stampacchia A. (2006). « Pierre-Benjamin Lafaye théoricien de la synonymie », *Introduction sur la théorie des synonymes* (P.-B. Lafaye), texte établi, présenté et annoté par A. Aruta Stampacchia. Fasano/Paris : Schena/Lanore.

Auroux S. (1984). « D'Alembert et les synonymistes », *Dix-Huitième siècle* 16 : 93-108.

Auroux S. (1985). « Deux hypothèses sur les sources de la conception saussurienne de la valeur linguistique », *Travaux de linguistique et de littérature* 23/1 : 295-299.

Auroux S. (1998). *La raison, le langage et les normes*. Paris : PUF.

Beaujot J.-P. (1984). « Le travail de la définition dans quelques maximes de La Rochefoucauld », *Les Formes brèves de la prose et le discours discontinu*, Études réunies et présentées par J. Lafond. Paris : Vrin, 98-99.

Biber D. (1993). « Representativeness in Corpus Design », *Literary and Linguistic Computing* 8/4 : 243-257.

Bourion E. (1995). « Le réseau associatif de la peur », *in* F. Rastier (dir.) *L'analyse thématique des données textuelles. L'exemple des sentiments*. Paris : Didier Erudition, 107-145.

Charaudeau P. (2009). « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique », *Corpus*, « Corpus de textes, textes en corpus », 8 : 37-66.

Doualan G. (2015). *Étude historique, épistémologique et descriptive de la synonymie*, Thèse de doctorat, Université Paris-Sorbonne.

Dalbera J.-P. (2002). « Le corpus entre données, analyse et théorie », *Corpus*, « Corpus et recherches linguistiques », 1.

Erlich D. (1995). « Une méthode d'analyse thématique. Exemples de l'ennui et de l'ambition », *in* F. Rastier (dir.) *L'analyse thématique des données textuelles. L'exemple des sentiments*. Paris : Didier Erudition, 85-103.

Flowerdew L. (2004). « The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings », *in* U. Connor et T. Upton (éd.) *Discourse in the Professions : Perspectives from Corpus Linguistics*. Amsterdam : John Benjamins, 11-33.

Flowerdew L. (2008). « Corpora and Context in Professional Writing », *in* V. K. Bhatia, J. Flowerdew et R. H. Jones (éd.) *Advances in Discourse Studies*. London : Routledge, 115-131.

García-Hernández B. (1997). « La sinonimia, relación onomasiológica en la antesala de la semántica », *Revista Española de Lingüística* 27/2 : 381-407.

Habert B. (2000). « Des corpus représentatifs : de quoi, pour quoi, comment ? », *in* M. Bilger (coord.) *Linguistique sur corpus. Études et Réflexions*, Cahiers de l'Université de Perpignan, 31. Perpignan : Presses Universitaires de Perpignan, 11-58.

Halliday M. A. K. et James Z. (1993). « A Quantitative Study of Polarity and Primary Tense in the English Finite Clause », *in* J. M. Sinclair, M. Hoey et J. Fox (éd.) *Techniques of Description : Spoken and Written Discourse*. London : Routledge.

Honeste M.-L. (2007). « Entre ressemblance et différence : synonymie et cognition », *Le Français moderne* 75/1 : 160-174.

Kennedy G. (1998). *An Introduction to Corpus Linguistics*. Harlow : Addison Wesley Longman.

Koester A. (2010). « Building small specialised corpora », *in* A. O'Keeffe et M. McCarthy (éd.) *Routledge Handbook of Corpus Linguistics*. New York : Routledge, 66-79.

Lafond J. (1984). *Les Formes brèves de la prose et le discours discontinu : XVI^e^-XVII^e^ siècles*. Paris : Vrin.

Lafond J. (dir.) (1992). *Moralistes du XVII^e^ siècle*. Paris : Laffont.

Maingueneau D. et Cossutta F. (1995). « L'analyse des discours constituants », *Langages* 117/29 : 112-125.

McCarthy M. et O'Keeffe A. (2010). « What are corpora and how have they evolved ? », *in* McCarthy M. et O'Keeffe A. (éd.) *The Routledge Handbook of Corpus Linguistics*. Oxford : Routledge, 3-13.

Meleuc S. (1969). « Structure de la maxime », *Langages* 13 : 69-99.

Mellet S. (2002). « Introduction », *Corpus*, « Corpus et recherches linguistiques », 1.

Nelson M. (2010). « Building a written corpus : What are the basics ? », *in* McCarthy M. et O'Keeffe A. (éd.) *The Routledge Handbook of Corpus Linguistics.* Oxford : Routledge, 53-65.

O'Keeffe A. (2007). « The Pragmatics of Corpus Linguistics », communication présentée à la *4^e^ Conférence de linguistique de corpus*, Université de Birmingham, Birmingham, juillet 2007.

O'Keeffe A., McCarthy M. et Carter R. (2007). *From Corpus to Classroom*. Cambridge : Cambridge University Press.

Rastier F. (1987). *Sémantique interprétative*. Paris : PUF.

Rastier F. (2001). *Arts et sciences du texte*. Paris : PUF.

Reppen R. (2010). « Building a corpus : what are the key considerations ? », *in* McCarthy M. et O'Keeffe A. (éd.) *The Routledge Handbook of Corpus Linguistics*. Oxford : Routledge, 31-37.

Saussure F. de (1916/1986). *Cours de linguistique générale*. Paris : Payot.

Sinclair J. (1991). *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

Sinclair J. (1996). *Preliminary recommendations on Corpus Typology*, Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards).

Sinclair J. (2001). « Preface », *in* M. Ghadessy, A. Henry et R. L. Roseberry (éd.) *Small Corpus Studies and ELT. Theory and practice*. Amsterdam/Philadelphia : John Benjamins Publishing company.

Stein-Zintz S. (2009). « La synonymie dans l'expression de la partition : le trio part, partie et portion » *Pratiques* 141/142 : 195-207.

Tognini Bonelli E. (2001). *Corpus Linguistics at Work*. Amsterdam : John Benjamins.

Tognini Bonelli E. (2010). « Theoretical overview of the evolution of corpus linguistics », *in* McCarthy M. et O'Keeffe A. (éd.) *The Routledge Handbook of Corpus Linguistics*. Oxford : Routledge, 14-27.

Tourrette É. (2008). *Les formes brèves de la description morale. Quatrains, maximes, remarques*. Paris : Champion.

Van Delft L. (2006). « Le fragment et les formes brèves », *in* J.-C. Darmon et M. Delon (dir.) *Histoire de la France littéraire. Classicismes XVII^e^-XVIII^e^ siècles*, t. 2. Paris : PUF, 762-792.

Zipf G. K. (1935). *The Psychobiology of Language*, Houghton Mifflin (reprinted 1965, Cambridge, MA, MIT Press).

NOTES

1. On oppose au corpus échantillonné le corpus exhaustif et clos, qui « sera étudié en tant que tel, sans prétendre à être représentatif d'autre chose que de lui-même ni à ouvrir sur aucune forme de généralisation ou modélisation. Un tel corpus est aussi, généralement, très homogène. On le rencontre notamment dans les études stylistiques ou en analyse du discours » (Mellet 2002 : § 3).
2. La représentativité supplante l'exhaustivité : « on sait que l'hypothèse de l'exhaustivité – vieux rêve de l'attitude positiviste – n'est plus tenue, et ce malgré le développement récent de la dénommée linguistique de corpus initiée dans le monde anglo-britannique, et prolongée en France par quelques auteurs avec une certaine prudence » (Charaudeau 2009 : § 4).
3. La réflexion de Dalbera peut être poussée plus en avant, puisque le corpus peut servir à la fois en phase liminaire et en phase de validation et non pas seulement dans l'un ou l'autre cas.
4. *Technology has been the major enabling factor in the growth of corpus linguistics but has both shaped and been shaped by it. The ability to store masses of data on relatively small computer drives and servers meant that corpora could be as big as one wanted* (McCarthy & O'Keeffe 2010 : 6).
5. « *The question of the size of corpora has been central to recent corpus development, and there has been the overriding belief among many corpus creators that 'biggest is best'* » (Nelson 2010 : 54).
6. Le caractère spécialisé du petit corpus repose aussi bien sur des spécificités de genre, de discours, de formes que de thématiques.
7. Se référant à Tognini Bonelli (2001), Koester rappelle : « *However, in most cases, there is some degree of variability even within a given genre, and it is therefore important to ensure that the full range of variability found is included in the corpus. For example, there may be different subgenres, or perhaps the genre is used in different types of organisations, or by different people* » (2010 : 69). Le corpus présenté dans la troisième partie de cette étude en est un exemple.
8. Cette implication vaut plutôt pour les corpus constitués à partir de types de texte plutôt qu'à partir de faits de langue.
9. La constitution d'un corpus d'énoncés est peu contraignante puisqu'il s'agit de constituer un corpus de type lexicographique, c'est-à-dire qui donne à voir des occurrences des items étudiés, sans tenir compte des types de discours, des genres, etc. qui pourraient influencer les observations menées.
10. La synonymie en langue n'est qu'une tentative pour saisir les relations sémantiques qui traversent la langue, elle n'a pas vocation à être valide dans tous les contextes qu'offrent le discours étant donné la polysémie inhérente à toute unité lexicale.
11. La synonymie vue sous cet angle a des implications quant au fonctionnement des facultés cognitives : pourquoi les locuteurs considèrent tel et tel item comme synonymes et non tel autre ? Sur quoi se fondent-ils pour établir des relations de synonymie ?

12. Même si les lexicalisations sont essentielles à la méthodologie mise en place, les étapes de la collecte des lexicalisations nécessaire à cette méthodologie ne sont pas détaillées, car la présente contribution se focalise sur la question du corpus. Pour le détail de la collecte des lexicalisations, voir Doualan (2015).

13. Il ne faut pas oublier les contraintes institutionnelles qui pèsent sur la recherche désormais : la recherche doctorale étant financée pour une durée déterminée, il est préférable d'être en mesure de traiter les résultats obtenus dans ce temps imparti.

14. https://www.gutenberg.org/. Toutefois, le texte de Pascal a été examiné à partir de la base Frantext. En effet, aucune version de l'édition scientifique des *Pensées* en texte intégral n'est disponible en libre accès ; seule est disponible l'édition remaniée de Port-Royal, seule version libre de droit. La base Frantext ne permet pas un accès au texte intégral mais les réseaux lexicaux ont été recherchés à partir des résultats fournis par la plate-forme suite à des requêtes sur les items *vice* et *vertu*.

15. La forme brève est un fragment textuel qui se caractérise par son autonomie et sa concision linguistique ; elle participe du discours discontinu (Van Delft 2006 ; Tourrette 2008).

16. L'étude des réseaux lexicaux au sein de la forme brève plutôt que sur l'ensemble du texte se justifie dans la mesure où il n'existe aucun protocole de lecture établi pour ce type de textes : leur lecture n'est pas nécessairement linéaire, chaque parcours du texte offre des connexions et des combinaisons imprévues (Van Delft 2006 : 781).

17. Les moralistes s'intéressent aux mœurs de leur siècle, pour mettre en lumière les travers de leurs contemporains, voire pour leur suggérer une ligne de conduite plus « morale ».

18. Le genre est un paramètre textuel susceptible d'introduire un biais quant à la représentativité du corpus lors de l'examen des données, puisqu'il peut influencer les phénomènes linguistiques observés, du moins ceux-ci peuvent se comporter d'une manière spécifique en raison du genre des textes du corpus. Pour donner un exemple mentionné par Tognini Bonelli (2010 : 16) : « *The large corpora of today often privilege material from an essentially unlimited source – journalism. This feature maintains the controversies about 'balance' and 'representativeness' which have been important issues since computer typesetting became almost universal. There is a clear risk that some features presented as characteristic of a language are actually characteristic mainly of its journalism. More recently the growth of electronic communication has given rise to several new and equally abundant sources, notably web pages, e-mail and blogging. All of these are uncharted territories whose communicative properties are, at the time of writing, largely unknown* ».

RÉSUMÉS

La notion de petit corpus nécessite une réflexion épistémologique pour se situer dans le paysage des sciences du langage. La taille du corpus ne pouvant suffire pour départager les petits corpus des grands corpus, la ligne de partage se situe au niveau de la représentativité et des objectifs de recherche. Le petit corpus est constitué en vue d'un objectif de recherche mené sur un domaine précis de la langue et du discours pour tenir compte des influences du type de discours et du genre sur les phénomènes linguistiques étudiés. Le petit corpus doit donc rendre compte de ce champ d'investigation particulier. La spécialisation du corpus lui confère donc sa représentativité. Pour aboutir à cette spécialisation, il est nécessaire d'établir un corpus homogène concernant les paramètres textuels, (discours, type de texte, genre, etc.). Par ailleurs, le petit corpus rend possible un retour au contexte linguistique et situationnel : une connaissance

approfondie des textes favorise l'interprétation des données. Afin d'illustrer ces propos sur les petits corpus, un corpus construit dans le cadre d'une recherche menée sur la synonymie selon une approche onomasiologique et syntagmatique est présenté. Au-delà de sa petite taille, ce corpus tire sa spécialisation de son homogénéité : les paramètres textuels ont été neutralisés, puisque les textes du corpus appartiennent tous au même discours, au même type de texte et au même genre. La recherche onomasiologique suppose d'ajouter un paramètre thématique lors de la constitution du corpus : les textes choisis doivent manifester de façon saillante la même thématique. Si les paramètres textuels semblent les plus évidents pour spécialiser un corpus, d'autres paramètres inhérents à l'objectif de recherche interviennent.

Small corpora need epistemological reflection to be set into the field of language sciences. The size of the corpus can not be enough to make a distinction between small corpora and big corpora, because the real distinction is about representativity and research objectives. A small corpus is built according to a research objective about a specific area of language to take into account the effects of type of discourse and genre on linguistic phenomena. The small corpus must be a representation of this specific field of research. As it is specialised, a small corpus can be representative. But to be specialised, a corpus must be homogeneous about textual parameters such as discourse, type of text or genre. Besides, with a small corpus, it is possible to be aware of the linguistic and situational context of texts : a detailed knowledge of the texts facilitate the interpretation of data. To illustrate these reflections about small corpora, a corpus built for a research about synonymy according to an onomasiological and syntagmatic approach is presented. Besides its small size, this corpus is specialised because of its homogeneity : the textual parameters have been neutralised, given that the texts of the corpus belong to the same discourse, the same type of text and the same genre. The onomasiological approach needs another parameters to build the corpus : a topic that must be relevant in all the texts of the corpus. If the textual parameters seem obvious and necessary to build a specialised corpus, other parameters, such as topic, depend on specific research objectives.

INDEX

Keywords : small corpus, synonymy, specialisation, representativity, homogeneity.
Mots-clés : petit corpus, synonymie, spécialisation, représentativité, homogénéité.

AUTEUR

GAËLLE DOUALAN

Université Paris Sorbonne