Les paradoxes de la fréquence

Sylvain Loiseau

Sylvain Loiseau, Paris*

Les paradoxes de la fréquence

Abstract
This article is an enquiry (mostly doxographic) about the role of the frequency in linguistic description. Frequency is mostly seen as the object of quantitative methods, but it may also be intuitive and, sometimes, it is a purely abstract argument. Under these different forms it is an ubiquitous notion in linguistics, used by linguists of almost all theoretical persuations and in almost all fields of the discipline, but it is rarely regarded as a central concept of linguistics. Taking into account this diversity of the usages of the frequency, this article tries to investigate the paradoxical status of frequency, both ubiquitous and weakly elaborated. Two theoretical frameworks may be of particular interest for a general account of the frequency dimension of the linguistic phenomena: the Coserian framework (and its concept of norm), and the usage-based view of grammar developped in a cognitive perspective. These two frameworks gound frequency respectively in the historicity of the language use and in the universality of the cognitive processing.

Keywords
frequency, repetition, quantitative methods, norm, corpus

Résumé:
Dans cet article je propose une enquête, en grande partie doxographique, sur la notion de fréquence en linguistique. La fréquence peut certes être un objet empirique et mesuré (l'objet des « méthodes quantitatives »), mais elle peut également être le résultat d'une évaluation « intuitive » ou un argument abstrait, auquel ne correspond aucune grandeur. Sous ces différentes formes, elle est présente dans un grand nombre de domaines et de traditions linguistiques, bien qu'elle soit rarement envisagée comme l'une des problématiques communes de la discipline. Derrière la diversité des arguments ayant recours à la fréquence, il s'agira donc d'essayer de construire un point de vue général sur cette notion et de s'interroger sur les aspects paradoxaux de la fréquence en linguistique, à la fois omniprésente et peu élaborée. Deux cadres théoriques en particulier me semblent donner aujourd'hui un statut fort à la dimension de fréquence des faits linguistiques : dans une perspective structurale, la linguistique unitaire de Coseriu, particulièrement à travers la notion de norme ; dans une perspective cognitive, les modèles dits « usage-based ». Ces deux cadres fondent la fréquence respectivement dans l'historicité de l'activité de parler ou dans l'universalité du fonctionnement cognitif.

Mots-clefs :
fréquence, répétition, méthodes quantitatives, norme, corpus

0. Introduction

0.1 La linguistique dite interne est classiquement définie par l'objectif de décrire des systèmes de règles aux différents niveaux de description, d'opposer possibilités et impossibilités

* L'auteur travaille au laboratoire Modyco (CNRS/Université Paris Ouest Nanterre La Défense). Ses principaux domaines de recherche sont la sémantique (et particulièrement la sémantique discursive), la linguistique textuelle et la linguistique de corpus, utilisant les corpora et les méthodes quantitatives pour décrire des propriétés des textes et des genres et essayant d'élaborer ces observables dans la perspective des typologies textuelles et dans un cadre variationniste ; il s'occupe également de développer une réflexion critique sur l'usage des méthodes quantitatives. Deux articles récents : « Investigating the interactions between different axes of variation in text typology », 2010 et « Contextualité et tactique sémantique dans un texte philosophique », à paraître. Courriel: sylvain.loiseau@wanadoo.fr

langagières indépendamment de toute considération de fréquence. Peu de cadres de travail donnent un statut important à cette propriété des faits de langue d'être répétés ou la considèrent comme une dimension centrale et constitutive des faits de langue[1]. Pourtant, la considération des propriétés empiriques les plus primaires de l'activité de parler oblige à reconnaître la dimension omniprésente, constitutive de la répétition ; ce qui caractérise particulièrement les faits de langue à tous les niveaux de description c'est qu'ils sont l'objet d'une intense répétition, aussi caractéristique et spécifique que leurs propriétés structurales. D'un point de vue méthodologique, par rapport aux autres sciences humaines et sociales, l'utilisation des méthodes quantitatives est limitée en linguistique[2] et principalement confinée aux disciplines de la linguistique dite externe, comme la sociolinguistique, la lexicométrie ou la psycholinguistique. L'élaboration théorique, la réflexion critique sur la notion de fréquence est pratiquement inexistante ; cette notion n'est pas considérée de façon consensuelle comme une notion centrale en linguistique, intéressant tous les cadres théoriques et tous les niveaux de description.

0.2 Pourtant, derrière ce constat sommaire, on peut également observer que la fréquence est mobilisée, même seulement de façon incidente, dans les contextes théoriques les plus divers. Elle est d'une certaine façon omniprésente, mais si cette omniprésence n'a jamais été prise en considération. C'est le premier paradoxe. Comme nous le verrons ci-dessous, la fréquence est régulièrement liée à certaines des notions principielles de la discipline. Elle est liée par exemple, en phonologie, à la théorie de la marque ; en morphologie elle peut être mobilisée pour établir la différence entre deux notions aussi cardinales que dérivation et composition ; en lexicologie elle ne peut jamais être évacuée totalement pour définir les unités polylexicales, c'est-à-dire pour établir les unités lexicales elles-mêmes. Dans le cadre de la description du changement linguistique elle est également un des paramètres essentiels, que le cadre récent de la grammaticalisation a remis en pleine lumière, mais qui a toujours été souligné. Dans ces différents exemples, la fréquence reçoit un statut important, instituant les catégories descriptives elles-mêmes. Cette fréquence est « théorique » au sens où elle n'est pas forcément une grandeur empirique : la légitimité de l'utilisation de la notion de fréquence pour distinguer entre deux pôles extrêmes d'un continuum, la flexion et la dérivation, ne tient pas à la possibilité d'exhiber ou d'établir concrètement des seuils de fréquence distinguant les deux phénomènes.

Je distinguerai ci-dessous différents types de fréquence : on peut opposer la *fréquence empirique*, qui est une grandeur, à la *fréquence théorique*, qui est un critère distinguant des catégories mais peu susceptibles de mesure. Enfin, je distinguerai dans la fréquence empirique la *fréquence mesurée*, quantifiée, et la *fréquence intuitive*, basée sur un savoir épilinguistique. J'essaierai de montrer que, si l'on prend en compte la diversité des références à la fréquence et les différents statuts qu'elle peut recevoir, fréquence théorique, mesurée ou intuitive, il semble que celle-ci soit omniprésente et ne puisse être assignée à telle ou telle sensibilité analytique, époque, ou objet de la linguistique. Elle peut être considérée comme une catégorie centrale, qui dépasse la question des « méthodes quantitatives ». Pour l'élaboration théorique

1 Dès que l'on donne une importance à la notion d'impossibilité, on doit recourir aux informations de l'intuition et non plus seulement à des données observables (Laks 2009 : 17).

2 Ajoutons que les méthodes de quantification sont absente, du moins en France, de la formation dans les cursus en Sciences du langage. La linguistique se distingue en cela de la plupart des sciences sociales.

de la notion de fréquence, il semble nécessaire de considérer ensemble ces différents types de fréquence.

J'exclus par contre de mon champ de recherche un certain nombre de domaines qui utilisent des méthodes quantitatives mais où l'objet n'est pas la fréquence linguistique proprement dite. C'est notamment le cas de la linguistique variationniste (labovienne). En effet, plusieurs méthodes quantitatives utilisées en sociolinguistiques sont certes quantitatives, mais ne quantifient pas une fréquence linguistique, une répétition. Par exemple, si l'on sélectionne 1 000 individus, que l'on observe la réalisation d'un phonème donné dans un énoncé, que l'on classe ces différentes réalisations en fonction de traits phonétiques, et enfin que l'on relève des corrélations entre ces traits et des propriétés socio-démographiques des locuteurs, on n'a pas observé à proprement parler une répétition, mais une corrélation entre variantes d'un phonème et groupes sociaux. Les grandeurs manipulées sont des grandeurs sociales, non linguistiques.

0.3 Le paradoxe du désintérêt des linguistes pour les questions de fréquence tient encore à ce que les données langagières ont des propriétés fréquentielles qui ne semblent pas moins difficiles à établir (à « coder » et éventuellement à compter) que les autres sciences humaines – du moins du point de vue du statisticien, qui considère des faits déjà données, comme la fréquence des mots graphiques dans un texte. Guiraud (1959 : 15) dit ainsi de la linguistique qu'elle constitue « la sciences statistique type »[3]. Si les méthodes quantitatives sont peu utilisées en linguistique, les « données langagières » intéressent les statisticiens. L'intérêt statistique de ces données est démontré par les nombreux développements méthodologiques élaborés sur des données langagières devenus des méthodes statistiques d'usage courant en dehors de la linguistique. Par exemple, pour modéliser les probabilités de transitions entre lettres-consonnes et lettres-voyelles dans *Eugène Onéguine*, Markov élabore en 1913 la méthode dite des chaînes de Markov. Pour résumer de grands tableaux notant les fréquences de mots graphiques dans différents sous-ensembles d'un corpus, Benzécri (1973) développe l'analyse factorielle dite des correspondances. Pour modéliser en terme probabiliste la distribution des fréquences de mots graphiques dans un texte, Zipf (1935) élabore une loi de distribution dite « loi de Zipf » (Baayen 2001, Petruszewycz 1973), qui est aujourd'hui utilisée dans de nombreux domaines (Newman 2005). Alors que les faits de fréquence intéressent peu les linguistes, les propriétés fréquentielles des données langagières sont statistiquement remarquables et ont permis, plus qu'aucune autre science humaine, le développement de méthodes statistiques.

Ces données langagières ne sont pas cependant nécessairement des objets linguistiques ; elles en sont parfois des approximations (par exemple, les lettres pour des phénomènes phonétiques ou les mots graphiques pour des phénomènes lexicaux). C'est, là encore, un aspect paradoxal de la fréquence (en tout cas de la fréquence mesurée) en linguistique : pour mesurer des fréquences, qui constituent une dimension empirique des faits de langue, il faudrait souvent appauvrir voire perdre ces faits de langue et se rabattre sur des « objets approximatifs », comme les mots graphiques. Cette difficulté à établir les unités n'est cependant pas propre aux méthodes quantitatives : on ne fait ici que retrouver les difficultés

3 Cf. également Heilmann (1972 : 306) : « […] in a classification dividing sciences into natural – dealing with laws – and statistical ones – dealing with trends – linguistics may appear to be a typical statistical science in that it deals with communicative behavior, with events neither completely predictable nor completely unpredictable. »

bien connues de l'élaboration des unités de la description en linguistique, dont « [la] délimitation est un problème si délicat qu'on se demande si elles sont réellement données. » (Saussure 1995 : 149). On peut faire l'hypothèse que le quantitatif fait construire des objets spécifiques[4], mais dans tous les cas, la question de leur articulation à des catégories qualitatives ou interprétatives reste l'une des difficultés les plus constantes, et nourrit la plupart des objections à l'utilisation de fréquences mesurées.

0.4 La question de la fréquence a une actualité particulière dans le contexte du développement de la « linguistique de corpus ». La disponibilité aujourd'hui de corpus importants et d'outils statistiques peut faire présager que des changements profonds sont en cours dans la discipline dans son rapport au quantitatif. Glessgen (2001 : 424) prédit ainsi que « les méthodes de quantification sont sur le point de transformer la recherche et l'histoire linguistique comme elles ont transformé, il y a quelques décennies, la recherche sociologique ou psychologique ». Laks (2008 : 23) souligne de même : « Bien que souvent inaperçus par les linguistes eux-mêmes, le champ conceptuel de la linguistique contemporaine internationale a connu des changements récents très substantiels : la relation entre théorie et données, modèles abstraits et observables, s'est totalement inversée. Des modèles sous-déterminés par les données, on est ainsi passé à des approches qui placent au premier plan les observables et les phénoménologies construites explicitement. Les modélisations quantitatives [...], statistiques [...] et probabilistes [...] se sont développées. ». Des cadres relativement récents comme la grammaticalisation, les grammaires « usage-based », la sociolinguistique historique ou la dialectométrie font entrer l'intérêt pour les questions quantitatives dans des champs de recherche où ils étaient naguère marginalement utilisés.

Cependant, pour apprécier cette mutation il faut sans doute la rapporter à l'histoire, déjà longue, de l'utilisation de données quantitatives dans la discipline. Peut-être les renouvellements paraîtront-ils moindres ou différents qu'escomptés. D'autre part, alors que des méthodes quantitatives sont de plus en plus mobilisées pour accéder aux régularités que les corpus permettent d'observer, il paraît nécessaire d'ouvrir une réflexion critique sur la notion de fréquence également pour se prémunir contre certaines visions positivistes. Celles-ci consistent, notamment, dans la tentation récurrente de prendre toute utilisation de méthodes quantitatives comme une garantie de scientificité, comme semble le faire par exemple Johnson (2008 : 4), qui voit dans l'utilisation des « [...] same quantitative methods used in science and engineering » une garantie pour la linguistique d'être « […] a member of the scientific community ». Là encore, le débat est-il vraiment nouveau, et faisons nous vraiment autre chose que rouvrir des dossiers largement patrouillés, même si les questions apparaissent aujourd'hui « nimbée de l'aura de technologies et d'outils très sophistiqués » (Laks 2008 : 8) ? Les critiques du recours au quantitatif comme à un critère de scientificité en lui-même sont largement présentes au moins depuis le milieu du siècle dernier. Citons, parmi d'autres, Wells (1957 : 136) qui rappelle utilement que « Not every scientific treatment is mathematical. And neither is every mathematical treatment scientific; for it is possible to miss-apply mathematics, and to misinterpret its results. », ou Coseriu (1965 : 95) : « l'exactitude réside en réalité dans la pensée et dans sa correspondance aux faits, et non pas dans les symboles et les chiffres, qui sont de simples instruments […] pour l'expression de la pensée ». Un argument particulièrement récurrent aujourd'hui est la garantie de cumulativité

4 « The regularities, both lexical and grammatical, that can be identified in the context of the node define the boundaries a *new unit of currency* in linguistic description. » (Tognini-Bonelli 2001 : 18)

que donnerait l'utilisation de quantifications (« Corpus work is cumulative […] » Stubbs 2001 : 223). Or, la cumulativité n'est pas l'accumulation de faits ou de relevés quantitatifs, d'ailleurs souvent incomparables parce que basés sur des hypothèses complètement différentes. De fait, la linguistique de corpus ne paraît pas être une discipline particulièrement cumulative.

Il faut ajouter enfin que les questions de fréquences, du moins quand elles prennent un tour empirique, font l'objet de l'un des préjugés défavorables les plus solidement établis de la disciplines : les questions de fréquences « ne [produisent] que des énoncés de probabilité, et non des règles, ce qui, pour bien des linguistes, est dépourvu d'intérêt. » (Labov 1976 :128). Ce mépris n'est d'ailleurs pas incompatible avec une certaine fascination pour les méthodes quantitatives.

0.5 Cette enquête sur la notion de fréquence a donc pour objectif de s'interroger sur les constantes dans la discipline dans l'utilisation de la fréquence, et sur les points communs entre les différents types de fréquence. Elle a également pour objectif de s'interroger sur les différents fondements théoriques possibles pour la prise en compte des faits de fréquence et leur articulation à des catégories descriptives. Les enjeux d'une élaboration critique de la notion de fréquence sont notamment le progrès dans la prise en compte et la description de l'activité de parler concrète, sur lesquelles mettent l'accent de nombreuses approches contemporaines (dans des approches cognitives comme sociolinguistiques ou textuelles).

Enfin, cette enquête essaiera d'apporter des éléments de réponse à des questions qui dépassent naturellement le cadre du présent article : quels sont les enjeux des méthodes quantitatives pour la description ? Avec quels objectifs descriptifs s'accordent-elles ? Sur quelles propriétés de l'objet décrit peuvent-elles s'appuyer ? Quels sont les statuts des données quantitatives et les règles de leur interprétation ? Comment articuler faits quantitatifs et catégories descriptives ? Comment « qualifier » ou « valider » des faits quantitatifs ? Mon information reste naturellement biaisée dans le sens des domaines que j'ai le plus travaillés : sémantique et linguistique textuelle.

Dans la suite de cet article, je montrerai d'abord la diversité des formes que prend la notion de fréquence puis son caractère à la fois omniprésent et peu élaboré. Dans la troisième partie je rappellerai certaines spécificités du quantitatif (de la fréquence mesurée) en linguistique par rapport à d'autres sciences humaines. Dans la quatrième partie, je montrerai ce que la notion de norme – notamment l'élaboration qu'en a proposé Coseriu – peut apporter pour la description des faits de fréquence. Dans cette perspective, les faits de fréquence sont rapportés à la dimension d'historicité de la langue. On peut lui opposer le cadre des grammaires cognitives usage-based, où la notion de fréquence est davantage fondée sur l'universalité du fonctionnement cognitif. Ces deux cadres illustrent les deux termes de l'alternative fondamentale quant au fondement possible de la dimension de répétition de l'activité de parler.

1. Les différents types de fréquences

1.0 La première distinction qu'il importe peut-être de faire est la distinction entre la fréquence comme objet d'une méthode expérimentale (*fréquence mesurée* ou fréquence objective), la fréquence comme évaluation intuitif (*fréquence intuitive* ou fréquence subjective[5], 1.1) et la

5 Un chiasme est possible : Germain (1981 : 158) parle « d'objectivité relative » pour un sens relativement

fréquence comme propriété générale attribuée à des faits de langue indépendamment de toute grandeur, objective ou subjective (*fréquence théorique*, 1.2).

1.1 Les jugements intuitifs sur le fait que tel phénomène est fréquent ou rare sont eux-mêmes... fréquents dans la littérature. Une forte réflexion sur cette « quantification intuitive » (*informal quantification*) et sur les difficultés de la mobilisation de données quantitatives a été proposée par Schegloff (1993) dans le cadre de l'analyse conversationnelle (AC). En effet, l'AC comme discipline interprétative accorde une importance particulière à l'établissement des phénomènes pertinents par les locuteurs, en fonction de toutes les potentialités signifiantes de l'activité verbale[6] ; les phénomènes (ce qu'en termes statistiques on peut appeler le « codage » des données) ne sont donc pas donnés préalablement de façon univoque. Les approximations inhérentes à tout codage sont particulièrement inacceptables ou inappropriées dans cette perspective.

Schegloff (1993 : 118-119) défend, dans ce cadre, l'utilisation de quantifications intuitives et montre les apories auxquelles peut mener l'utilisation de méthodes quantitatives en AC. Il souligne que la différence entre fréquence intuitive et quantification est une différence de nature, et non pas seulement une différence de degré de précision :

> [informal quantification *vs* more formal quantitative techniques] are *not* simply weaker and stronger versions of the same undertaking; they represent different *sort* of accounts. Formal quantitative analysis is the outcome of a set of procedures focused on « precise » numerical characterization […] Terminology such as *occasionally* or *massively* report an *experience* or *grasp* of frequency, not a count: an account of an investigator's sense of frequency over the range of a research experience, not in a specifically bounded body of data.

Si *count* et *experience* traduisent deux *types* de fréquences différentes, si la fréquence intuitive a une légitimité (et une spécificité) il me semble que c'est, plus généralement, parce qu'elle relève d'un savoir épilinguistique. En effet, des jugements et des représentations portant sur la fréquence sont comme on le sait accessibles à l'intuition de chaque locuteur (en donc, en second lieu, du linguiste-locuteur ou *investigator*) : ils font partie d'une certaine façon du savoir sur la langue, au même titre que le savoir épilinguistique syntaxique ou lexical. En tant que savoir épilinguistique, la fréquence perçue fait partie de l'objet de la discipline[7]. L'existence épilinguistique de la dimension de fréquence est l'un des fondements de sa pertinence et de son importance linguistique.

Ces jugements épilinguistiques de fréquence peuvent prendre des formes et des objets aussi variés que les « mesures objectives » : par exemple, ils peuvent porter sur la fréquence d'une unité (telle unité est-elle fréquente ?), c'est-à-dire résulter d'une abstraction à partir de l'ensemble des expériences linguistiques du locuteur. Mais ils peuvent porter également sur la fréquence d'une unité dans un texte ou une interaction particulière. Ils intéressent également la catégorisation des variétés par les locuteurs. Plusieurs descriptions soulignant que cette catégorisation fait intervenir une perception quantitative intuitive[8] (Labov 1976 : 172-173) :

---

proche.

6 De ce fait l'AC est assez proche d'une linguistique « de l'activité de parler » universelle selon les termes de Coseriu ; cf. Schlieben-Lange 1998 : 261.

7 « […] tout ce que le sujet parlant naïf pense de sa langue est déterminant pour le fonctionnement de celle-ci : les opinions du sujet parlant à propos de la langue appartiennent, à la rigueur, à l'objet « langue » et, par conséquent, on ne saurait les ignorer. » (Coseriu 2001 : 17)

8 *Contra* Gadet (2003), qui aborde cependant plutôt la stigmatisation que l'identification des variétés :

> L'emploi isolé d'une variante, même très stigmatisée, telle une diphtongue centralisée dans « *boid* » au lieu de « *bird* » (oiseau), ne produit pas d'ordinaire une forte réaction sociale ; il peut tout au plus installer une expectative, telle que l'auditeur commence à percevoir une structure socialement significative. Il n'est pas de locuteur qui, à l'occasion, ne prononce pas un (dh) initial avec une attaque forte, produisant ainsi un son interprétable comme une affriquée, [d đ]. […] En fait, c'est la fréquence avec laquelle Bennie N. emploie de telles formes qui possède une signification sociale, et c'est essentiellement entre un niveau de fréquence et un autre que le contraste se manifeste dans la structure.

La fréquence a donc une double dimension à la fois objective et subjective. Une question centrale est celle de l'articulation entre le versant subjectif et le versant objectif. Les jugements de fréquence des locuteurs diffèrent-ils de fréquences mesurées, et, dans ce cas, ces divergences ont-elles un caractère systématique (une explication) ? Troubetzoy (1986 : 8-9) expose clairement les incompatibilités entre les mesures et la fréquence subjective.

> Il va de soit que de telles « normes » [normes de la parole, issues de statistiques] ne peuvent avoir qu'une valeur de moyenne et qu'on ne peut les assimiler aux valeurs de la langue. […] Si dans un texte on examine soigneusement quant à leur degré de souffle tous les « k » qui s'y présentent, qu'on exprime par un chiffre le degré de souffle dans chaque cas particulier, et qu'on calcule ensuite la valeur moyenne du souffle de *k*, cette valeur moyenne ne correspondra à aucune réalité : tout au plus représentera-t-elle la fréquence relative de l'apparition d'un *k* devant une voyelle accentuée. Des résultats non ambigus ne pourraient être obtenus que si l'on calculait deux valeurs moyennes différentes, l'une pour *k* devant voyelle accentuée, l'autre pour *k* devant voyelle inaccentuée. mais la norme à laquelle se réfèrent les sujets parlants est « *k* en général », et celui-ci ne peut être établi par des mesures et des calculs.

Ces incompatibilités sont dues au fait que les unités ne peuvent être les mêmes.

1.2 Enfin, à côté de la mesure et des jugements intuitifs, on peut distinguer encore le recours à la fréquence comme à un critère théorique central, indépendamment de toute *grandeur* (mesurée ou intuitive). Meillet (1936 : 130-131, cf. Germain 1981 : 9) propose par exemple de définir le sens d'un mot comme une moyenne des différents emplois de ce mot, c'est-à-dire une moyenne qui n'est pas présentée comme susceptible d'être mesurée empiriquement. La grandeur dont cette moyenne serait une moyenne n'est pas définie. Ce qui est posé, c'est l'idée générale que certains phénomènes existent et ne peuvent être décrits que comme une *abstraction* d'un grand nombre d'occurrences singulières. Ce type de fréquence est paradoxal au sens où il s'agit d'une dimension à la fois empirique de la langue, et en même temps inaccessible à l'observation. Cette fréquence, que l'on appellera « fréquence théorique », ne désigne pas ce qui est « non quantifié » par opposition aux quantifications, mais ce qui est non quantifiable par rapport à ce qui est quantifiable. Sous l'espèce « théorique », les recours à la fréquence paraissent omniprésents, et concernent de nombreux domaines. Lüdtke (1996 : 538) indique ainsi, pour distinguer entre les situations de contact entraînant ou n'entraînant pas de changement linguistique : « Entre l'interférence [de deux langues dans le cerveau d'un locuteur bilingue] et le changement, il y a la statistique ».

Cette fréquence abstraite et principielle est présente dans de nombreux concepts centraux de la discipline. Sans tergiverser, venons-en directement aux plus éminents d'entre eux : on la

---

« Ainsi, on peut supposer que les variables qui permettent d'évaluer ou d'affirmer la différence entre ingroup et outgroup ont une saillance qui les rend pour le locuteur socialement signifiantes. Loin que la fréquence soit toujours significative, une seule occurrence de formes fortement stigmatisées peut suffire à connoter négativement un discours (pour le français, on peut penser à des « fautes » de morphologie, comme des chevals, il a s'agi ou il éteignera, ainsi qu'à certaines liaisons fautives). »

trouve notamment chez Saussure dans l'exposition de la dichotomie langue / parole (1995 : 29-30) :

> Entre tous les individus ainsi reliés par le langage, il s'établira une sorte de moyenne : tous reproduiront, – non exactement sans doute, mais approximativement – les mêmes signes unis aux mêmes concepts. […] Si nous pouvions embrasser la somme des images verbales emmagasinées chez tous les individus, nous toucherions le lien social qui constitue la langue.

La « moyenne » (puis la « somme »[9]) a sans doute un statut métaphorique (marqué par l'enclosure « une sorte de »). Cette métaphore est cependant récurrente dans le CLG[10]. Elle occupe une place stratégique, celle de l'articulation entre parole (« reproduiront ») et langue, du passage de l'une à l'autre[11] ; c'est bien sûr dans cette articulation que se concentrent toutes les difficultés de la dichotomie, après que les deux termes ont été séparés par un « abîme » (Coseriu 1982 : 14). Toute interrogation sur la fréquence porte en même temps sur la nature de l'articulation entre les deux termes de la dichotomie, sur la nature de l'abstraction linguistique et sur la solution de l'antagonisme entre « la libertà delle uso singolo et il determinismo dell'insieme » (Heilmann 1983 : 218). Le choix le plus radical, à cet égard, consiste à simplement rabattre l'opposition langue / parole sur le couple statistique « population » / « échantillon » (corpus) et à réduire le processus d'abstraction de la description linguistique à la généralisation statistique (Herdan, Guiraud, cf. ci-dessous 3.6.1). La phonométrie de Zwirner de même vise à « jeter un pont par dessus l'opposition existant entre phonologie et phonétique » (Troubetzkoy 1986 : 7)

Chez Saussure, puis chez les saussuriens, la perspective fréquentielle est donc indissociable de l'élaboration de la dichotomie langue/parole. Elle occupe une place stratégique dans l'articulation des deux termes de la dichotomie. Cette place a été peu relevée par rapport aux interprétations de la dichotomie en termes pyschologique ou institutionnel. La fréquence est indissociable de la définition du processus d'abstraction sur lequel repose la description linguistique. Rappelons que le statut des « lois synchroniques » auxquelles aboutit la description est fondé, pour Saussure, sur une dimension de régularité, par opposition à une universalité : « En résumé, si l'on parle de loi en synchronie, c'est au sens d'arrangement, de régularité » (1995 : 131)[12].

1.3 Sans chercher l'exhaustivité, distinguons entre plusieurs formes concrètes que prend la fréquence (qu'elle soit mesurée, intuitive ou théorique) : fréquence absolue et relative, moyenne, covariation, cooccurrence et récurrence. Dans les utilisations de statistiques pour la description de fréquence, les observations de corrélations, de cooccurrences ou de récurrences ont été particulièrement privilégiées.

---

9 Si la langue est la somme des images verbales, la parole est « la somme de tout ce que disent les gens » (1995 : 37)

10 La langue est « une somme d'empreintes déposées dans chaque cerveau, à peu près comme un dictionnaire dont les exemplaires identiques seraient répartis entre les individus » (38) ; « l'ensemble des habitudes linguistiques qui permettent à un sujet de comprendre et de se faire comprendre » (112).

11 « C'est dans la parole que se trouve le germe de tous les changements : chacun d'eux est lancé d'abord par un certain nombre d'in[divi]dus avant d'entrer dans l'usage. [...] Un fait d'évolution est toujours précédés d'un fait, ou plutôt d'une multitude de faits similaires dans la sphère de la parole » (1995 : 138-139).

12 Cf. également (Mańczak 1969 : 19) : « […] il faut constater que la linguistique est née le jour où l'on a formulé la première règle de grammaire, et des notions linguistiques aussi élémentaires et fondamentales à la fois que la "règle" et "l'exception" ont un caractère quantitatif : la règle est ce qui est fréquent, alors que l'exception est ce qui est rare. »

1.3.1 Une covariation entre phénomènes est observée à travers une variation concomitante de plusieurs fréquences. Ces covariations sont particulièrement utilisées dans le contexte de la linguistique variationniste, où elles impliquent d'un côté un phénomène linguistique et de l'autre un phénomène social. La covariation comme méthodologie est définitoire de cette tradition descriptive, elle présuppose une catégorisation indépendante des deux ordres de phénomènes.

La covariation peut également être observée entre des phénomènes uniquement linguistiques. C'est une dimension importante de l'appréhension quantitative des faits de langue : ceux-ci, particulièrement dans le domaine du texte, sont souvent caractérisés par un très grand nombre de phénomènes simultanés. Par exemple, comparer quantitativement des auteurs ou des genres implique des méthodes comparant plusieurs milliers de variables (par exemple les différentes formes lexicales ou les différents traits morphosyntaxiques, etc.). Si une seule fréquence (la fréquence d'une forme lexicale donnée par exemple) ne suffit pas à caractériser un genre, en revanche la prise en compte de nombreuses fréquences garantit de saisir une « signature » du genre décrit. Cette particularité des phénomènes linguistiques est à l'origine de l'élaboration de la méthode factorielle dite des correspondances (Benzecri 1973). L'attention aux phénomènes de corrélations est particulièrement importante pour la description de variétés ou de genre, une variété n'étant pas caractérisée par une seule fréquence, mais par des corrélations entre un grand nombre de fréquences (Biber 1995 : 13) :

> By themselves, however, frequency counts cannot identify linguistic dimensions. Rather, a linguistic dimension is determined on the basis of a consistent co-occurrence pattern among features. That is, when a group of features consistently co-occur in texts, those features define a linguistic dimension.

Une certaine « confiance » dans les résultats est souvent gagée sur la stabilité de ces rapports entre fréquences, au-delà des fréquences elles-mêmes. L'enjeu de la covariation est donc de permettre de trouver une stabilité, une robustesse, des rapports constants au-delà de variation de fréquences, c'est-à-dire, à nouveau, de pouvoir réaliser une abstraction à partir de fréquences.

L'intérêt pour les méthodes factorielles est ancien : Wells (1957 : 120) par exemple relève les avantages d'une méthode statistique permettant d'observer des variations concomitantes d'un grand nombre de variables ; cette variation concomitante rapportée ici à la redondance du « code » linguistique (120) : « [In factor analysis] there is *redundancy*, in a sens essentially like the sens in which Information Theory speaks of redundancy. »

1.3.2 La notion de cooccurrence intéresse la question de la définition des unités de la description. À partir de phénomènes isolés (tels des mots graphiques), on s'interroge sur leur tendance à apparaître ensemble plus souvent (ou plus rarement) que ce à quoi on peut s'attendre dans l'hypothèse d'une répartition indépendante de leurs occurrences respectives. Les phénomènes qui « co-occurent » peuvent être considérés, à l'issues de telles procédures, comme des unités. De fait, la notion de cooccurrence est particulièrement utilisée dans le domaine lexical pour l'identification des unités polylexicales.

1.3.3 On peut enfin identifier la récurrence, c'est-à-dire le fait pour un phénomène d'apparaître plusieurs fois, d'être répété. C'est par exemple le critère définitoire de l'isotopie (« l'effet de la récurrence d'un même sème », Rastier 2001 : 299). Une condition nécessaire (mais non suffisante) d'une isotopie est donc qu'il y ait « au moins deux » occurrences du

sème. Relève également de cette dimension de récurrence la redondance classiquement attribuée au code linguistique[13].

1.4 Il peut être utile également de distinguer différents types de fréquence en fonction de la « généralité » donnée à cette fréquence : d'un côté, une propriété fréquentielle peut être posée comme universelle, trans-linguistique ; à un autre extrême, une fréquence peut caractériser un événement historique singulier (un texte, un style, un genre), à l'intérieur d'une langue. Il est remarquable que les travaux sur la fréquence mesurée privilégie d'un côté des propositions très universalistes et de l'autre, une utilisation de la quantification dans une perspective textuelle, voire herméneutique, tandis que peu de contributions s'intéressent aux régularités propres à une langue fonctionnelle.

Le premier cas, celui des lois quantitatives universelles, est représenté par excellence par les différentes « lois de Zipf », selon lesquelles il y a un rapport constant entre rang et fréquence d'une unité, entre fréquence et polysémie, et entre fréquence et « économie ». Ces rapports sont censés être des constantes mathématiques des langues, des propriétés internes et universelles des langues comme codes, fondés d'ailleurs dans une tendance au moindre effort qui ne s'appliquerait pas qu'aux langues. Des régularités très abstraites sont donc rapportés à des principes explicatifs extrêmement généraux.

D'un autre côté, des observations impliquant des faits de fréquences peuvent porter sur des objets historiques : texte, genre, style, etc. Dans ce cas, la généralisation, si besoin, passe par une méthode comparative. Cet usage de la fréquence me semble être commun aussi bien aux descriptions de « dimensions de variation » proposées par Biber (1995) qu'aux descriptions de discours ou de texte, comme dans la tradition lexicométrique avec par exemple dans cette dernière la méthode des spécificités (Lafon 1980).

Cette variation dans l'usage des faits quantitatifs, entre des perspectives universelles ou des perspectives plus historiques, montre la diversité des degrés d'abstraction que l'on peut tirer des faits de fréquence : la détermination du degré d'abstraction choisi est la question essentielle de toute utilisation de faits de fréquence.

On peut remarquer qu'il y a peu de descriptions quantitatives qui se donnent comme objet un niveau intermédiaire, en termes d'abstraction, entre ces deux pôles, et qui correspondrait en gros au niveau des langues. Peu de travaux utilisent des faits de fréquences pour caractériser une langue isolément. Il n'y a pas sans doute à s'en étonner puisque la notion de langue peut précisément être définie comme ce qui peut être décrit par un système d'oppositions fonctionnelles, à l'exclusion de tout critère fréquentiel. Les exceptions à ce constat sont par exemples les travaux de Martinet (1955) pour définir l'état d'équilibre d'un système fonctionnel, fondé sur la prise en compte du rendement fonctionnel des oppositions.

La recherche de lois quantitatives universalistes en linguistique suppose de pouvoir décrire une structure mathématique au cœur du langage qui serait abstraite de l'historicité des langues et des usages. Néanmoins, les lois ainsi dégagées restent le plus souvent des approximations des faits réellement observées dans des situations historiques concrètes. Dès lors, on peut se demander si ce niveau d'abstraction a une pertinence descriptive. C'est la critique que Martinet adresse aux lois de Zipf. Martinet (1955 : 132) reconnaît que « de façon générale, moins les apparitions d'un phonème sont prévisibles et attendues, ce qui veut dire, en gros, fréquentes, plus grande est sa valeur distinctive et moins les locuteurs sont tentés

13 Cf. (Jakobson 1963 : 89). L'isotopie elle-même ne relève pas de cette question puisque la redondance qui l'intéresse est une redondance d'un signifié, pas d'un signifiant (Rastier 1981 : 25-26).

d'en négliger l'articulation. » Mais peut-on traduire quantitativement ce rapport général sous la forme d'une loi mathématique ? Troubetzoy, dans la discussion des rapports entre marque et fréquence, préfère, selon Martinet (1955 : 133), « renoncer sagement à établir, entre les deux grandeurs, autre chose qu'un rapport assez vague ». Pour Troubetzkoy (1986 : 282), « Dans sa rédaction phonologique cette théorie [de Zipf] pourrait se présenter ainsi : " des deux termes d'une opposition privative le terme non marqué apparaît plus souvent dans le discours suivi que le terme marqué" ». Il ajoute (1986 : 282) : « S'il n'y a aucun doute que la distinction entre termes d'opposition marqués et non marqués, de même que la distinction entre oppositions neutralisables et non neutralisables, ont une influence sur la fréquence des phonèmes, il est toutefois également clair que ces faits ne suffisent pas à expliquer les rapports de fréquence. » C'est donc cette loi fréquentielle générale qui apparaît comme *plus abstraite* que le niveau de description phonologique, et non l'inverse : dans une telle loi, les faits de fréquence sont sous-déterminés par les catégories descriptives.

La critique que Coseriu adresse à la glottochronologie repose aussi sur la non prise en compte des faits historiques. La glottochronologie est une méthode proposée à la fin des années 50 pour dater le « moment » de la divergence entre deux langues apparentées. Elle postule une constante anhistorique : il y a, à chaque intervalle de mille ans, une proportion constante du « lexique fondamental » qui devient privatif entre deux langues issues d'un ancêtre commun (80 à 85 %). Étant donnée cette constante, il suffit d'observer la proportion du lexique fondamental qui est partagée, à un moment donné, entre deux langues apparentées, pour calculer le moment du début de leur divergence[14]. Outre la naïveté du cadre diachronique utilisé (qui repose sur la Stammbaumtheorie), Coseriu (1965 : 93) insiste sur l'absurdité à laquelle mène la tentation d'élaborer une constante indépendamment des phénomènes historiques : « La décadence des signifiants dépend des circonstances historiques particulières à chaque parler et n'a aucune relation définie avec la chronologie absolue. » Il souligne « [un] danger encore plus grave que ne l'est la glottochronologie en elle-même : c'est celui qu'implique la quantification de ce qui n'est pas quantifiable, la prétention de remplacer la méthode comparative et l'histoire par les mathématiques et le calcul. »

C'est donc le choix du niveau d'abstraction qui est en cause : les régularités extrêmement générales que l'on peut constater à un niveau indépendant des langues font négliger les réalités historiques, et ne donnent accès qu'à des généralités non nécessairement interprétables. Elles ne permettent pas non plus d'obtenir de réelles constantes mathématiques. Le meilleur exemple en est la plus connue des lois de Zipf, selon la quelle le produit du rang et de la fréquence d'une forme est constant. Si celle loi saisit sans doute, dans une forme mathématique simple, une tendance massive, elle reste difficile à formuler mathématiquement. La recherche d'une telle formulation mathématique est encore l'objet de travaux (Baayen 2001), mais on peut se demander si c'est un objectif raisonnable de continuer la complexification de modèles mathématiques pour décrire un niveau d'abstraction qui, de toute façon, néglige la réalité historique et reste ininterprétable.

1.5 On peut enfin opposer les travaux d'« attribution » et les travaux de description proprement dit. En effet, une des premières utilisations des statistiques linguistiques a été d'assister des questions d'attribution : des indicateurs quantitatifs qui résument des régularités stylistiques profondes inaccessibles à la lecture peuvent devenir des arguments pour trancher

14 La glottochronologie utilise donc des faits quantitatifs mais non des faits de fréquence linguistique (répétitions).

en faveur de l'attribution disputée de tel texte à tel auteur. Les sonnets de Shakespeare, ou jusqu'à très récemment les comédies de Molière, ont fait l'objet de tels travaux. Les faits (quantitatifs) qui permettent l'attribution ou non d'une œuvre à un auteur peuvent cependant n'intéresser en rien la description. Ces « signatures » peuvent être par exemple la coprésence de tels traits grammaticaux, permettant infailliblement de reconnaître un auteur, mais peu interprétables et peu intéressants pour décrire son idiolecte.

À propos de description stylistique, Rastier (2001b) oppose « identification » et « caractérisation » :

> […] distinguons […] l'identification et la caractérisation, ou si l'on préfère les traits « morelliens » et les traits « spitzériens ». Morelli, médecin italien, révolutionna à la fin du XIXe siècle les attributions de tableaux en décelant des traits, notamment anatomiques, comme les lobes d'oreille, dont la facture caractéristique échappait jusque là aux faussaires comme aux experts. Quant à Spitzer, on lui a maintes fois reproché de caractériser les œuvres par des traits formels qui paraissaient choisis arbitrairement, et lui permettaient pourtant d'entrer dans le cercle d'une interprétation révélatrice.

Cette opposition entre l'identification et la caractérisation s'applique très bien à la distinction entre utilisation de données fréquentielles pour l'attribution ou pour la description. Comme les « traits morelliens », les faits quantitatifs utiles pour trancher une question d'attribution sont des « patrons » arbitraires qui n'intéressent pas nécessairement la description. Les traits qui permettent d'identifier un sonnet comme shakespearien ne sont pas nécessairement ceux qui intéressent les exégètes de Shakespeare. Comme les « traits spitzériens », les faits quantitatifs utiles à la description, au contraire, doivent pouvoir être interprétés, c'est-à-dire rapportés à d'autres catégories descriptives.

Ce n'est certes pas un hasard que des travaux « d'identification » aient été les premiers à utiliser des données quantitatives : dans le cas de travaux d'identification, la nécessité d'articuler les données quantitatives à des catégories descriptives disparaît.

Il faut souligner qu'il y a un continuum entre ces deux pôles et que de nombreux travaux qui tendent à « caractériser » en restent peut être essentiellement à l'identification. Par exemple, on peut se demander si la définition contrastive d'un genre, au moyen d'une analyse factorielle, par un ensemble de traits, est un résultat en termes de caractérisation ou d'identification. Muni de ces informations, en effet, on peut infailliblement identifier le genre d'un nouveau texte (le « classer »). Mais le résultat descriptif pour la connaissance de ce genre – et surtout la validation de la pertinence de l'unité genre choisie – n'est pas garanti. Dans les travaux de Biber (1988), les corrélations de traits définissant chaque genre peuvent faire l'objet d'une interprétation et d'une analyse ; mais rien ne garantit que des regroupements et des interprétations similaires n'auraient pu être faits avec de tout autres corpus.

2. Diversité et omniprésence de la fréquence

2.0 Au risque de lui faire prendre un aspect « catalogue », l'objectif de cette partie est de montrer l'omniprésence de la notion de fréquence en linguistique et la variété des formes sous lesquelles elle est convoquée. Ce bilan n'est pas exhaustif, mais essaye de parcourir autant que possible le spectre de cette variation.

2.1 Nous avons vu ci-dessus que l'élaboration de l'opposition langue/parole mobilise, de façon incidente, une considération de fréquence ; une fréquence abstraite, « théorique ». Si

l'on prend en compte l'ensemble de ces considérations de fréquence, qu'elles soient théorique, intuitive ou mesurée, c'est un grand nombre de notions centrales de la linguistique qui apparaissent liés à la fréquence. Ces notions sont à l'origine de nombreux développements qui ne doivent rien à des considérations quantitatives ; elles sont souvent présentées sans référence à cette relation à la fréquence ; pourtant, si on examine leur élaboration, la fréquence n'en est jamais absente.

Au niveau phonologique, la question de la fréquence est particulièrement présente dans l'élaboration de la notion de marque chez Troubetzkoy, comme nous l'avons déjà souligné (ci dessus 1.4).

La distinction des différentes parties du discours mobilise traditionnellement une opposition entre des ensembles finis, énumérables (les morphèmes grammaticaux) et des ensembles « infinis » (les morphèmes lexicaux). Ainsi Martinet (1970 : 119) oppose d'abord lexème et morphème (grammème) par le fait que « [les] monèmes grammaticaux sont ceux qui alternent, dans les positions considérées, avec un nombre relativement réduit d'autres monèmes ». Ce critère, comme on le sait, est pratiquement peu opératoire. Le même Martinet indique ainsi plus tard (1985 : 106) que « les choses ne sont pas si simples » : « il n'est même pas sûr que les inventaires de ce que nous désignons comme des modalités, comme les articles, ou les temps et modes verbaux, soient aussi figés qu'on le croirait à première vue ». Un critère de fréquence est finalement donné : « Sans donc faire état d'une distinction catégorique fondée sur la délimitation des inventaires, on retiendra cependant le fait qu'on trouve partout une opposition entre des inventaires dont la fréquence moyenne des unités est élevée et ceux où cette fréquence est basse ».[15]

Dans le domaine lexical, les critères de fréquence sont toujours nécessaires pour définir les unités polylexicales. Ces unités polylexicales ont reçu des dénominations diverses (lexies complexes, synthèmes, etc.) et sont toujours nécessaires pour se donner une unité lexicale qui soit autre chose que le mot graphique. Par exemple – et pour rester chez Martinet – c'est encore la fréquence qui doit être invoquée en dernier recours pour fonder la distinction entre lexie et syntème : à partir de quel moment « la corne de l'Afrique » doit être considérée comme une unité ? Après avoir évoqué des tests, il faut recourir aux corpus (1985 : 39) : « Des données statistiques peuvent probablement permettre de préciser à quel moment de l'évolution de la langue un complexe atteint une fréquence comparable à celle d'unités simples du même type, ce qui va automatiquement tendre à dissuader les usagers de restituer une autonomie à ses composants. »

Dans ces trois exemples de la marque, des parties du discours et des unités polylexicales, la fréquence n'est pas un élément définitoire suffisant mais ne peut pas non plus être entièrement ignorée. À chaque fois, c'est la définition d'unité ou de catégorie fondamentales de la description qui est en jeu.

2.2 Si l'on prend maintenant comme perspective les différents domaines de la discipline, il semble qu'aujourd'hui les objets linguistiques sur lesquels sont appliqués des considérations de fréquences ou des méthodes quantitatives sont d'une extrême diversité : tous les aspects de la langue se prêtent ou se sont prêtés, de fait, à des investigations fréquentielles.

15 Dans Martinet (1970 : 119) déjà le critère de la taille des ensembles était immédiatement suivis d'un critère de fréquence : « [la] fréquence moyenne de monèmes grammaticaux [...] est bien supérieure à celle de monèmes lexicaux [...] ».

2.2.1 Les phénomènes distingués par les différents niveaux de description ont tous fait l'objet d'étude quantitative : la phonologie, la prosodie (Obin *et al.* 2008), la morphologie (avec par exemple les travaux sur la productivité morphologique de Baayen 2009) ; la lexicologie (Blumenthal 2009) et la lexicographie contemporaine lorsqu'elle recourt à de larges corpus ; la syntaxe enfin mais dans une moindre mesure (avec par exemple les approches probabilistes Manning 2003).

2.2.2 Les domaines de la linguistique externe sont plus systématiquement engagés dans les études quantitatives : c'est le cas de la psycholinguistique, de l'analyse du discours ou de la lexicométrie. C'est également le cas de certaines disciplines plus interprétatives ou herméneutiques comme la linguistique textuelle.

2.2.3 L'analyse de la variation a partie liée avec une dimension quantitative. Dans le contexte du développement de la linguistique de corpus, Glessgen (2007 : 102) indique ainsi que « [la] distinction concrète des différentes variétés est plus le résultat d'une étude statistique qu'une réalité linguistique reconnue par les locuteurs ». Lodge (1994 : 25) indique de même que « [le] "cockney" n'existe nulle part à l'état pur. Il s'agit, comme d'habitude, de différences quantitatives dans la distribution de variables linguistiques clefs ». Pour certaines disciplines, comme la sociolinguistique, les méthodes quantitatives sont constitutives de l'approche théorique (Labov 2004 ; Macaulay 2009) ; pour d'autres, comme la dialectologie (Goebl 2003) ou la typologie (Cysouw 2003), elles font l'objet de développements méthodologiques récents.

2.2.4 Deux dimensions de recherche semblent importantes pour le développement des méthodes quantitatives et leur extension en dehors de leurs champs d'applications traditionnels : la grammaticalisation et la sociolinguistique historique. Dans le cadre de la grammaticalisation, le rôle de la fréquence est prédominant et régulièrement mentionné[16]. Dans la sociolinguistique historique, l'utilisation de méthodes quantitatives textuelles est nécessaire pour l'estimation de phénomènes variationnels dans des états de langue anciens.

2.3 Enfin, du point de vue de l'histoire de la discipline, il n'est pas possible d'assigner la prise en compte des faits de fréquence à une époque et, en particulier, il paraît difficile de considérer que la prise en compte du caractère central des faits de fréquence soit accentuée par le développement des corpus.

En effet, les considérations de fréquence sont, en premier lieu, caractéristiques de toutes les époques de la discipline. De très nombreux travaux, commencés dans les années 30 (Zipf, Troubetzoy) sont antérieurs à toute instrumentation. Les questions de fréquence ne semblent d'ailleurs pas avoir fait jadis l'objet d'un préjugé aussi négatif qu'aujourd'hui. Les questions essentielles concernant l'usage de données quantitatives ont été identifiées et discutées dès les années 50, comme le montre les références utilisées dans le présent article. En 1962, Heilmann (1972 : 30) pouvait déjà déclarer : « […] the frequency of occurrence of linguistic forms has been universally recognized as a basic aspect of language ».

16 Entre autres : « Quant au troisième critère [distingué par Traugott et Heine], la fréquence, il est apparemment le plus déterminant. Une unité lexicale qui n'est pas d'usage fréquent n'aura pas l'occasion d'évoluer sémantiquement et certainement pas de se grammaticaliser. » (Bat-Zeev Shyldkrot 1995 : 76)

Différentes traditions nationales (par exemple en Allemagne avec G. Altmann, en Angleterre autour de J. Sinclair, ou en France avec C. Muller), qui ont élaboré l'essentiel des notions et des méthodes qui influencent encore la linguistique de corpus dans ces différentes communautés, sont fondées sur des moyens techniques préalables à la généralisation de l'ordinateur. Il en est de même pour l'usage de méthodes quantitatives dans les enquêtes en linguistique variationniste ou en dialectologie (Labov 1976, Davis 1990).

Dans de nombreux contextes l'outillage contemporain n'apporte pas de nouveautés décisives : pour accéder à – ou plutôt pour construire – des faits de fréquence, la complexité est théorique bien plus que méthodologique et pratique. C'est le « comptage » qui nécessite des statistiques, plus que les statistiques ne nécessitent des capacités techniques de comptage, comme le rappel le dictum statistique selon lequel « serious counting require statistics » (cité dans Davis 1990 : 4). Il ne paraît pas certain que les chercheurs bénéficiant aujourd'hui des « grands corpus » bénéficient par ailleurs d'un cadre méthodologique et théorique leur permettant d'en tirer des résultats descriptifs nouveaux, qui auraient été inaccessibles aux générations antérieures. L'un des meilleurs exemples en est le domaine de la variation : depuis les travaux de Biber, souvent cité comme l'instigateur de cette méthodologie, des analyses contrastives de sous-corpus au moyen de méthodes factorielles ont été maintes et maintes fois reproduites avec tout type de partition en sous corpus et tout type de variables pour les caractériser. Mais on ne constate pas, plus de 20 ans après les premières de ces expériences, qu'elles aient apporté de progrès décisifs dans la description de la variation ni qu'elles se soient succédées d'une manière cumulative (cf. Loiseau 2008). Il semble cependant qu'on puisse les continuer indéfiniment et en tirer à chaque fois des caractérisations variées, sans progresser vers la définition ou l'élaboration de catégories (genre, registre...) qui restent largement sous-déterminées par les données. L'essentiel du cadre conceptuel et des catégories descriptives de l'analyse de la variation est né avant ces méthodes. Alors que les grands corpus informatisés semblent donner accès à un matériau empirique *a priori* d'un nouvel ordre de complexité, il n'y a pas eu, à ma connaissance, un renouvellement si important des catégories de l'analyse de la variation ni de progrès majeur sur le plan conceptuel et méthodologique. Le paradoxe est que des données plus riches déterminent un état moins cumulatif de la discipline. Les domaines qui semblent le plus contribuer au renouvellement de la problématique de la fréquence, comme les modèles usages-based, sont par ailleurs prudents dans leur approches des méthodes quantitatives et des grands corpus (eg. Bybee 2010 : 97).

De ce point de vue, il faut reconnaître que la question de la fréquence en linguistique n'est en rien fondamentalement renouvelée par la démocratisation des moyens de calculs quantitatifs. On peut appliquer au statut de la quantification dans la description ce que Laks (2008 : 12) dit des corpus : « Il importe de reconnaître que le rapport aux données descriptives regroupées en vastes ensembles stabilisées, publiques et partagées ne date ni de l'apparition de l'ordinateur, ni de celle des outils sophistiqués de traitement automatique de ces bases ». Cependant, il faut noter également que les grands corpus numériques renouvellent les observables que l'on peut construire et cela principalement grâce au recours aux méthodes quantitatives.

2.4 Sans prétendre faire un bilan doxographique exhaustif, on doit donc souligner que les considérations de fréquences sont présentes à tous les niveaux de la discipline et qu'il est difficile d'attribuer la prise en compte de la fréquence à un cadre théorique ou même un style analytique particulier. Elle n'est pas propre, par exemple, à une tradition dans la discipline

que l'on pourrait dire plus empirique, ou à une tradition qui serait plus attachée à la variation, etc. Comparée à la variété des recours à la notion de fréquence dans la littérature, la faible élaboration théorique de cette notion est donc paradoxale.

3. Spécificité du quantitatif en linguistique

3.0 Un autre ensemble de paradoxes de la fréquence en linguistique tient aux propriétés quantitatives des données langagières. Ces données ont plusieurs spécificités statistiques importantes du point de vue de leur interprétation ou de leur utilisation pour la description. La première de ces spécificités est d'avoir une double dimension objective et subjective comme nous l'avons relevé plus haut (1.2). D'autres spécificités intéressent particulièrement la mise en œuvre de méthodes quantitatives.

3.1 La première de ces spécificités est le caractère redondant ou stable des faits de fréquence dans leur relation les uns aux autres (cf. ci-dessus 2.1.3). Un phénomène ou une argumentation, dans un raisonnement qui s'appuie sur des données quantitatives, est souvent gagé sur le fait que les données « convergent » toutes dans le même sens, comme le note par exemple Labov (1976 : 304-305) :

> Un statisticien expérimenté verra aussitôt qu'il est dépourvu de pertinence de vérifier que de tels résultats [des phénomènes assourdissements dans différentes classes d'âge/groupes sociaux] diffèrent significativement entre eux. La présentation des données permet au besoin une analyse statistique, mais il est clair que, même si on reste en dessous du niveau de différence significative dans un cas particulier, une telle convergence de tant d'événements indépendants suffit à nous porter à un degré de sûreté que la plupart des recherches sociales ou psychologiques continuent d'ignorer.

De ce fait, les observations quantitatives ont souvent la double propriété d'être à la fois stables ou « robustes » et en même temps difficile à rapporter à des catégories descriptives, du fait d'être basées sur des approximations de phénomène qualifiés. Dans l'établissement des phénomènes, on gagne donc sur un plan ce que l'on perd sur l'autre.

3.2 Cette observation de Labov souligne une limite des tests d'hypothèses appliqués aux données langagières. Le principe des tests statistiques est d'aider à décider si un phénomène quantitatif est conforme au libre jeu du hasard ou si, au contraire, il s'écarte significativement du hasard et permet de postuler l'œuvre d'un déterminisme sous-jacent. Selon Labov, même si les fréquences observées s'écartent peu des fréquences attribuables au hasard, la répétition de ces écarts permet d'exclure l'hypothèse d'un libre jeu du hasard.

Une critique plus systématique des tests d'hypothèse est proposée par Kilgarrif (2005). L'objet de l'article est une critique générale de l'usage de certaines méthodes quantitatives en linguistique et des tests d'hypothèse en particulier. Les tests d'hypothèse consistent en effet à postuler que les données pourraient être réparties selon une certaine homogénéité (que c'est une hypothèse raisonnable), et à mesurer l'écart avec cette répartition homogène, de façon à décider quelle est l'hypothèse la plus probable : ou bien que le hasard seul est en cause pour expliquer les données, ou bien qu'un déterminisme est à l'œuvre. Ce sont des tests d'hypothèse qui sont utilisés, par exemple, pour mesurer les phénomènes d'attirance entre mots (qu'il s'agisse du chi2, de la loi hypergéométrique, ou du test de Fischer). Or, les données linguistiques ne sont jamais réparties de façon homogène et, comme le dit le titre de

l'article de Kilgarriff, « language is never, ever, ever random »[17]. L'auteur rapporte ainsi une expérience qui consiste à contraster un sous-corpus d'anglais américain et un sous-corpus d'anglais britannique. De nombreux phénomènes ont des fréquences qui varient dans une proportion qui ne peut être attribuée au hasard et ils peuvent donc être attribués aux différences des deux variétés. Faut-il cependant considérer que la différence des deux variétés est l'explication (la cause) de ces divergences ? En prenant deux ensembles de textes tirés *au hasard* du même corpus, et mêlant donc des textes des deux variétés, on obtient pourtant autant de différences (Kilgarriff 2005 : 268 sqq.) : *avec des données suffisamment nombreuses*, il y a toujours une irrégularité de distribution significative[18]. Appliqués aux phénomènes linguistiques, les tests d'hypothèse montrent donc *toujours* que les phénomènes linguistiques divergent significativement d'une répartition au hasard. La condition de l'utilisation de ces tests – que le hasard seul s'exerce si le phénomène observé (ici, l'opposition des variétés) n'a pas d'effet – n'est donc jamais remplie.

Un autre exemple peut être proposé à partir du cas de l'observation de la variation des fréquences d'un phénomène à différentes positions d'une unité. Si l'on relève les fréquences d'un phénomène à différentes positions d'un texte (par exemple dans chaque dixième d'un ensemble de textes d'un même genre), un test statistique peut servir à décider si cette variation des fréquences est significative, ou pas ; autrement dit, si, dans le cadre du genre considéré, le phénomène observé est significativement attiré par une position textuelle (Loiseau à paraître). Là encore, une démonstration par l'absurde montre le danger d'utiliser les tests statistiques : même un phénomène grammatical extrêmement peu susceptible de spécialisation, comme un grammème de très haute fréquence, exhibe des sous-fréquences qui sont, selon ces tests, significativement différentes d'une distribution régulière, due au seul hasard.

L'emploi des tests d'hypothèse reste cependant valide si l'on est moins intéressé par la stricte alternative (la fréquence observée s'explique-t-elle par le hasard ou pas ?) que par *l'interclassement* des différentes fréquences observées : quelles sont les phénomènes *les plus* attirés par l'une ou l'autre variété. L'*ordre* est (globalement) interprétable (quels lexèmes sont les plus remarquables ?), alors que la simple valeur dichotomique (significativité ou pas) ne l'est pas : c'est ce qu'observe également Muller (1992 : 114).

Kilgarriff montre que cette non pertinence des tests statistiques tient entre autre au *volume* des données manipulées dans les grands corpus. En effet, même si les statistiques ont pour objectif de permettre de tirer des conclusions relativement indépendantes de la taille des échantillons, on sait que celle-ci n'est pas entièrement neutre. Les méthodes statistiques insistent généralement sur le fait que les échantillons ne doivent pas être trop petits. La difficulté, dans le cas des données linguistiques, est plutôt qu'ils sont trop grands. Les phénomènes statistiques observés en corpus sont en effet souvent des phénomènes de très haute fréquence (du moins la taille du corpus, la fréquence de référence, est-elle élevée), ce qui pose également des limites à l'utilisabilité des tests statistiques.

3.3 Une autre donnée fondamentale de l'utilisation de fréquences linguistiques est le caractère rare de certains événements. En effet, même les corpus les plus (déraisonnablement)

17 Cf. une critique analogue dans (Bybee 2010 : 97) : « […] lexemes do not occur in corpora by pure chance. Every lexeme was chosen by a speaker in a particular context for a particular reason. »

18 « Given enough data, H0 [l'hypothèse que les phénomènes sont régulièrement réparties] is almost always rejected however arbitrary the data » (Kilgarriff 2005 : 268).

volumineux ne donnent pas accès à des fréquences élevés pour certains phénomènes fins : « Many syntactic phenomena, especially those commonly of interest to theoretical syntacticians, are just incredibly rare in text corpora. » (cf. Manning 2003 : 295 ; cf. aussi Gadet 2000 : 37) : dix ans du journal *Le Monde* par exemple n'ont pas donné assez d'occurrences de *quelque* et *plusieurs* pour contraster leur valeurs à travers leur cooccurrents ; Habert rapporte que dans un corpus du journal *Le Monde* de 14 millions de mots il n'y a pas d'occurrence de l'acception « trahir » parmi les 1 345 occurrences de *vendre*, etc. De fait, la loi de Zipf implique qu'un corpus est majoritairement composé de phénomènes rares. Enfin et surtout, les phénomènes observables et leur quantification sont, naturellement, entièrement fonction des variétés ou des genres représentés dans le corpus, ce qui fait en définitive de chaque corpus un objet singulier difficilement commensurable aux autres.

3.4 En effet, les spécificités du quantitatif en linguistique tiennent également à l'extrême variété des normes linguistiques (dans un sens très général : niveaux de description, axes de variation, genre, etc.). Chaque corpus permet d'observer un entrecroisement singulier de ces normes. De ce fait, les faits quantitatifs ne permettent d'observer que le résultat d'une interaction complexe entre ces normes, ils mélangent ce que les cadres descriptifs distinguent, et il est donc particulièrement difficile de les rapporter à des catégories descriptives. C'est l'une des raisons pour lesquelles les faits quantitatifs excèdent souvent les catégories descriptives qui les prennent en charge : les mêmes phénomènes d'attirances entre mots par exemple sont utilisés aussi bien dans des perspective lexicologiques (Blumenthal 2006) que discursives (Lafon 1980) ou pour décrire des « constructions » (Stefanowitsch/Gries 2003) dans le cadre des grammaires de construction.
Dans la version la plus pessimiste, cette intrication des phénomènes conduit à voir dans tout phénomène quantitatif le résultat arbitraire d'un ensemble de déterminations impossibles à identifier. Ullmann (1951 : 294) conclut ainsi : « In language, quantity can never eclipse quality. All we may hope to achieve is, under exceptionally favourable circumstances, to discern a broad pattern behind the countless influences and accidents concealing and distorting it. Nor should it be forgotten that the individual act of speech, the only concrete realisation of the language system, is essentially indeterminate […]. »[19] Pourtant, si les catégories d'analyse, c'est-à-dire le qualitatif, est premier, les faits quantitatifs ne sont pas sans signification et les actes individuels ne peuvent pas être considérés comme entièrement indéterminés.

La reconnaissance de cette sous-détermination de la théorie par les faits fréquentiels est importante pour l'interprétation de ces faits de fréquence. Dans le cas de Stefanowitsch et Gries par exemple, les mesures d'association entre mots utilisées font naturellement rencontrer des phénomènes culturels qui excèdent de façon particulièrement problématique les catégories descriptives initialement posées. À l'issue d'une étude de la construction causative en *into*, Wulff/ Stefanowitsch/Gries (2007 : 279) croient pouvoir conclure que « […] the contrast between movement-initializing cause predicates in British English as opposed to movement-restricting cause predicates in American English may confirm the commonplace perception that British culture lacks the strong and explicit emphasis on mobility as an

19 Cf. également (Larochette 1981 : 139) : « Il y a diverses normes dont les isoglosses s'entrecroisent; si elles ne sont pas reconnues *avant* toute statistique, non seulement celle-ci gommera les isoglosses et ne distinguera plus la norme générale, les normes restreintes, les normes individuelles, mais elle ne distinguera plus les déviations de toutes ces normes. » ; (Stubbs 2001 : 224) « Frequency counts from large corpora may well mean that the researcher does not notice uses which cluster in particular genre. »

essential condition for a happy and free life as we find it in American culture. » La sur-interprétation des données est ici patente et donne libre cours aux plus éculés des clichés culturels.

3.5 La question plus généralement posée est celle du codage, c'est-à-dire de la décision de tenir pour identiques certains phénomènes (Desrosières 1989 : 227) :

> un codage est une décision conventionnelle de construire une classe d'équivalence entre divers objets, la « classe » étant jugée plus « générale » que tout objet singulier. La première condition pour cela est de supposer que ces objets peuvent être comparés, ce qui ne va pas de soi. [...] La plus ancienne tradition de ce problème taxinomique vient sans doute de la jurisprudence : juger et coder reviennent tous deux à classer un *cas* dans une catégorie légale, ou, comme le dise les juristes, à le *qualifier*. Ainsi apparaît le lien étroit entre les dimensions politiques et cognitives du codage : la catégorie légale réfère à la fois au roi et au savoir.

Les choix de codage sont des choix interprétatifs : « La transcription n'est pas simplement une activité sélective, mais plus radicalement une entreprise interprétative » (Mondada 2000 : 131). Cette activité interprétative impliquée dans les choix de codage renvoie à la double dimension des statistiques, indissociablement moyen de connaître et moyen d'agir sur les objets étudiés.

Dans le domaine linguistique, les conventions de codage sont un enjeu central : elles décident particulièrement des observables qui sont rendus disponibles. Mondada (2000) montre les enjeux attachés à tous les actes de transcriptions de l'oral, jusqu'à la répartition dans l'espace de la page, qui incorporent des interprétations ou des choix théoriques. Elle oppose deux stratégies de codage : l'une issue des modèles théoriques et analytique, préexistant aux données ; l'autre dont la pertinence est fondée sur une analyse du corpus et des locuteurs et des catégories qu'ils rendent pertinentes.

> Alors que la première question renvoie à une approche exogène, étique de la transcription et de son analyse, la seconde renvoie à un traitement endogène, émique, considérant les orientations des participants vers tel ou tel phénomène.
> Si la première permet un codage des transcriptions en vue de relevés quantitatifs, ainsi que le traitement des exemples comme confirmation d'hypothèses formulées ailleurs, la seconde est plutôt orientée vers la « découverte » de phénomènes dont la pertinence n'est pas affirmée a priori. Les deux types de transcription n'échappent pas à leur dimension indexicale et interprétative, mais la seconde décide de suspendre certains choix de transcription ou de privilégier une notation relevant de la transcription plutôt que de la description, en considérant qu'il revient à l'analyse de tirer certaines conclusions.

Cette alternative se rencontre dans d'autres sciences humaines (outre, bien sûr, la sociologie, on peut mentionner le champ de l'histoire : cf. Lemercier/Zalc 2008 : 14-15 pour l'opposition entre micro-histoire et histoire quantitative, « labroussienne »).

Dès 1981 Fénelon note (1981 : 110) : « Les opérations de codage et de recodage constituent les 9/10e du travail réel Analyse des Données, et les 999/1000e de l'arbitraire des résultats. » La question des enjeux des choix de codage a été particulièrement posée par la lexicométrie : la démarche explicite est le pari selon lequel une perte dans la finesse de la catégorisation est compensée par ce que la quantification des données ainsi codées permet d'observer (Lafon 1981 : 99-100) :

> [...] le dénombrement des cooccurrences n'est pas une opération innocente. Il met parfois implicitement au même niveau des phénomènes de nature assez différente, qu'il faudrait peut-être distinguer. Ainsi, il nous arrivera quelquefois d'additionner des poires avec des pommes. [...] la démarche lexicométrique [...] consiste en ceci : quantifier l'objet d'étude au prix de certaines réductions délibérément acceptées, car nous pensons qu'elles n'affectent pas les faits lexicaux les plus massifs, mais, en contrepartie, disposer de l'apport de l'automatisation et du calcul pour explorer exhaustivement le texte et hiérarchiser les résultats obtenus afin de n'en retenir que les plus remarquables.

En résumé, la problématique du codage des données en linguistique rencontre deux ordres de difficulté. Le premier est que les choix de codage, aussi préconçus, fermés et simplificateurs soient-ils, restent relatifs à un corpus : c'est le caractère commensurable des résultats de différentes analyses qui fait défaut. Autrement dit, c'est la difficulté à faire, selon l'expression de Desrosières, « des choses qui tiennent ». La sociologie a établi, grâce à une procédure de construction de classes d'équivalence, des objets comme les catégories sociaux-professionnelles qui, même si elles sont en permanence remises en question – et, peut-être, grâce à cela – sont néanmoins une construction commune de la discipline, sur laquelle peuvent s'appuyer quantification et description, et qui rend les travaux comparables. En linguistique, on ne peut pas dire que le gain heuristique de la quantification ait permis de convenir, même temporairement ou « par provision », d'objets sur lesquels appuyer des descriptions ultérieures.

Le second est le coût de la « simplification » inhérente à un codage simplificateur. Lafon à nouveau (1981 : 132) exprime bien les conséquences des codages trop grossiers et de l'observation superficielle des interactions de normes dont nous parlions ci-dessus (3.4) : « Les associations retenues [les cooccurrences] sont, tout comme les phénomènes textuels qui les engendrent, d'une extrême diversité. En outre, comme tous les traitements statistiques exhaustifs de cette nature, celui-ci fournit un certain nombre d'associations dont la signification demeure obscure : nous ne savons pas leur attribuer une interprétation satisfaisante. » Les conséquences des simplifications méthodologiques sont de rendre ininterprétables les faits produits.

3.6 Un dernière spécificité des données langagières est la difficulté à mettre en œuvre l'un des principes cardinaux du raisonnement statistique, à savoir l'échantillonnage. Tout résultat statistique est censé permettre une généralisation contrôlée depuis l'échantillon étudié vers l'« ensemble parent » dont cet échantillon est issu. Or, dans le cas des données langagières, la relation entre l'échantillon soumis aux méthodes quantitatives et un ensemble qui serait représenté est très difficile à établir.

3.6.1 La version la plus « ambitieuse » de la question de l'échantillonnage consiste à assimiler les notions statistiques d'« échantillon » et d'« ensemble parent » aux notions linguistiques de « parole » et « langue » : tout texte attesté est un « échantillon » de la langue. L'articulation parole/langue (cf. 1.2) est ainsi définie comme purement quantitative et statistique. C'est la position de Muller (1992 : 14) : « si nous introduisons la distinction classique entre langue et discours, entre la virtualité et l'actualisation, nous serons amenés à considérer tout discours comme une réalisation, comme un échantillon de la langue de son auteur ». C'est également la position de Herdan (1966 : 28) « "la parole" […] appears […] as a term for statistical samples withdrawn from "la langue" as the population ». De ce point de vue, les mots auraient une fréquence « en langue », la fréquence serait un attribut du mot, que l'on pourrait inférer à

partir de l'observation d'échantillon : « La fréquence est un attribut positif et concret du mot et fait partie de sa définition. » (Guiraud 1959 : 29) ; « what " la langue " comprises are not only engrams as lexical forms, but these engrams plus their respective probabilities of occurrence. » (Herdan 1966 : 27). On sait que l'hypothèse de l'existence de « fréquences en langue » a été abandonnée en faveur de l'idée que les fréquences ne peuvent être rapportées qu'à une « norme » (Lafon 1980 : 127)[20], à des « usages » (Tournier 1980 : 194), « en entendant par là des systèmes d'habitudes qui gèrent les emplois des mots en situation » ou encore à des « facteurs thématiques et stylistiques » (Heger 1969 : 56). Ce qui est abandonné avec l'idée de fréquence en langue c'est une application « littérale » de certaines des formules saussuriennes qui font de la langue la somme des actes de paroles : « La langue […] n'est jamais une somme de paroles, même hétérogènes à souhait. » (Tournier 1980 : 192).

3.6.2 La discussion sur les questions de la représentativité des corpus a été particulièrement développée dans la tradition britannique dite « contextualiste ». La relation fondamentale reste pourtant, la encore, celle d'une généralisation statistique d'un échantillon à une population parente.

Pour Tognini-Bonelli par exemple, un corpus doit être représentatif pour permettre une « généralisation ». Un corpus est une « population » dont l'analyse doit être applicable « to a larger sample or to the language as a whole » (Tognini-Bonelli 2001 : 57). Dans la perspective des corpus « équilibrés », un corpus est représentatif s'il a une hétérogénéité interne qui lui fait représenter la langue ou la variété visée : « Biber […] defines representativeness as « the extend to which a sample includes the full range of variability in a population » (Tognini-Bonelli 2001 : 59). La méthodologie proposée pour cela, suivant Biber (1993), est de structurer le corpus en différents *registers*, c'est-à-dire différents ensembles définis par la situation de communication (la référence à des paramètres extra-linguistique est nécessaire pour éviter tout cercle vicieux). Ce corpus, une fois constitué, permet d'étudier des types de textes, c'est-à-dire des regroupements de textes fonctionnellement proches du point de vue de propriétés linguistiques. La notion de représentativité d'une langue est donc abandonnée : « […] the concept of a representative sample of the English language makes little sense. […] A sample can be representative only if the population to be sampled is homogeneous […] » (Stubbs 2001 : 223). La représentativité n'a un sens que pour un genre (Stubbs 2001 : 224, cf. aussi 120-121). Pourtant les difficultés restent sensiblement les mêmes : comme nous l'avons vu, la présence de plusieurs normes, de plusieurs axes de variation, dans tout corpus, rend très théorique la notion d'homogénéité. De plus, la difficulté reste entière de définir quelle est le genre visé, la « target population the corpus aims to represent » (Tognini-Bonelli 2001 : 59).

En définitive, il est assez paradoxal d'observer combien les différentes présentations des corpus persistent à défendre la notion de représentativité tout en s'avouant incapables de la définir précisément. En effet, la difficulté à définir des règles pratiques claires conduit Tognini-Bonelli à reconnaître qu'il s'agit essentiellement d'un vœux pieu (« […] representativeness must be regarded largely as an act of faith [...] » (2001 : 139)). La discussion de cette notion chez Tognini-Bonelli se conclut régulièrement sur l'impératif de documenter ses choix, ce qui est certes essentiel, mais sans lien avec la question de la représentativité (« It is imperative to be explicit about how a corpus is constructed » (2001 : 88) ; « worker in the field should be as explicit as possible » (2001 : 62)). Gries (2009 : 8) de même finit par reconnaître benoîtement et en toute naïveté : « In sum, balanced corpora are a

20 Il s'agit ici d'une norme statistique un corpus de référence.

theoretical ideal that corpus compilers constantly bear in mind, but the ultimate and exact way of compling a balanced corpus has remained mysterious so far ».

S'il s'avère finalement impossible de dire en quoi consiste un corpus équilibré et représentatif, peut-être est-ce la notion elle-même qui est inadéquate ou non opératoire. Il faut bien reconnaître que cette notion, admise comme une évidence, n'a débouché sur aucun résultat tangible, tels que des critères concrets de constitution de corpus ou d'interprétation des résultats, ou une cumulativité des résultats descriptifs. L'insistance à préserver la notion de représentativité semble une tentative de préserver un modèle scientifique finalement assez positiviste, où la généralisation (l'abstraction) est le produit d'un simple calcul. Plusieurs difficultés peuvent être identifiées qui montrent qu'on ne peut parler de représentativité dans le sens statistique courant, sinon par métaphore. Il me semble qu'on peut distinguer deux types de difficultés qui limitent l'utilité du modèle statistique pour encadrer l'usage de corpus. Le premier tient aux spécificités des données langagières, le second à la nature des catégories (langue, genre, etc.) que l'on veut représenter par leur biais.

1/ tout d'abord la question est posée comme si des notions linguistiques pouvaient correspondre directement aux notions statistiques (« population », « échantillon », « généralisation »). Or, si la question de la représentativité est intrinsèquement liée à l'usage des statistiques, elle ne peut être posée sans réflexion sur les spécificités du domaine. On ne peut dire en effet qu'un corpus est un « échantillon » ou « représente » tel fait de langue au sens où un échantillon de tulipes peut représenter toutes les tulipes cultivées dans le même champ (ou l'espèce tulipe). Dans les données langagières, comme on l'a vu ci-dessus, il y a toujours un grand nombre de normes en interactions : l'« échantillon » est peut-être homogène du point de vue, par exemple, de la variété diastratique, mais il ne représente pas cette variété puisqu'il réalise en même temps une interaction singulière de différentes normes : « le champ d'application du diatopique, du diastratique et du diaphasique ne peuvent facilement être distingués les uns des autres, parce qu'ils sont totalement imbriqués et interdépendants dans les pratiques des locuteurs. » (Gadet 2003 : 99)[21]. Même à l'intérieur d'une même langue fonctionnelle (synchronique, syntopique, synstratique...), on ne voit pas pourquoi la création de corpus « homogène » serait plus simple : il y a là encore différentes normes en interactions (idiolectes, genres, discours ; cf. Loiseau 2010). La variation ne peut relever d'une classification dans un ensemble de classes disjointes (la classification en *registers* proposée par Biber par exemple).

Par ailleurs, le caractère définitoire d'un d'échantillon est d'être identique à la population parente *modulo* la taille. Or, comme le montre par exemple Baayen (2001) avec la notion de « Large Number or Rare Event », la taille est difficilement neutralisable avec des données langagières. Prolongeant la loi de Zipf, la notion de Large Number of Rare Event formalise l'idée que les faits langagiers contiennent un très grand nombre de phénomènes de très faible fréquence. Une augmentation de la taille d'un corpus change donc toujours les phénomènes qu'on peut y observer, même si le corpus est déjà de taille élevée. Augmenter la taille change la nature des données. C'est la relation échantillon / ensemble parent qui est compromise. De plus, comme nous l'avons vu (3.3) alors que le moindre corpus (par exemple, 100 000 mots, l'équivalent d'un livre de poche) permet d'observer un très grand nombre d'occurrences des

21 Cf. également Manning (2003 : 295) : « There is no easy answer to the problem of getting sufficient data of just the right type: language changes across time, space, social class, method of elicitation, etc. There is no way that we can collect a huge quantity of data (or at least a collection dense in the phenomenon of current interest) unless we are temporarily prepared to ride roughshod over at least one of these dimensions of variation. »

phénomènes les plus fréquents (de plusieurs ordre de grandeur supérieur aux données que l'on peut collecter par enquête), les phénomènes rares restent toujours rares, et donc difficiles à caractériser quantitativement, même dans des corpus de taille déraisonnable. La représentativité n'est donc pas la même pour les différents phénomènes que contient un corpus.

2/ D'autre part, la « généralisation » est présentée comme un procédé identique qu'il s'agisse de généraliser à un ensemble plus vaste ou à la langue elle-même (« to a larger sample or to the language as a whole », Tognini-Bonelli 2001 : 57). On peut douter que la méthode soit la même pour ces deux types de « généralisation », c'est-à-dire qu'on généralise de la même façon ou au même sens vers un corpus plus vaste, vers la langue, ou encore vers un genre. C'est la notion de « population parente » qui est maintenant en cause : on ne sait jamais de quel ensemble (population parente) un corpus est représentatif et ils n'entretiennent pas un rapport « échantillon/ensemble parent » parce que cet ensemble parent est une *abstraction* (Koch/Oesterreicher 2001 : 602, Coseriu 1982 : 61), jamais un fait empirique au même sens que le corpus. Alors que l'ensemble des tulipes d'un champ (ou l'ensemble de toutes les tulipes) est un ensemble réel, une langue (ou un genre) n'est pas un ensemble mais une abstraction ; or, on ne voit pas comment un ensemble pourrait être un sous-ensemble d'une abstraction. L'ensemble qu'il s'agit de représenter n'est pas donné mais construit.

À travers cette notion de représentativité, la question posée est celle de la nature de l'opération d'abstraction : les faits observés dans le corpus doivent-ils faire l'objet d'une pure transposition à un sur-ensemble, ou faire l'objet d'une interprétation ? La notion de représentativité retire toute dimension interprétative, contextualisante, historique, à l'opération d'abstraction.

Il n'est donc pas certain que la relation du corpus aux catégories décrites soit différente dans le cas d'une description à base quantitative que dans le cas de toute description « qualitative » qui, à partir d'un corpus, énonce des régularités selon des modes d'abstraction qui ne doivent rien à la généralisation au sens statistique. De même, le moment de la généralisation (à un genre, ou à des faits de langue) dans l'utilisation de méthode quantitative sur un corpus est peut être un moment interprétatif plutôt qu'une généralisation au sens statistique.

A contrario, même si un corpus est manifestement peu représentatif d'une norme, rien n'empêche d'en tirer des enseignements sur cette norme. Par exemple, dans Loiseau (2007) j'ai proposé une description d'un corpus de discours philosophique représentant principalement l'œuvre de G. Deleuze ; ce corpus n'est représentatif en rien du discours philosophique, ni même du discours philosophique en France dans les années 50-70, mais il est évident qu'on peut en tirer des enseignements sur ce discours. Si, par exemple, les concepts philosophiques apparaissent dans ce corpus comme partageant de nombreuses propriétés avec les thèmes littéraires, c'est un résultat qui intéresse le discours philosophique : le corpus étudié est au moins une réalisation que le discours philosophique rend possible ; il paraît relativement gratuit de supposer que ces propriétés des concepts philosophiques soient strictement spécifique au corpus étudié.

4. Répétition et norme

4.0 Dans cette partie, je souligne l'intérêt de la notion de norme pour la prise en compte des faits de fréquence dans la description. La notion de norme est commune à de nombreuses disciplines des sciences humaines et doit être replacée dans ce contexte. Son intérêt est en

particulier de permettre de prendre en compte les faits de fréquence (1). En linguistique, la *norme* de Coseriu et l'*usage* des contextualistes britanniques sont deux notions parmi les plus élaborées permettant d'articuler les faits de fréquence à des catégories descriptives (2). Enfin, je comparerai à la notion de norme cosérienne le modèle « usage-based » de la linguistique cognitive qui développe également une prise en charge des faits de fréquence (3). Dans ce dernier modèle, la fréquence est rapportée à des propriétés cognitives universelles plutôt qu'à la dimension historique de l'activité de parler.

4.1.1 La notion de norme est commune à de nombreuses sciences humaines. En France, c'est le plus souvent Canguilhem (2006) qui est cité pour l'illustrer. Canguilhem propose la notion de norme pour articuler faits de fréquence et catégories interprétatives. Il s'agit de rendre compte du fait que certaines caractéristiques anthropomorphiques sont particulièrement fréquentes. Par exemple, les membres d'une population ont une taille ou une durée de vie typique, « normale ». Ces caractéristiques sont à la fois naturelles et sociales, dans la mesure où un « genre de vie », un ensemble de choix sociaux conditionnent ces propriétés. Il montre en effet (2006 : 102) que, derrière les moyennes physiologiques des populations, qui peuvent passer pour un fait biologique, c'est une normativité sociale qui est à l'œuvre :

> S'il est vrai que le corps humain est en un sens un produit de l'activité sociale, il n'est pas absurde de supposer que la constance de certains traits, révélée par une moyenne, dépend de la fidélité consciente ou inconsciente à certaines normes de la vie. Par suite, dans l'espèce humaine, la fréquence statistique ne traduit pas seulement une normativité vitale mais une normativité sociale. Un trait humain ne serait pas normal parce que fréquent mais fréquent parce que normal.

D'une part, Canguilhem insiste donc sur le sens dans lequel doit procéder l'explication : la fréquence en elle-même, dans les faits sociaux, n'est pas une explication ni un résultat ; elle doit être rapportée à une normativité (une catégorie descriptive), même si c'est cette fréquence qui permet d'objectiver cette normativité (« la norme ne se déduit pas de la moyenne, mais se traduit dans la moyenne » 2006 : 104). Cette position est poursuivie par Bachelard (cf. Loiseau 2008).

D'autre part, à l'inversion du rapport explicatif entre fréquence et catégories descriptives s'ajoute une inversion de l'ordre des déterminations : le social conditionne des caractéristiques qui pourraient être tenues pour « naturelles » (« La durée de vie moyenne n'est pas la durée de vie biologiquement normale, mais elle est en un sens la durée de vie socialement normative. » 2006 : 104).

La normativité, en quelque sorte, est l'analogue de la causalité dans les domaines où interviennent des dimensions sociales et historiques, où seules s'observent des régularités, et non des faits entièrement prédictibles.

Soulignons enfin que la question centrale de la norme, l'articulation d'une dimension objective et d'une dimension subjective des faits sociaux, est intéressante pour prendre en charge la double dimension de la fréquence (fréquence objective et fréquence subjective). Dans une certaine mesure, ce sont les catégories subjectives qui expliquent les catégories objectives.

4.1.2 La notion de norme a de nombreuses déclinaisons dans différentes sciences humaines – on peut penser à la notion d'habitus chez Bourdieu par exemple. Celle-ci peut s'observer à travers des faits quantitatifs, qu'elle contribue à rendre intelligibles ; elle permet de

« dénaturaliser » l'analyse du comportement et son inscription dans les corps ; à travers la notion d'incorporation, elle permet de rendre compte de la subjectivisation des valeurs.

Le fait que la pure fréquence ne soit pas un élément explicatif est également fréquemment souligné. Auroux indique ainsi (1998 : 240) :« Si vous observez une régularité vous laissez entier le problème de l'explication des phénomènes. Or le recours à la norme est une façon d'expliquer les régularités ; comme le note Canguilhem, "la norme ne se déduit pas de la moyenne, mais elle se traduit dans la moyenne" [...]. ». Ou, dans une perspective variationniste, Gadet (2003 : 105) : « On est donc amenés à mieux réfléchir sur la notion d'explication, étant entendu que celle-ci ne saurait résider dans une mise en relation de deux ordres de faits, de type causal, ce que font les modèles qui quantifient des corrélations entre des objets statistiquement évalués. »

4.2.1 C'est surtout chez Coseriu que la notion de norme a reçu en linguistique un développement important. À partir de la dichotomie langue/parole, il propose de distinguer deux éléments dans la langue : le *système normal* et le *système fonctionnel*. Dans les actes de paroles concrets, « […] hay elementos que no son *únicos* u *ocasionales*, sino *sociales*, es decir, *normales* y *repetidos* en el hablar de una comunidad, y que, sin embargo, no pertenecen al sistema functional de la formas lingüísticas […] » (1982 : 55-56). C'est le système normal (ou *norme*). Un second ensemble de faits, à un degré d'abstraction supérieur, est constitué des éléments fonctionnels : ils constituent le « système fonctionnel » (ou *système*) (2001 : 205) :

> La *norme* comprend tout ce qui, dans la 'technique du discours', n'est pas nécessairement fonctionnel (distinctif), mais qui est tout de même traditionnellement (socialement) fixé, qui est usage commun et courant de la communauté linguistique. Le système, par contre, comprend tout ce qui est objectivement fonctionnel (distinctif).

Les faits de normes sont seulement « traditionnels », tandis que les faits de systèmes sont « fonctionnels ». Autrement dit, le non respect d'un fait de norme ne compromet pas l'intelligibilité, mais fait reconnaître le discours comme « anormal », non conforme à la façon de parler d'une communauté ; le respect du système, au contraire, a un enjeu communicatif (1986 : 322, 2007a : IV 6).

Dans cette conception, norme et système fonctionnel sont distingués comme deux niveaux d'abstraction et de formalisation ; ce sont deux degrés d'abstraction possible à partir du discours ou actes de paroles concrets. Ils sont toujours rapportés à un objet unitaire et premier, l'activité de parler : « Vale decir que le *sistema* y la *norma* no son realidades autónomas y opuestas al hablar y tampoco « aspectos del hablar », que es una realidad unitaria y homogénea, sino *formas* que se comprueban en el mismo hablar, abstracciones que se elaboran sobre la base de la actividad lingüística concreta, en relatión con los modelos que elle utiliza. » (1982 : 94-95) Norme comme système sont des isoglosses, c'est-à-dire des faits communs aux discours[22] (1982 : 61-62) :

22 « […] en el acto lingüístico se comprueban los llamados « hechos de lengua » […], es decir, isoglosas entre el acto considerado y actos lingüísticos anteriores, del mismo individuo o de otros individuos, que se han tomado como modelo. » (1982 : 57) ; « *sistema de isoglosas* (aspectos comunes comprobados en los actos considerados) » (1982 : 92) ; « [...] el concepto de *lengua* no encuentra su justificación en la visión retrospectiva desde el acto lingüístico y en la formalización « en profundidad » de ese mismo acto, sino, más bien, en generalización que se establece « en amplitud » sobre la base de una serie de actos lingüísticos, abarcando los aspectos comunes que en ellos se comprueban. » (1982 : 102) .

> En todas las analogías aducidas pueden distinguirse siempre tres series de características, según el grado de abstracción o formalizatión: 1) las características concretas, infinitamente variadas y variables, de los objetos observados; 2) las características normales, comunes y más o menos constantes, independientemente de la función específica de los objetos (primer grado de abstracción); 3) las características indispensables, es decir, functionales (segundo grado de abstracción).

Coseriu donne des exemples relevant de tous les niveaux de description (1982 : 70 sqq.). Distinguons, dans les exemples qu'il propose, trois types différents de relation entre le système fonctionnel et le système normal.

Un premier type possible d'articulation est le cas où la norme spécifie ce que le système laisse indéterminé. Prenons l'exemple de la neutralisation en finale de l'opposition entre sourdes et sonores en russe. Puisque l'opposition est neutralisée, la réalisation peut être sourde ou sonore sans conséquence fonctionnelle ; cependant elle est toujours réalisée sous la forme de sourde : c'est un fait de norme[23].

Un autre type de relation consiste à opposer les trois degrés d'abstraction que sont le système, la norme et le discours du point de vue de la variabilité des phénomènes. Les trois degrés d'abstraction s'ordonnent selon le nombre d'unités qu'ils font distinguer. « Tenemos, por consiguiente, un único fonema /*o*/ en el sistema, dos variantes típicas, dos tipos de *o*, en la norma y, finalmente, una infinidad de realizaciones distintas (variantes individuales y ocasionales) en el hablar concreto, en los actos lingüísticos » (1982 : 72). En morphologie (1982 : 75 sqq.), les exemples concernent principalement des construits ou des dérivés possibles mais non existants, c'est-à-dire des analogies fautives, notamment par des apprenants. En syntaxe, il s'agit de tous les aspects phraséologiques et des figements, de l'ordre normal ou préférentiel des lexèmes (1982 : 80 sqq. ; cf 2001 : 248).
Les faits de normes relèvent enfin d'aspects fréquentiels ; par exemple, la fréquence d'un phonème peut être « anormale » (1982 : 71) :

> El fonema /*x*/ (en la grafía corriente *j*, o *g*, delante de *e*, *i*) es un elemento común del sistema fonológico español; sin embargo, una frase como *Artajo trajo la valija abajo* produce un extraño efecto « estilístico », porque la *frecuencia relativa* del fonema es mucho menor en la norma española.

4.2.2 Une dimension particulièrement intéressante de l'élaboration de la notion de norme est son inscription dans des phénomènes de fréquence et de répétition. La norme est ainsi faite d'éléments « no accesorias y esporádicas » (1982 : 89) ; « simple tradición constante, elemento común » (1982 : 96) ; « normal, repetido en una comunidad » (1982 : 57) ; « *normales* y *repetidos* en el hablar de una comunidad » (1982 : 55) ; « elementos normales y constantes » (1982 : 69) ; « repetición de un modelo anterior » (1982 : 57 et *passim*) ; « traditionnel et "usuel" » (1987 : 53-54), etc. Comme dans toutes les sciences humaines, la normativité se traduit quantitativement par des fréquences remarquables. Les faits normaux sont des faits fréquents.

L'importance de la norme pour la description des propriétés fréquentielles des faits linguistiques est ainsi soulignée par Coseriu (1982 : 71−72) :

> Consideramos, justamente, que todo lo que se refiere a la frecuencia de los fonemas en una lengua, todos los hechos de estadística fonológica, conciernen a la *norma* y no al *sistema*; en efecto, se trata de

23 « […] pero la realización de los fonemas correlativos implicados (/b/-/p/, /d/-/t/, etcétera) no es de ninguna manera indiferente desde el punto de vista de la norma, puesto que ellos se realizan siempre como sordos. » (1982 : 66)

> hechos que caracterizan una lengua, pero no pertenecen al conjunto de sus intrínsecas oposiciones fundamentales.

Dans cette perspective, la norme est donc le lieu de la prise en charge de la dimension première, originelle, de la fréquence : la dimension de la répétition. La référence aux statistiques dans la définition de la norme par Coseriu a été plusieurs fois relevée (cf. Lara 1983, Heger 1969 : 55). Elle rend compte, par exemple, des phénomènes collocatifs ou de la fréquence relative entre paires de quasi-synonymes (2001/1966 : 248), et relevant partiellement d'une description quantitative.

La distinction de la norme et du système permet également d'aborder, d'un point de vue proche de celui déjà souligné chez Martinet, la notion d'équilibre du système, c'est-à-dire les influences des faits quantitatifs sur le système (1982 : 106)[24] :

> El estudio estadístico, estudio cuantitativo de la norma, adquiere cada vez más importancia, pues la norma representa el equilibrio de un sistema en un momento dado, y los cambios cuantitativos suelen llevar a cambios cualitativos: los cambios en la norma llevan a cambios en el sistema.

L'enjeu de l'articulation de la norme et du système est de pouvoir choisir le second terme de l'alternative posée par Heilmann (1983 : 218) : « In altre parole, nel linguaggio gli aspetti qualitativo e quantitativo sono indipendenti, ovvero tra numero e funzione esistono rapporti determinati? »

Les faits de fréquence sont rapportés à la dimension de répétition primaire de l'activité de parler. C'est dans la caractéristique de la parole d'être hautement répétitive que sont ancrés les phénomènes de fréquence. Deux notions importantes à cet égard sont celles de « tradition » et de « langue antérieure », c'est-à-dire la part de réutilisation et de la répétition des actes de parole. Dans le cadre d'une conception unitaire l'activité de parler est une activité de reconfiguration et de resystématisation : l'activité de parler est création et non seulement usage (Coseriu 1982 : 94)[25] :

> [los actos lingüísticos] son […] – por la misma condición esencial del lenguaje, que es la comunicación –, actos de re-creación; no son invenciones *ex novo* y totalmente arbitrarias del individuo hablante, sino que se estructuran sobre modelos precedentes, a los que los nuevos actos contienen y, al mismo tiempo, superan.

L'étude des aspects de fréquence des faits linguistiques est donc fondée dans cette propriété fondamentale du langage d'être produit et configuré par la répétition dans l'activité de parler. La notion de norme est liée à la dimension d'historicité de la langue et les phénomènes quantitatifs sont fondés dans cette dimension d'historicité. L'activité de parler a une propriété particulière, celle d'être hautement répétitive : à l'échelle d'un idiolecte, ces « moules »

24 Cf. également « [...] le système représente la dynamicité de la langue, sa façon de se faire, et, par conséquent, sa possibilité d'aller plus loin que ce qui a déjà été réalisé ; la norme, en revanche, correspond à la fixation de la langue en modèles traditionnels ; et en ce sens, précisément, la norme représente à tout moment l'équilibre synchronique (« externe » et « interne ») du système. » (2007a : II, 3.1.3)

25 Cf. également (Coseriu 2007a : III 5.1) « La langue se refait parce que l'activité de parler se fonde sur des modèles antérieurs et elle est parler-et-comprendre ; elle est dépassée par l'activité linguistique parce que l'acte de parler est toujours nouveau ; elle est rénovée parce que comprendre est toujours comprendre au-delà de ce qui est déjà su au moyen de la langue antérieure à l'acte. La langue réelle et historique est dynamique parce que l'activité linguistique ne consiste pas à parler et comprendre une langue, mais parler et comprendre quelque chose de nouveau par l'intermédiaire d'une langue. »

répétés peuvent atteindre une proportion extrêmement élevée du discours[26]. L'activité de parler est sans doute l'activité qui repose le plus fortement sur cette dimension de répétition.

Parole, norme et système peuvent être vus comme correspondant à trois types d'historicité, et rapportés aux trois types d'historicité distingués dans (Kabatek 2005 : 151 sqq.). Le premier type d'historicité correspond à la langue : « La langue en tant que langue particulière est l'histoire d'une communauté intériorisée dans l'individu. » (2005 : 151). Le second type d'historicité correspond à la dimension de répétition et correspond peut-être particulièrement au niveau de la norme (2005 : 152) :

> Le second type d'historicité, en revanche, concerne tout type de phénomène culturel, voire langagier. Cela implique les traditions au sein d'une communauté, la création récurrente d'objets culturels à partir de certaines similarités ou de l'identité partielle de phénomènes culturels antécédents. Il s'agit donc d'objets culturels dont l'actualisation, qui ne correspond jamais exactement à la tradition, est disponible à la communauté. La langue en tant qu'objet se manifeste par des textes qui se rapportent à la tradition par la répétibilité d'une certaine finalité textuelle et surtout par certains caractères formels. La répétabilité de formes textuelles comprend une échelle continuelle de marquage de tradition minimaux - p.ex. une certaine dénomination de texte ou une formule dans un texte qui par ailleurs n'est pas fixé par la tradition -, l'organisation formelle et le figement définitif d'un texte.

Enfin le troisième type d'historicité concerne « les phénomènes individuels, non reproductibles, c'est-à-dire le texte en tant qu'individu, puisque chaque texte exprimé peut être situé historiquement. » (2005 : 153).

À la lumière de cette distinction le rapport entre quantitatif et qualitatif, l'utilisation de faits quantitatifs dans la description, doit être défini comme relevant d'une perspective historique : la répétition est une tradition qui s'interprète dans l'historicité plus large de la langue et qui permet d'interpréter les textes singuliers. Le rapport qui les unis est alors un rapport de contextualisation.

Comme nous l'avons souligné plus haut (1.2), toute prise en compte des faits de fréquence suppose une position relativement à la dichotomie langue/parole, et la notion de norme a un intérêt quantitatif du fait d'être un palier d'abstraction intermédiaire entre la parole et la langue. Loin d'être condamné à l'antinomie d'une imprédictibilité absolue des événements langagiers concrets et d'un stabilité des faits de système, comme l'exprimant Ullmann ci-dessus (3.4), l'un et l'autre sont liées par la traditionnalité de la norme.

Les particularités du cadre de la norme cosérienne pour l'analyse des faits de fréquence tient d'abord au fait que la distinction du système normal et du système fonctionnel n'implique pas l'établissement d'une solution de continuité entre le fonctionnel et le non-fonctionnel : les deux éléments s'inscrivent dans un continuum de niveau d'abstraction. Il s'agit ainsi de dépasser une opposition quantitatif/qualitatif, exprimée par exemple par Malmberg (1974 : 68) : « Il faut pourtant dans toute langue – et par conséquent dans la description linguistique – distinguer entre deux aspects différents qui pourtant se complètent. Il s'agit d'un côté de savoir ce qui se peut et de l'autre ce qui est fréquent ».

---

26 On trouve chez Meillet (1936 : 10) une attention au caractère fortement répétitif de l'activité de parler, qui ne va pas cependant jusqu'à en faire un lieu de configuration de la langue elle-même : « Un sujet de culture médiocre parle surtout par formules qui ne varient guère ; chez la plupart des gens les associations de mots ne sont ni libres ni personnelles. Le mot n'est qu'une partie de combinaisons pratiquement constantes ; la valeur du mot dans un pareil ensemble ne s'explique pas par le sens universel et général de ce mot, mais par l'habitude que l'on a de le voir dans certaines combinaisons. »

Dans cette perspective, les faits de fréquence sont des faits empiriques, observables, issus du discours concrets ; ils ne sont pas rapporté à une instance extralinguistique (comme une instance cognitive, cf. ci-dessous 4.4). Ils ne peuvent être formulés comme des lois universelles indépendamment d'une circonstance historique.

4.3 La notion de norme cosérienne peut être rapprochée d'autres propositions d'élaboration. On peut notamment la comparer avec la notion d'usage dans l'école contextualiste britannique. Dans cette tradition, l'objectif de la description sur corpus est de dégager des normes, des régularités, des usages typiques et routinisés : « Corpus linguistics offer new ways of studying linguistic routines: what is expected, predictable, usual, normal and typical in the utterance-by-utterance of spoken and written language in use. » (Stubbs 2001 : 241 ; cf. aussi 59, 100, 221, 222)

Comme dans la notion de norme cosérienne, l'usage, tel qu'il peut être observé à travers des fréquences, articule, ou plutôt permettrait de dépasser, l'opposition entre langue et parole : « Many features of an individual text (=*parole*) are idiosyncratic; if the were not, the text would convey no information and there would be no point in reading it. But a corpus is not mere performance: as a sample of language use, it reveals typical and repeated patterns. » (Stubbs : 2001) ; « *parole* is, therefore, becoming amenable to systematic observation » (Tognini-Bonelli 2001 : 86).

Les deux perspectives de la norme cosérienne et de l'usage contextualiste sont cependant rapidement inconciliables. Les divergences portent notamment sur le caractère résolument non structuraliste de la seconde : l'opposition entre langue et parole n'est pas dépassée par une tentative d'articuler les deux, mais de les remplacer par les seules régularités de l'usage.

Dès lors la fréquence intuitive ne peut être prise en compte et est simplement disqualifiée pour la description (« Speakers have strong intuitions about such characteristics of language use, but the basis for these feelings – about what is natural, native-like, authentic, typical and representative – is not well understood. », Stubbs 2001 : 59)

De ce fait encore, la description basée sur corpus produit des catégories descriptives autonomes, à l'exclusion de toutes autres catégories descriptives produites par une analyse structurale : « The corpus-driven approach […] aims to derive linguistic categories systematically from the recurrent patterns and the frequency distributions that emerge from language in context. (Tognini-Bonelli 2001 : 87)
Enfin la description n'ayant comme base que les régularités ne peut saisir ce qui est reconfiguration du système dans l'activité de parler.

4.4 Dans la perspective de Coseriu, c'est la notion de norme qui prend en charge la dimension de répétition. Elle prend en charge la dimension quantitative la plus primitive de la langue : celle de la répétition historique, entre l'acte singulier et la langue fonctionnelle, qui n'ont ni l'une ni l'autre de dimension de répétition.

Les grammaires usage-based proposent également d'intégrer une dimension « première », constitutive, de la répétition. Comme dans la perspective cosérienne, l'accent est mis sur l'activité de reconfiguration du système dans l'activité de parler ; Bybee/Hopper (2001 : 2) citent d'ailleurs Coseriu à l'appui de l'idée que « The fixing of linguistic groups of all kinds as recognizably structural units (word and phrase units) is an ongoing process; it is the result at any point in time of the "constant resystematization" of language. (Coseriu 1954) »

La dimension de répétition et la prise en compte de la fréquence sont dans ce cadre rapportés à l'activité cognitive : si la répétition importe, c'est à travers l'hypothèse que ses effets cognitifs rendent compte de la stabilisation et de la systématisation du système (Schmidt 2007 : 119) : « The frequency of occurrence of concepts or constructions in a speech community has an effect on the frequency with which its members are exposed to them. The (tacit rather than explicit) implication is that this results in some kind of collective automatization effect, which makes it possible to talk of the degree of entrenchment of a concept or construction in a given language. »

Ainsi la fréquence détermine les unités pertinente de l'analyse, en fonction d'hypothèses sur le « stockage cognitif » des unités : « High-frequency words and phrases grow strong with repetition and loom large, forming looser connections with other items, while low-frequency words and expressions are less prominent but grain stability by conforming to patterns used by other items. » (Bybee 2007 : 9)

Plusieurs différences avec le cadre cosérien méritent d'être soulignées. On observe une alternative entre deux modèles de prise en charge de la fréquence : dans le cas de Coseriu, la fréquence est rapportée à un plan d'analyse historique, tandis que dans le cas des grammaires usage-based, elle est rapportée à un plan cognitif.

L'opposition porte aussi sur le statut empirique de la fréquence : elle est directement observable dans le premier cas, mais non dans le second, puisque la fréquence qui compte est celle qu'« enregistre » l'activité cognitive, dont le fonctionnement demeure évidemment largement obscur. Les catégories descriptives, dans ce cadre, sont gagées sur les hypothèses relatives au fonctionnement cognitif : « The general claim is that this more frequent sequence gradually moved away from its source construction becoming more autonomous. » (Bybee 2006 : 11)

L'opposition porte enfin sur le régime interprétatif dont relèvent les deux conceptions de la fréquence. En tant que forme historique, les normes relèvent partiellement d'une interprétation et d'une herméneutique. Dans le second cas, la fréquence relève d'une explication de type causal.

Le recours au fonctionnement cognitif comme lieu d'organisation des catégories de la description fait retrouver un paradoxe de la fréquence dans le cadre d'une stricte dichotomie langue/parole : elle est à la fois empirique et non directement observable. Ici, d'après la distinction que nous avons établie ci-dessus (1.4), il s'agit plutôt d'une fréquence universelle, que d'une fréquence historique et située : les principes de l'entrenchment sont universels et indifférents aux langues et aux situations historiques, tandis que la norme n'a aucune réalité hors d'une configuration historique donnée. De plus, la « localisation » cognitive de la fréquence oblige à ne pas distinguer entre différents types de normes à l'origine des régularités observables, à faire de la fréquence un tout indécomposable.

Le cadre des grammaires usage-based ont peu développé l'usage des méthodes quantitatives, mis à part, notamment, les travaux de Gries (cf. eg. Wulff/Stefanowitsch/Gries 2007), qui restent pourtant marginaux dans ce cadre même. Les critiques que Bybee adresse à cet auteur sont d'ailleurs des critiques classiques à l'encontre de l'usage des méthodes quantitatives : les données fréquentielles ne sont pas articulées aux bonnes catégories descriptives ; la fréquence mesurée, en définitive, n'a aucun statut (Bybee 2010 : 98). Le cadre usage-based, même s'il inclut une attention à la variation, n'inscrit pas la fréquence dans la diversité des usages : en ce sens, il s'agit d'un retour à une notion de fréquence « en langue », secondairement paramétrée par des paramètres variationnels.

On peut relever que Coseriu semble fournir une objection anticipée à cette perspective (1982 : 93-94), où la notion historique de langue antérieure est opposée à la notion, psychologique et extra-linguistique, « d'acquis linguistique », qui reste néanmoins ici encore propre à une langue :

> Ahora bien, adoptando el punto de vista de un acto lingüístico concreto, podemos considerar una lengua que comprenda en una isoglosa ese mismo acto, pero también una « lengua anterior », sistema establecido, en la misma comunidad, sobbre la base de los actos lingüísticos precedentes al acto al que nos enfrentamos: el sistema en el que se encuentran los modelos de ese mismo acto, o con respecto al cual el acto se presenta como innovación. Ese concepto de « lengua anterior » es importante, porque corresponde, justamente, a una realidad histórica continuada por el nuevo acto considerado, al cuadro en el que se realiza como hablar una intuición individual e inédita; es un concepto lingüístico, por constituirse desde un punto de vista estrictamente lingüístico, pero, por su contenido, coincide prácticamente (por lo menos en gran parte), en el individuo o en el grupo de individuos considerados, con el concepto psicológico de « saber » o « acervo lingüístico ». Aquí también, se trata de modos distintos de encarar los mismos objetos, más bien que de objetos distintos: por un lado, se elabora una generalización sobre la base de fenómenos concretos; por el otro, se considera la misma generalización como saber depositado en la memoria de uno o más individuos. Pero, por eso mismo, el concepto de « acervo lingüístico » resulta exterior a la lingüística, que estructura sus abstracciones exclusivamente sobre la base de hechos concretamente registrados, y no sobre virtualidades o conjuntos de representationes no investigable con medios glotológicos.

Ces deux propositions d'articulation de la langue et de l'activité de parole concrète sensibles à la dimension de fréquence s'opposent donc comme une perspective universaliste et naturalisante à une perspective historique sur les faits de langue.

Conclusion

La prise en charge de la fréquence est paradoxale : les considérations de fréquence sont au cœur de nombreuses propositions descriptives, sans que cette notion ait été le plus souvent élaborée et posée comme une dimension centrale des faits de langue. Les objectifs et les méthodes présidant à l'utilisation de faits de fréquence dans la description font l'objet d'un consensus très faible. La diversité des approches souligne néanmoins le caractère incontournable et fondamental des faits de fréquence, que l'on peut aborder à un très grands nombres de degrés d'abstraction différents, depuis approches textuelles jusqu'à des degrés d'abstraction trans-linguistiques.

L'examen des usages de la fréquence a montré que celle-ci impliquait toujours une interprétation de la dichotomie saussurienne langue/parole, la langue pouvant presque être définie, négativement, comme « ce qui peut être décrit sans considération de fréquence. » L'utilisation de faits de fréquence pour la description suppose de les lier à des catégories descriptives et interprétatives. Le concept de norme tel qu'élaboré par Coseriu, qui insiste sur la traditionnalité de l'activité de parler concrète, intermédiaire entre la parole et la langue, permet particulièrement de contextualiser les faits de fréquence.

On peut lui opposer une théorie cognitive contemporaine qui insiste également sur les phénomènes de fréquence, mais du point de vue de leur effet cognitifs. La fréquence sert alors au passage depuis les phénomènes observables vers un arrière-plan cognitif. La répétition se traduit en faits cognitifs qui ne sont pas eux-mêmes directement observables. Alors que dans le cadre d'une théorie de la norme la fréquence est quelque chose d'essentiellement

empirique, historique et relève d'une interprétation, dans un cadre cognitif la fréquence n'est pas directement observable et elle relève de phénomènes cognitifs universel.

Références

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

Bachelard, G. (1993 [1938]) : *La Formation de l'esprit scientifique*. Paris: Vrin.

Baayen R. Harald (2001) : *Word Frequency Distributions*. Dordrecht: Kuwer.

— (2009) : « Corpus linguistics in morphology: morphological productivity », in : Lüdeling, A./ Kyto, M./ McEnery, T. (ed.) : *Corpus Linguistic, An International Handbook*. New York, Berlin: Walter de Gruyter, 899-919.

Bat-Zeev Shyldkrot, H (1995) : « *Tout* : polysémie, grammaticalisation et sens prototypique », in : *Langue française*, 107, 72-92.

Benzécri, J.-P. (1973) : *L'analyse des données. L'analyse des correspondances*. Paris: Bordas.

Biber, D. (1988) : *Variation across speech and writing*. Cambridge: Cambridge University Press.

— (1995) : *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press.

— (1990) : « Methodological issues regarding corpus-based analyses of linguistic variation », in : *Literary and Linguistic Computing*, 5, 4, 257-270.

— (1993) : « Using Register-Diversified Corpora for General Language Studies », in : *Computational Linguistics*, 19, 3, 219-241.

Blumenthal, P. (2006) : « De la logique des mots à l'analyse de la synonymie », in : *Langue française*, 150, 14-31.

Blumenthal, P. (2009) : « Éléments d'une théorie de la combinatoire des mots », in : *Cahiers de lexicologie*, 94, 11-29.

Bod, R./Hay, J./Jannedy, S. (2003): *Probabilistic Linguistics*. Cambridge: MIT Press.

Bybee, J./Hopper, P. (2001): *Frequency and the Emergence of Linguistic structure*. John Amsterdam: Benjamins.

Bybee, J. (2006) : « From usage to grammar: The mind's response to repetition », in : *Language*, 82, 4, 711-733.

— (2007) : *Frequency of use and organization of language*. Oxford: Oxford University Press.

— (2010) : *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Canguilhem, G. (2006 [1966]): *Le normal et le pathologique*. Paris: PUF.

Coseriu, E. (1965) : « Critique de la glottochronologie appliquée aux langues romanes », in : Straka, G. : *Actes du Xe congrès international de linguistique et philologie romanes (Strasbourg, 1962)*. Paris: Klincksieck, 87-96.

— (1982 [1952]) : « Sistema, norma y habla », in : *Teoría del lenguaje y lingüística general*. Madrid: Gredos, 11-113.

— (1986) : « Sistema, norma y tipo », in : *Lecciones de lingüística general*. Madrid: Gredos, 316-327.

— (1987) : « Le latin vulgaire et le type linguistique roman », in : Herman, J. (ed.) : *Latin vulgaire- latin tardif. Actes du Ier Colloque international sur le latin vulgaire et tardif (Pécs, 2 – 5 septembre 1985)*. Tübingen: Max Niemeyer Verlag, 53-64.

— (2001) : *L'Homme et son langage*, Dupuy-Engelhardt H./Durafour J.-P./Rastier F. (ed.), Peeters, Leuven. « Vers l'étude des structures lexicales » (1966/2001 : 215-252 [« Structure lexicale et enseignement du vocabulaire », *in Actes du Premier colloque International de Linguistique appliquée*, 175-217]). « Détermination et entour » (1955/1982 : 31-67 [« Determinación y entorno. Dos problemas de una lingüística del hablar », *Romanistisches Jahrbuch*, 7, 1955-1956, 29-54.]).
— (2007a [1958]) : « Synchronie, diachronie, histoire », in : *Texto !*, 12, 3/4, non paginé.
— (2007b [1980]) : « Du primat de l'histoire », in : *Texto !*, 12, 2, non paginé.
Cysouw, M. (2003) : « Quantitative Methods in Typology », in : Altmann, G./ Köhler, R./ Piotrowski, R. (ed.) : *Quantitative Linguistics: An International Handbook* (HSK). Berlin: Walter de Gruyter, 554-578.
Davis L. M. (1990) : *Statistics in Dialectology*. Alabama: University of Alabama Press.
Desrosières, A. (1989) : « Comment faire des choses qui tiennent : Histoire sociale et statistique », in : *Histoire & Mesure*, 4, 3, 225-242.
— (2000 [1993]) : *La politique des grands nombres. Histoire de la raison statistique*. Paris: La découverte.
— (2001) : « Entre réalisme métrologique et conventions d'équivalence : Les ambiguïtés de la sociologie quantitative », in : *Genèse*, 4, 112-127.
— (2008): *Pour une sociologie historique de la quantification. L'argument statistique*. Paris: Presses de l'École des Mines.
Fénelon, J.-P. (1981) : *Qu'est-ce que l'analyse des données ?* Paris: Lefonen.
François, F. (1968) : « La description linguistique », in : Martinet, A. (ed.) : *Le langage*, Encyclopédie de la Pléiade. Paris: Gallimard, 171-282.
Gadet, F. (1971) : « Recherches récentes sur les variations sociales de la langue », in : *Langue française*, 9, 74-81.
— (2000) : « On n'en a pas fini avec les problèmes de recueil de corpus », *in* : Andersen, H. L & A. B. Hansen (ed.) : *Le français parlé. Actes du colloque international, Université de Copenhague, 29 au 30 octobre 1998.* Copenhague: Museum Tusculanum Press, 29-44.
— (2003) : « La signification sociale de la variation », in : *Romanistisches Jahrbuch*, 54, 98-114.
Germain, C. (1981) : *La sémantique fonctionnelle*. Paris: PUF.
Goebl, H. (2003) : « Dialektometrie / Dialectometry », in : Altmann, G./ Köhler, R./ Piotrowski, R. (ed.) : *Quantitative Linguistics: An International Handbook* (HSK). Berlin: Walter de Gruyter, 498-532.
Guiraud, P. (1969) : *Essais de stylistique*. Paris: Klincksieck.
— (1959) : *Problèmes et méthodes de la statistique linguistique*. Dordrecht: R. Reidel.
Gumperz (1982) : *Discourse strategies.* Cambridge: Cambridge University Press.
Glessgen, M.-D. (2007) : *Domaines et méthodes en linguistique française et romane*. Paris: Armand Colin.
Gries S. Th. (2009) : *Quantitative Corpus Linguistics with R*. New York: Routledge.
Heger, K. (1969) : « La sémantique et la dichotomie de langue et parole », in : *Travaux de linguistique et de littérature*, 7, 1, 47-111.
Heilmann, L. (1972 [1962]) : « Statistical Considerations and Semantic Content », in : B. Malmberg (ed.) : *Readings in Modern Linguistics*. Stockholm: Läromedelsförlagen/ Mouton.
— (1983) : « Considerazioni statistiche e contenudo semantico », in : *Linguistica et umanismo*, Bologna: Mulino, 217-230.

Herdan, G. (1966 [1956]) : *The Advanced Theroy of Language as Choice and Chance*. Berlin/Heidelberg/New York: Springer-Verlag.

— (1960): *Type-token Mathematics*. The Hague: Mouton.

Hjelmslev, L. (1971) : *Essais linguistiques*, Minuit, Paris. « La stratification du langage » (1954/1971 : 44-76 [*Word*, 10, 163-188]). « Langue et parole » (1943/1971 : 77-89 [*Cahiers F. de Saussure*, 2, 29-44]).

Jakobson, R. (1963 [1961]) : « Linguistique et théorie de la communication », in : *Essais De Linguistique Générale.* Minuit, Paris, 87-99.

Johnson, K. (2008) : *Quantitative Methods in Linguistics*. Malden: Blackwell.

Kabatek, J. (2005) : « À propos de l'historicité des textes », in : Murguía, A. : *Sens et référence, Mélanges pour Georges Kleiber*. Tübingen: Gunter Narr Verlag, 149-158.

Kilgarriff, A. (2005) : « Language is never ever ever random », in : *Corpus Linguistics and Linguistic Theory*, 1, 2, 263-276.

Koch, P./ Oesterreicher, W. (2001) : « Gesprochene Sprache und geschriebene Sprache/Langage parlé et langage écrit », in : Holtus, G./ Metzeltin, M./ Schmitt, C. (ed.) : *Lexikon der Romanistischen Linguistik*, vol. 2. Tübingen: Max Niemeyer Verlag, 584-627.

Labov, W. (1976) : *Sociolinguistique*. Minuit, Paris (traduction A. Kihm).

— (2004) : « Quantitative Analysis of Linguistic Variation », in : Ammon, U./ Dittmar, N./ Mattheier, K. J/ Trudgill, P. (ed.) : *Sociolinguistics/Soziolinguistik* (HSK). Berlin: Walter de Gruyter, 6-21.

Lafon, P. (1980) : « Sur la variabilité de la fréquence des formes dans un corpus », in : *Mots,* 1, 127-165.

— (1981) : « Analyse lexicométrique et recherche des cooccurrences », in : *Mots*, 3, 95-148.

Laks, B. (2008) : « Pour une phonologie de corpus », in : *French language studies*, 18, 3-32.

Lara, L. F. (1983) : « Le concept de norme dans la théorie Eugenio Coseriu », in : Bédard, É./ Maurais, J. (ed.) : *La norme linguistique*. Québec: Gouvernement du Québec, non paginé.

Larochette, J. (1981) : « ‹ Normal › et ‹ anormal › dans la syntaxe », in : Geckeler H./ Schlieben-Lange, B./ Trabant, J./ Weydt, H. : *Logos semantikos*. Berlin, New York, Madrid: Walter de Gruyter / Gredos, vol. 2, 131-140.

Lebart, L./Salem, A./Berry, L. (1998) : *Exploring textual data*. Dordrecht: Kuwer.

Lemercier, C./Zalc, C. (2008) : *Méthodes quantitatives pour l'historien*. Paris: La découverte.

Lodge, A. (1994) : « Parlers populaires et normalisation politique et sociale : poissard, parigot, cockney », in : *Romantisme*, 24, 86, 25-32.

Loiseau, S. (2006) : *Sémantique du discours philosophique chez Deleuze : du corpus aux normes*. Thèse de doctorat, Université Paris X Nanterre.

— (2008) : « Corpus, quantification et typologies textuelles », in : *Syntaxe et sémantique*, 9, 73-85.

— (2010) : « Investigating the interactions between different axes of variation in text typology », in : Grzybek, P./ Kelih, E. (ed.) : *Text and Language: Structures, Functions, Interrelations*. Wien: Praesens Verlags, 109-118.

— (à paraître) : « Contextualité et tactique sémantique dans un texte philosophique », in : Valette, M. (ed.) : *Concepts en contexte*. Nancy: Presses universitaires de Nancy.

— (en préparation) : « Les faits statistiques comme objectivation ou comme interprétation : statistiques et modèles usage-based », in : *Travaux de linguistique*.

Lüdtke, H. (1996) : « Changement Linguistique », in : Goebl, H./ Nelde, P. H./ Starý, Z./ Wölck, W. (ed.) : *Kontaktlinguistik* (HSK). Berlin: Walter de Gruyter, 526-540.

Macaulay, R. (2009) : *Quantitative Methods in Sociolinguistics*. Basingstoke: Macmillan.
Malmberg, B. (1974) : *Manuel de phonétique générale*. Paris: Picard.
Mańczak, W. (1969) : « Quelques réflexions sur la doctrine de Noam Chomsky », in : *Linguistics*, 49, 18-27.
Manning, C. (2003) : *Probabilistic Syntax*, *in* Bod *et al.* (2003), 289-341.
Marcellesi, J.-B. (1971) : « Linguistique et groupes sociaux », in : *Langue française, Linguistique et groupes sociaux*, 9, 119-122.
Martinet, A. (1970) : *Éléments de linguistique générale*. Paris: Armand Colin.
— (1974 [1965]) : *La linguistique synchronique*. Paris: PUF.
— (1985) : *Syntaxe générale*. Paris: Armand Colin.
— (1955) : *Économie des changements phonétiques*. Berne: Francke.
Meillet, A. (1982 [1905]) : « Comment les mots changent de sens », in : *Linguistique historique et linguistique générale*. Genève: Slatkine, 237-271.
— (1936) : *Linguistique historique et linguistique générale*, vol. II, Paris: Klincksieck.
Mondada, L. (2000) : « Les effets théoriques des pratiques de transcription », in : *Linx*, 42, 131-150.
Muller, C. (1992 [1973]) : *Initiation aux méthodes de la statistique linguistique*. Paris: Champion.
Newman, M. E. (2005) : « Power laws, Pareto distributions and Zipf's law », in : *Contemporary Physics*, 46, 5, 323-351.
Obin, N./ Lacherey, A./ Veaux, C./ Rodet, X./ Simon, A.-C. (2008) : « A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features », in : *Proceedings Interspeech 2008*, 1204-1207.
Petruszewycz, M. (1973) : « L'histoire de la loi d'Estoup-Zipf : document », in : *Mathématiques et sciences humaines*, 44, 41-56.
Rastier, F. (1981) : « Le développement du concept d'isotopie », in : *Actes Sémiotiques*, 3, 29, 5-29.
— (2001) : *Arts et sciences du texte*. Paris: PUF.
— (2001 b) : « Vers une linguistique des styles », in : *L'information grammaticale*, 89, 3-6.
Saussure, F. de (1995 [1916]): *Cours de linguistique générale*. Paris: Payot.
Schegloff, E. (1993) : « Reflections on Quantification in the Study of Conversation », in : *Research on Language and Social Interaction*, 26, 2, 99-128.
Schmid, H.-J. (2007) : « Entrenchment, salience, and basic levels », in : Geeraerts, D./ Cuyckens, H. (ed.) : *The Oxford handbook of cognitive linguistics*. Oxford: Oxford University Press, 117-138.
Stefanowitsch, A./ Gries, St. Th. (2003) : « Collostructions : Investigating the interaction of words and constructions », in : *International Journal of Corpus Linguistics*, 8, 2, 209-243.
Stubbs, M. (2001): *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
Schlieben-Lange, B. (1998) : « Les hypercorrectismes de la scripturalité », in : *Cahiers de Linguistique Française*, 20, 255-273.
Tognini-Bonelli, E. (2001) : *Corpus Linguistics at Work*. Amsterdam/Philadelphia: Benjamins.
Tournier, M. (1980) : « D'ou viennent les fréquences de vocabulaire ? », in : *Mots*, 1, 189-212.
Troubetzkoy, N. S. (1986 [1939]): *Principes de phonologie*. Paris: Klincksieck.

Troubetzkoy, N. S. (1969) : « La phonologie actuelle », in : Pariente, J.-C. (ed.) : *Essais sur le langage*. Paris: Minuit, 141-164.

Ullmann, S. (1951) : *The Principles of Semantics*. Jackson, Glasgow : Basic Blackwell.

Völker, H. (2009) : « La Linguistique variationnelle et la perspective intralinguistique », in : *Revue de linguistique romane*, 73, 27-76.

Wells, R. (1957) : « A mathematical Approach to Meaning », in : *Cahiers Ferdinand de Saussure*, 15, 117-136.

Wulff, S./ Stefanowitsch, A./ Gries, St. Th. (2007) : « Brutal Brits and persuasive Americans: variety-specific meaning construction in the into-causative », in : Radden, G./ Köpcke, K.-M./ Berg, T./ Siemund, P. : *Aspects of meaning construction*. Amsterdam: John Benjamins, 265-281.

Zipf, G. K. (1935) : *The Psychobiology of Language. An Introduction to Dynamic Philology*. Boston: Houghton-Mifflin.