# Théories de la fréquence linguistique et interprétations des faits quantitatifs en sémantique

Sylvain Loiseau

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100332




# Théories de la fréquence linguistique et interprétations des faits quantitatifs en sémantique

Loiseau, Sylvain

LDI (UMR 7187 CNRS / Université Paris 13-Nord)
sylvain.loiseau@univ-paris13.fr

## 1 Introduction : utilisation des faits de fréquence en sémantique

La sémantique a longtemps tenu pour négligeable les faits de fréquence. Rappelons la position à ce sujet de Ullmann (1951 : 294) :

> « 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 […]. »

La notion de *fréquence* a reçu depuis quelques décennies une importance accrue dans la théorie linguistique et les pratiques descriptives. Sur le plan des pratiques descriptives, elle a reçu le renfort de l'engouement pour les grands corpus. Sur le plan théorique – et avant que ne commence cet engouement – elle est revenue sur le devant de la scène théorique avec, notamment, la diffusion de la linguistique cognitive, notamment sous l'espèce des « modèles basés sur l'usage » (cf. Legallois & François, 2011).

En effet, ces modèles basés sur l'usage » ont mis au premier plan la fréquence comme facteur expliquant l'émergence et la systématisation des faits linguistiques (notamment des unités lexicales et grammaticales) à partir de la répétition dans l'usage. L'importance de la répétition tient à des hypothèses sur les effets cognitifs de la répétition : la fréquence serait source de systématisation à travers l'effet de la répétition sur le stockage et le traitement cognitif des unités. C'est ce que résume le concept d'*enracinement cognitif* (*entrenchment*) introduit par Langacker (1987: 59) : « With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit; moreover, units are variably entrenched depending on the frequency of their occurrence. ». La fréquence détermine les unités de la description. Hoffmann peut donc indiquer (2004 : 172) : « In a theory which considers language use and language structure to be interrelated, frequency of occurrence is likely to play an important role ».

Les remarques sur le fait que la fréquence est, en même temps, mal connue, et ses effets mal définis, sont cependant nombreuses. Ainsi, Mair (2004 : 126) indique : « [...] we are still far from a precise understanding of the role of frequency in grammaticalisation ». Hoffmann (2004 : 185) fait un constat similaire : « [i]t is a basic tenet of functional linguistics that language use shapes grammar. However, the exact mechanisms which underlie the interaction between language use and language structure are a matter of continuous debate. One of the variables in the equation is certainly "frequency of occurrence" ».

À côté de cette insistance sur l'importance théorique de la notion de fréquence, les descriptions qui s'inscrivent dans le cadre de l'usage sont de plus en plus nombreuses à recourir à des corpus, aux méthodologies de la linguistique de corpus, et à des méthodes statistiques. Schmid (2007 : 127) parle d'un « quantitative turn in the investigation of entrenchment and salience effects » qu'il fait remonter aux travaux de Geeraerts, Grondelaers, & Bakema (1994) (cf. également Schmid, 2010 : 102 pour des jalons sur les premiers travaux dans cette perspective). Ces travaux sont parfois désignés sous le terme de « cognitive corpus linguistics » ou « quantitative corpus-driven cognitive semantics » (cf. Glynn & Fischer, 2010). Ils reposent sur l'objectif d'« opérationnaliser » la fréquence définie dans le cadre cognitif, c'est-à-dire sur l'hypothèse selon laquelle « frequency in text more or less directly instantiates salience or



Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20120100332

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100332


entrenchment in the cognitive system » (Schmid, 2010 : 102). La fréquence est ici une grandeur mesurée, empirique. Or, sur ce plan méthodologique également, les difficultés rencontrées dans le traitement de la fréquence mesurée dans des corpus sont fortes. Schmid indique ainsi (2010 : 125) : « so far we have understood neither the nature of frequency itself nor its relation to entrenchment, let alone come up with a convincing way of capturing either one of them or the relation between them in quantitative terms ». Les règles d'interprétation des résultats et de mise en œuvre de ces méthodes sont donc encore largement indéterminées. Une meilleure compréhension de la fréquence et de son utilisation dans la description est sans doute aujourd'hui une condition de progrès dans l'utilisation de corpus pour la description.

Nous proposons dans cet article de faire le point sur la notion de fréquence. D'abord, en rappelant son importance théorique à différents niveaux de description et en proposant une typologie de différents types de fréquence. Celle-ci nous permettra de situer plus précisément les modèles basés sur l'usage parmi les tentatives de rendre compte de la fréquence linguistique.

Ensuite nous ferons le bilan des difficultés dans l'« opérationnalisation » de cette notion. La fréquence est une notion paradoxale. C'est (*a priori*) une propriété empirique, immédiatement observable, supposée plus observable et moins construite en tout cas que de nombreuses catégories et de nombreux faits de systèmes traditionnels. C'est à ce titre, parce qu'elle serait une garantie d'empiricisme, qu'elle se trouve promue dans les modèles basés sur l'usage au rôle de notion théorique centrale, à travers l'hypothèse que les formes linguistiques émergent de l'usage. Et pourtant, d'un point de vue pratique, cette notion est difficile à opérationaliser. Dans la perspective de « quantitative corpus-driven cognitive semantics » de nombreux travaux ont été amenés à souligner que les faits de fréquence ne peuvent être interprétés qu'à l'aide d'un cadre textuel.

Nous illustrerons cela avec une expérience sur l'analyse sémantique des fréquences de cooccurrents lexicaux. On montre que, dans un genre donné, les cooccurrents d'une unité lexicale peuvent varier en fonction de la position de cette unité en début ou en fin de texte. Cette variation des cooccurrents peut s'interpréter comme des emplois sinon des acceptions, différents de l'unité considérée. Cette expérience est une contribution à la démonstration de la nécessité de prendre en compte les normes textuelles dans l'utilisation des faits de fréquence pour la modélisation sémantique.

## 2 Théorie de la fréquence

Devant l'importance théorique prise par la notion de fréquence il semble nécessaire d'entreprendre un travail doxographique et un travail d'élaboration théorique de cette notion. De fait, plusieurs types de fréquence peuvent être distinguée en linguistique.

### 2.1 Différents types de fréquence

La fréquence est présente dans la définition de nombreuses notions. Du fait du faible intérêt traditionnellement accordé aux questions de fréquence[1], l'omniprésence de cette dernière est sans doute sous-estimée. Un inventaire rapide de quelques concepts cardinaux, au risque de produire un catalogue un peu superficiel, permet cependant d'illustrer cette omniprésence. Elle est liée par exemple, en phonologie, à la théorie de la marque (Troubetkoy, 1986 : 282 ; Greenberg, 1966 : 64-71 et *passim*). La distinction des différentes parties du discours mobilise traditionnellement une opposition entre des faits fréquents et des faits moins fréquents (Martinet, 1985 : 119). Dans le domaine lexical, les critères de fréquence sont souvent mentionnés pour définir les unités polylexicales, c'est-à-dire pour se donner les éléments de base de la description. 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. Ce sont les définitions d'unités ou de catégories fondamentales de la description qui sont en jeu[2]. 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, comme 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[3].

De tels exemples de mobilisation du critère de la fréquence dans la définition de phénomène peuvent être multipliés. Elle est particulièrement présente, par exemple, dans les perspectives variationnelles/variationnistes. Glessgen (2007 : 102) prétend que « [la] distinction concrète des différentes variétés [d'une langue] est plus le résultat d'une étude statistique qu'une réalité linguistique reconnue par les locuteurs ».

La fréquence enfin est présente dans l'exposé de la dichotomie langue / parole, quand la langue est définie comme une moyenne (Saussure, 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. ». « Moyenne » (puis « somme »[4]) ont sans doute un statut métaphorique (marqué par l'enclosure « une sorte de »). Cette métaphore est cependant récurrente dans le *CLG*[5]. Elle occupe une place stratégique, celle de l'articulation entre parole (« reproduiront ») et langue, du passage de l'une à l'autre[6] ; 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) ou un « hole in the middle » (Geeraerts, 2010 : 74)[7]. De fait, comme nous le verrons ci-dessous, toutes les utilisations de méthodes quantitatives s'accompagnent d'une prise de position par rapport à cette dichotomie.

Ce sont donc un grand nombre de notions cardinales en linguistique qui mobilisent des considérations de fréquence. Si elles renouvellent, certes, la question, les perspectives « basées sur l'usage », en donnant une importance théorique à la fréquence, s'inscrivent cependant aussi dans une tradition bien établie.

## 2.2 La fréquence et l'abstraction linguistique

La relation de la description linguistique aux faits de fréquence peut être également posée du point de vue du statut des abstractions produites par la description linguistique. Dans quelle mesure les résultats de la description, les systématicités, sont-elles d'abord des régularités, c'est-à-dire dans quelle mesure la description linguistique est-elle quantitative ? Mańczak (1969 : 19) estime que « 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. ». En d'autres termes, la linguistique comme procédure d'abstraction (Coseriu, 1982 : 61) est fondée sur des répétitions, sur la prise en compte de régularités. Dans la discussion du statut des « lois » synchroniques dans le *CLG*, par opposition aux lois diachroniques, Saussure (1995 : 131) fonde les lois synchroniques (par oppositions par exemple aux lois physiques) sur une dimension de régularité, donc sur une considération de fréquence : « En résumé, si l'on parle de loi en synchronie, c'est au sens d'arrangement, de régularité ». Les lois synchroniques n'ont donc pas d'autre statut que celui de régularités statistiques. Les modèles basés sur l'usage soulignent particulièrement cette dimension. Du Bois (1985 : 363) indique que « grammars code best what speakers do most ». Langacker (1987 : 62) que « [...] the centrality of a given unit and the importance of including it in the description consequently vary depending on the proportion of speakers familiar with it. ».

La question de l'articulation de la répétition et du systématique, de leur relation dialectique, que les modèles basés sur l'usage contribuent particulièrement à élaborer, n'est donc pas neuve. Heilmann (1983 : 218) la formule par exemple ainsi : « In altre parole, nel linguaggio gli aspetti qualitativo e quantitativo sono indipendenti, ovvero tra numero e funzione esistono rapporti determinati? ». Cohen (1949 : 10) exprime ainsi l'enjeu de la prise en compte des faits de fréquence :

> Mais il convient de se rendre compte [...] que la manière dont les questions ont été posées depuis quelques décades implique pour l'avenir la nécessité de statistiques. L'école de Prague, développant les idées « phonologiques » qui pointaient déjà auparavant, a nettement défini le rôle des oppositions ; mais très vite elle a aperçu que ces oppositions elles-mêmes avaient des valeurs très différentes suivant leurs rendements : ceux-ci ne peuvent être appréciés sans précisions.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100332


Toute interrogation sur la fréquence porte donc sur la nature de l'articulation entre les deux termes de la dichotomie langue / parole, sur la nature de l'abstraction linguistique et sur la solution de l'apparent paradoxe entre « la libertà delle uso singolo et il determinismo dell'insieme » (Heilmann, 1983 : 218). La version la plus pessimiste, où aucune articulation n'est possible, est représentée par exemple par Ullman cité au début de cet article. À l'autre extrême, un choix tout aussi radical 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 celui de la généralisation statistique (Herdan, 1966 : 28 ; Muller, 1992 : 14).

Au terme de ce bref coup d'œil, on peut dégager trois façons distinctes d'établir un lien entre les faits linguistiques et les questions de fréquence. En premier lieu, ce lien peut être établi au niveau de l'objet, dans l'affirmation que la dimension de répétition est constitutive des objets linguistiques. En second lieu, il peut être établi au niveau de la discipline, quand on affirme que les descriptions linguistiques ont une dimension d'approximation statistique – indépendamment de l'usage de méthode quantitative. Enfin, ce lien peut être établi au niveau méthodologique et tenir à l'utilisation de méthodes proprement quantitatives. Ce sont trois niveau d'implication croissante du quantitatif : chacun suppose les précédents.

## 2.3 Fréquence textuelle, fréquence sociale et fréquence typologique

Pour commencer, plusieurs types de fréquence mobilisés dans la description linguistique peuvent être distingués. Une première distinction concerne la nature de la fréquence : à la fréquence comme répétition, ou « fréquence textuelle », sur laquelle je reviens plus bas, on peut opposer une autre fréquence, également importante pour la description et qui porte sur des effectifs de locuteurs. C'est cette fréquence, que l'on peut appeler « fréquence sociale », que l'on manipule quand on relève des corrélations entre des traits phonétiques et des propriétés socio-démographiques des locuteurs. Cette fréquence intervient dans un grand nombre de perspectives, dans des problématiques variationnelles, de contact, de changement, etc. : il est remarquable qu'il s'agit, tout autant que la fréquence textuelle, d'une dimension quantifiable et fondamentale des faits de langue[8].

Enfin des quantifications peuvent porter, non plus sur l'observation du fonctionnement de la langue, mais sur des types linguistiques observés à partir d'une description. Par exemple, quand Cohen (1949 : 12) s'intéresse à la proportion de racines bilitères, trilitères et quadrilitères dans une langue sémitique, quand il se demande le nombre de « bons rapprochements étymologiques » sur lequel est construite la grammaire comparée indo-européenne (1949 : 14), quand on s'intéresse à la proportion du lexique héréditaire ou des emprunts dans le lexique d'une langue, il s'agit de grandeurs qui s'obtiennent moins à partir d'un texte qu'à partir d'une description linguistique déjà donnée (dictionnaire, grammaire). Ce type de grandeur intéresse particulièrement la typologie linguistique. La distinction entre fréquence textuelle et ce qu'on pourrait ainsi appeler des « fréquences typologiques » s'inscrit cependant peut-être dans un continnum dans certains cas.

La distinction entre « type frequency » et « token frequency » utilisée particulièrement par Bybee gagne me semble-t-il à être comparée à cette distinction : dans la plupart des cas, seul le « token frequency » est une fréquence textuelle, une répétition, tandis que le « type frequency » est davantage une « fréquence typologique ». Les deux éléments pourraient donc être éloignés d'un point de vue d'une typologie plus générale des types de fréquences.

La distinction de ces trois types de fréquence paraît sans doute aller de soi mais le fait que les mêmes méthodes statistiques puissent être utilisées pour chacune d'elle (notamment des méthodes factorielles) peut parfois dissimuler cette différence fondamentale de statut. Les règles d'interprétation, les fondements théoriques, la réalité prise en compte dans chacun de ces trois cas sont évidement entièrement différents.

## 2.4 Fréquence émique et fréquence étique

Si l'on se restreint maintenant à la seule fréquence comme répétition (également appelée « fréquence textuelle ») on peut encore proposer plusieurs distinctions.

*Fréquence émique et fréquence étique* – Une distinction importante est celle qui oppose la fréquence issue d'une intuition (d'un jugement intuitif) et la fréquence issue d'une mesure, d'une méthode quantitative. On peut appeler ces fréquences *fréquence intuitive* et *fréquence mesurée*. On pourrait également proposer le couple *fréquence émique* et *fréquence étique*, pour éclairer cette distinction par le couple notionnel étique/émique.

En effet la différence ne tient pas seulement au moyen d'obtenir ces fréquences, mais plus radicalement à leur place « à l'intérieur » ou « à l'extérieur » de l'objet de la description : les fréquences émiques font partie de l'objet à décrire, et le linguiste y accède en tant que locuteur[9].

De plus, tout locuteur a des intuitions sur la fréquence de tel ou tel phénomène – tout locuteur du français pense sans doute que les occurrences du verbe *être* sont plus nombreuses que les occurrences du verbe *danser* sans avoir eu besoin de compter effectivement, dans des textes, les occurrences de ces deux lexèmes. Il y a donc tout autant des intuitions (des connaissances épilinguistiques) sur la fréquence des phénomènes que des intuitions sur des propriétés systémiques.

La question essentielle est naturellement celle du rapport entre ces deux fréquences ; peut-on dire, en particulier, que la fréquence émique (intuitive) est seulement une version imprécise de la fréquence étique (mesurée) ? Portent-elles, pour commencer, sur les mêmes objets ? Il me semble important au contraire de distinguer fortement les deux types de fréquence. Schegloff (1993 : 118-119) 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. »

Troubetzkoy (1986 : 8-9) étend la différence entre les deux fréquences à leurs objets mêmes.

> Il va de soi 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.

Les objets de la fréquence intuitive et de la fréquence mesurée ne sont donc pas les mêmes. Comme le montre Troubetzkoy, la fréquence mesurée nécessite ce qu'en termes statistiques on appelle une procédure de *codage*, c'est-à-dire une convention pour établir les phénomènes décomptés, qui est nécessairement une approximation ou une interprétation des faits par le linguiste. À l'inverse, la fréquence intuitive se fonde sur une appréhension des faits qui ne nécessite pas un tel codage. D'autre part, comme l'indique Schegloff, la fréquence mesurée porte sur un corpus, nécessairement fini, assemblé pour pouvoir être interprétable (et représentant donc une variété, un genre, etc. ou un ensemble raisonné de tels unités), tandis que la fréquence intuitive ne porte pas sur un ensemble documenté. Ainsi, alors que l'on s'accorde aujourd'hui à dire qu'il n'y a pas de fréquence (quantifiée) d'un lexème « en langue »[10], une fréquence émique quant à elle, la « norme des sujets parlants » de Troubetzkoy, peut porter sur la langue comme sur n'importe quelle abstraction dont un locuteur peut avoir l'intuition. Ce sont donc les objets qui diffèrent tout comme l'interprétation à donner à cette grandeur.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100332


Enfin, la fréquence émique peut diverger de façon systématique de la fréquence étique, particulièrement quand des représentations linguistiques sont en jeu. C'est ce que montre par exemple le fait bien connu de la sous-évaluation systématique par les locuteurs de la fréquence des traits stigmatisés dans leur propre production. Les jugements intuitifs sont donc sensibles aux représentations.

Le fait que la fréquence est l'objet d'un savoir épilinguistique des locuteurs (et donc également des linguistes, en tant qu'ils sont locuteurs) est sans doute l'un des principaux arguments en faveur de son importance empirique et de sa prise en compte dans la description. Ces intuitions importent au fonctionnement de la langue même, comme à son évolution : on peut supposer qu'entre deux formes de mêmes fréquences mesurées mais dont l'une est massivement sous-estimée dans les représentations partagées, tandis que l'autre est perçue comme fréquente, la seconde aurait un attrait analogique beaucoup plus fort – ce que dit, par exemple, Mair (2004 : 139) : « After all, grammaticalisation did not proceed because a form was becoming more and more frequent, but because successive generations of speakers perceived the phenomenon and developed it further [...] ».

Ainsi, dans certains cas (et particulièrement dans l'étude des phénomènes de grammaticalisation), on peut dire pour conclure que *ce sont les mesures quantifiées (étiques) qui sont des approximations de mesures intuitives (émiques)*, et non l'inverse. Les objets pertinents sont en effet les fréquences intuitives, les fréquences quantifiées n'ont de statut et d'intérêt que comme approximation des fréquences perçues par les locuteurs. D'autre part, quand on quantifie des faits linguistiques, on ne procède pas à une modélisation ou une mathématisation quantitative (qui vaudrait un supplément de scientificité à la discipline), mais on interprète, par des approximations quantitatives, une dimension déjà présente dans le fonctionnement de l'objet.

De ce point de vue, dans une perspective basée sur l'usage, les données quantitatives issues de procédures de mesures et de décompte de fréquences textuelles ne peuvent être considérées comme des données premières.

## 2.5 Fréquence et niveaux d'abstraction

Une autre distinction importante porte sur le degré de « généralité » des descriptions basées sur la prise en compte de faits de fréquence.

D'un côté, on peut parler de propriétés de fréquence sur un plan très abstrait, trans-linguistique. Les lois quantitatives universelles sont particulièrement représentées 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 » (moindre effort articulatoire par exemple). Ces rapports sont censés être des constantes mathématiques, des propriétés internes et universelles des langues, fondées sur le principe, plus général encore, de la tendance au moindre effort.

À un autre extrême, des observations impliquant des faits de fréquences peuvent porter sur des objets historiques singuliers : texte, genre, style, etc. Dans ce cas, la généralisation, si besoin est, passe souvent par une méthode contrastive. Les premiers usages de la quantification par la philologie relevaient de ce second pôle (Cohen, 1949 : 8) : par exemple, pour caractériser le lexique d'un texte ou d'un auteur. Cet usage de la fréquence caractérise également les descriptions de « dimensions de variation » proposées par Biber (1995), bien qu'il s'agisse également, dans ce cas, d'identifier également des co-fréquences de phénomènes qui caractérisent des types de discours au-delà du corpus étudié.

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

Ces distinctions peuvent nous permettre de mieux caractériser la fréquence des modèles basés sur l'usage. Elle est à la fois émique et universaliste, ce qui tient simplement à son caractère cognitif. D'une part en effet son objet n'est pas les propriétés matérielles des textes mais le fonctionnement des locuteurs. D'autre

part les principes dans lesquels les faits de fréquences sont engagés sont des principes universels, puisqu'ils concernent le fonctionnement cognitif. Les principes de l'*entrenchment* sont universels et indifférents aux langues et aux situations historiques. 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. La fréquence, en ce sens, n'est pas un fait historique dans cette perspective[11].

Le recours au fonctionnement cognitif comme lieu d'organisation des catégories de la description fait retrouver le paradoxe récurrent dans les considérations de fréquence : cette dernière est à la fois posée comme une dimension empirique primaire de l'objet (et, de fait, elle est au moins approximativement observable à travers des méthodes quantitatives) et en même temps elle est mobilisée principalement sur un plan théorique, voire même pour fonder un plan qui échappe à l'observation (celui du fonctionnement cognitif).

## 3 Le développement d'un point de vue textuel

En conclusion, il apparaît que la fréquence, au sens des modèles « basés sur l'usage » présente de nombreux points communs avec plusieurs autres élaborations où la fréquence reçoit un statut universel. À travers les notions d'*entrenchment*, d'*automatisation* ou de *routinization*[12], la fréquence est construite comme une cause, de type mécanique, et interprétable univoquement.

Nous avons montré également que les lois ainsi dégagées ne peuvent que rester des approximations des faits réellement observés dans des situations historiques concrètes. Cette fréquence n'a pas encore pu être « opérationnalisée » : le rôle mécanique[13] et général postulé pour la fréquence est difficile à rendre opératoire dans le cadre de descriptions utilisant corpus et méthodologies quantitatives. Mair (2004 : 126) constate par exemple l'absence de résultat stable et généralisable pour caractériser empiriquement et quantitativement le rapport entre fréquence et grammaticalisation : « In the published literature, for example, there is little to guide us on the most crucial questions: Is an increase in discourse frequency a prerequisite for and concomitant of ongoing grammaticalisation […] ? » Hoffmann de même, (2004 : 188) remarquant que la fréquence n'est pas un critère opératoire à elle seule (« frequency alone cannot be decisive factor in the selection of lexical items as sources for grammaticalization. ») remet en cause le modèle d'une causalité mécanique (2004 : 205) : « Rather than seeing entrenchment (or lexical strength) in a purely mechanical light, such an approach focuses on the concept of saliency as context-dependent variable. ».

Dès lors, on peut se demander si ce niveau d'abstraction a une pertinence descriptive. C'est la critique que Troubetzkoy (1986 : 282, cf. aussi Martinet, 1955 : 132) adressait déjà aux lois de Zipf : « [d]ans 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" […]. 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. »

De fait, on peut souligner que l'argument de la fréquence, quand elle reste une notion purement théorique, est mobilisé pour expliquer des processus largement contradictoires. Hoffmann (2004 : 188-189) souligne ces contradictions : « Hopper and Traugott (1993: 103), for example, state that "[t]he more frequently a form occurs in texts, the more grammatical it is assumed to be. Frequency demonstrates a kind of generalization in use patterns." However, several authors have also noted that extremely high frequency can have the opposite effect: [...] 'conserving effect' […]. A typical example of this conserving effect is seen in the cas of highly frequent irregular verbs [...] which retain their conservative irregular past tense forms rather than being replaced by the highly productive regular *-ed* pattern. » Un très grand nombre de lois universelles fréquentistes ont été proposées. On a pu défendre que ce qui est fréquent est court (Zipf, 1935), ou ancien (Käding, 1897 ; Zipf, 1949) ou irrégulier (Greenberg, 1966: 68; cf. the “conservative effect” Bybee & Thompson, 2000: 380) ou grammatical (Hopper & Traugott, 1993: 103 ; Haiman, 1994)

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100332


ou, comme nous l'avons vu, plus facilement accessible cognitivement (« entrenched ») (Langacker, 1987 59-60 ; Bybee, 2003, 2007 ; Croft 2008). Les phénomènes de fréquence ont aussi été réputés moins marqués (Zipf, 1935 ; Haspelmath, 2008). Les types fréquents sont plus productifs que ceux qui sont rares (Baayen, 2009). La diffusion d'une innovation est supposée avoir une signature quantitative (la basse fréquence initiale, par définition, de l'innovation est suivie d'une croissance rapide et, finalement, d'une stabilisation) (Rogers, 1962 ; Croft, 2000).

Plusieurs auteurs, interrogeant les difficultés dans l'opérationalisation de la fréquence, sont amenés à prendre davantage en compte le contexte textuel. Par exemple, Stefanowitsch et Gries (2003), représentants emblématiques d'une position qui insiste sur l'intérêt de l'utilisation de véritables méthodes statistiques et d'indicateurs probabilistes, ont proposé une méthode pour apprécier au moyen d'indices quantitatifs la force de la liaison entre une construction et des items lexicaux susceptibles de saturer les places de cette construction. À l'issue d'une étude de la construction causative en *into* (2007 : 279), ils rencontrent ainsi des faits qu'ils estiment de l'ordre du stéréotype et qu'ils doivent prendre en compte pour interpréter leurs données. Schmid (2010 : 123) prend acte du fait que « we seem to be quite far from having a good grip on the relation between frequency and entrenchment. This is mainly due to the unclear interaction between absolute and relative frequency, or cotext-free and cotextual entrenchment, respectively. » La notion de *cotextual entrenchment*, qu'il introduit, étend considérablement la notion d'*entrenchment* en lui demandant de rendre compte de l'autonomisation et de la conventionnalisation, à travers des processus de routinisation, non plus seulement d'unités du lexique, mais également de phénomènes textuels. On peut faire l'hypothèse que l'utilisation de méthodes quantitatives par les modèles fondés sur l'usage rend nécessaire d'articuler ce modèle avec une perspective textuelle : les fréquences étiques (mesurées empiriquement) ne peuvent être que des indices reflétant une conjonction complexe de causes historiques diverses, interprétables seulement en prenant en compte la diversité des phénomènes variationnels et textuels, et relevant d'une herméneutique plus que d'une modélisation mécaniste.

## 4 Une expérience : fréquence et tactique textuelle

Dans le cadre d'un genre donné, certains mots peuvent acquérir un emploi spécifique en début ou fin de texte, ou à des positions définies par rapport à d'autres unités (chapitres, paragraphes, etc.). Ces faits sont observables quantitativement et apportent de nouveaux matériaux pour la caractérisation des traditions discursives.

L'importance de la linéarité textuelle est soulignée dans de nombreuses perspectives s'intéressant à la définition de types de texte ou, plus précisément, de traditions discursives (entendues de façon générale (Koch/Oesterreicher, 2001: 588) comme un «terme qui englobe les types de textes, les genres (littéraires et non-littéraires), les styles, etc., qui transcendent d'ailleurs les communautés linguistiques»). Ainsi, l'un des exemples souvent utilisés pour illustrer les régularités formelles des traditions discursives (Aschenberg 2003: 1, *inter alia*) est l'*incipit* «il était une fois» des contes. La formule est non seulement ritualisée mais elle doit aussi, pour être valide, être la première phrase du conte. Les propositions descriptives de Rastier pour les genres (Rastier, 1989) distinguent quatre dimensions de description des textes, dont une dimension, dite «tactique», qui porte sur l'enchaînement des éléments sémantiques sur l'axe de la linéarité textuelle. Dans la tradition contextualiste, Hoey (2005: 130) propose la notion de *priming* (amorçage), c'est-à-dire le fait qu'un mot soit associé à des paramètres textuels. Il propose de prendre en considération un amorçage positionnel: «Just as a word may be primed to occur (or to avoid occurring) in first or last position in a sentence, so it may also be primed to occur (or avoid occurring) in first or last position in a paragraph, a section, or a text.».

La mise au jour de telles spécialisations des fonctions lexicales dans les textes peut bénéficier de l'utilisation de grands corpus et de méthodes quantitatives. En effet, les phénomènes en question sont, le plus souvent, davantage des régularités que des contraintes fixes ; elles ne sont observables qu'à des «échelles» élevées (le texte entier), pour lesquelles le dépouillement manuel des phénomènes est malaisé.

## 4.1 Méthode

Pour décrire des spécialisations lexicales dans les textes, j'utiliserai une méthode extrêmement fruste. Il s'agit, d'abord, de «découper» les textes d'un corpus en un certain nombre de «tranches». Par exemple, chaque texte est découpé en dix segments de taille égale. Les premières tranches, correspondant à tous les premiers dixièmes de textes, sont ensuite regroupées pour former un sous-corpus. C'est ce sous-corpus qui est comparé à l'ensemble du corpus pour mettre au jour les mots particulièrement attirés par cette position. J'utilise comme mesure de la sur-représentation des mots dans le sous corpus la loi hypergéométrique (Lafon, 1980).

Ce dispositif permet d'identifier des mots irrégulièrement distribués. Il reste cependant possible qu'un mot, bien que régulièrement distribué, acquière à différentes positions un emploi spécialisé. Afin d'accéder à de tels phénomènes, j'emploie, dans un second temps, les mesures d'attirance entre mots pour identifier les cooccurrents les plus fréquents d'un mot lorsqu'il est en première position, cooccurrents qui peuvent être comparés avec les cooccurrents de ce mot lorsqu'il est, par exemple, en dernière position.

Une difficulté immédiate de ce dispositif est ce «découpage» des textes en un nombre fixé de segments de taille égale. Bien évidemment, ce découpage est un mauvais traitement fait au texte: ces segments ne correspondent à aucune unité interprétable. Il serait plus approprié d'articuler cette quantification à des unités linguistiques: par exemple, de contraster les fréquences d'un mot entre introduction et conclusion, plutôt que de contraster les fréquences entre premiers et derniers dixièmes de textes. Il n'y a pas deux textes pour lesquels «premier dixième» veut dire la même chose. Schegloff (1993) a parfaitement exposé les apories d'un raisonnement où la quantification est appliquée à des objets non-construits: il est absurde d'établir quel est le «nombre de rires moyen par minute» entre deux interactions, puisqu'il faut rapporter un événement comme le rire à des unités interactionnelles, elles-mêmes de longueurs irrégulières, et non à des minutes.

Cependant, outre le fait que des unités comme «introduction» et «conclusion» sont elles-mêmes délicates à établir – et que la méthode proposée vise, précisément, à donner des moyens de définir linguistiquement ces unités dans le corpus considéré –, il s'agit essentiellement, pour l'instant, de vérifier l'hypothèse d'une variabilité des mots en fonction de leur position. Je n'essayerai pas de justifier le nombre de 10 «tranches», ni de faire varier ce nombre pour trouver un «bon» nombre: il n'y a certainement pas de bon chiffre autrement que permettant d'attester cette variation, et il est plus légitime d'assumer son côté linguistiquement arbitraire. L'avantage de cette méthode est qu'elle se prête à différentes unités: il est possible de découper ainsi des textes entiers, mais également les paragraphes d'un texte donné, ou toute autre unité. Les deux corpus utilisés permettront précisément d'observer des variations d'abord au niveau du paragraphe, puis au niveau du texte.

Deux corpus seront utilisés. Le premier est un essai philosophique de G. Deleuze et F. Guattari publié en 1972: *l'Anti-Œdipe*. Ce texte philosophique articule une thématique psychiatrique, dans la perspective de l'anti-psychiatrie, et une perspective politique, et élabore un parallèle entre différents stades psychiatriques et différents stades politiques. Le discours philosophique possède l'intérêt pour une telle recherche d'être peu décrit du point de vue des propriétés matérielles de ses textes. Le discours philosophique est en effet généralement conçu sous les catégories de la terminologie et de l'argumentation, et les sciences humaines et sociales peinent à ne pas hériter du lieu commun philosophique selon lequel seule la philosophie elle-même rend compte des conditions dernières, et donc d'elle-même, et qu'elle ne pourrait être l'objet de déterminations. Citons, entre mille exemples, ce début de compte-rendu d'ouvrage (Goetz, 1998): «La philosophie a une propriété qui la distingue des autres savoirs (et peut-être du savoir en général), et des autres pratiques: on ne peut la définir sans commencer à philosopher. La question de la définition de la philosophie est une question interne à la philosophie. Pour définir la nage ou l'histoire, il faut s'arrêter de nager ou de raconter des histoires.». Ici, il ne s'agit certes pas de définir la philosophie, mais de caractériser certaines propriétés matérielles de l'un de ses textes, sans lui reconnaître de privilège particulier. Dans ce corpus, c'est l'unité paragraphe qui sera considérée: chaque paragraphe est découpé en 10 segments de taille égale, et les spécialisations en début ou fin de

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100332


paragraphe seront recherchées. Peut-on caractériser, par ce biais, le fonctionnement de cette unité dans le texte considéré ?

Le second corpus présenté est constitué de l'ensemble des articles de la rubrique *Opinion* du journal *Le Monde* publiés entre 1987 et 2003. Ce sont ici les spécialisations en début ou fin de textes qui sont considérées, permettant la mise au jour de certaines caractéristiques de ce genre textuel. Si le premier corpus permet d'observer des régularités idiolectales, le second permet d'observer des régularités d'ordre générique.

## 4.2 Variabilités lexicales dans les paragraphes d'un texte philosophique

Afin de caractériser la variation du lexique sur la linéarité du paragraphe, on peut dans un premier temps relever les formes les plus sur-représentées dans les premiers dixièmes de paragraphe d'un côté, et les formes sur-représentées dans les derniers dixièmes de paragraphe de l'autre.

Les mots les plus sur-représentés dans les premiers dixièmes sont ainsi : *Synthèse, nous, comment, production, tâche, formation, avoir, trop, premier*. À un seuil encore significatif : *si, usage, remarquer, désirant, dualisme, postulat, territorial, pratique, empire, infinité, catégorie, thèse, individuel, célèbre, ndembu, souvent, trois, deux, abord, pourtant, comprendre, pouvoir, quantité, éliminer, prétendu, grand*.

Mots les plus sous-représentés dans les premiers dixièmes: *partie, ou, je, homme, ni, et, autres, soi, bref, inscription, alors, son, désir, qui*.

Mots les plus sur-représentés dans les derniers dixièmes : *ni, narcissique, théorie, retirer, inconscient, désir, disperser, grégarité, intermédiaire, réalité, monde, analytique, machine, usine, feu, énonciation, théâtre, alors, chanson, artistique, garant, évolution, forme, cesse, dans, inférieur, schizophréniser, hanter, égaler, centre, analyse, sexe, dur, multiplier, outre, schizes, succession, enculer, brûler, Carroll, Castro, concourir, crotte, disant, faillir, interception, journée, kodak, pessimiste, realitatis, rythme, sain, semblant, signes, traduction, Weismann, art*.

Mots sous-représentés dans les dixièmes dixièmes de paragraphe de *l'Anti-Œdipe* : *comment, si, un, délire, pourtant, ne, pouvoir, valeur*.

L'opposition des mots positivement et négativement associés d'une part, des premiers et derniers dixièmes d'autres part, fait apparaître des constantes sémantiques. Ainsi l'initiale des paragraphes est caractérisée par un vocabulaire philosophique, de la construction d'un raisonnement et de la polémique (*dualisme, postulat, premier, thèse, si, remarquer, prétendu, comprendre, reproche*). Le seul pronom personnel présent est *nou*s. À l'inverse, le lexique sous-représenté concentre tous les pronoms autres que *nous*: *je* et *son*, puis à des taux d'association inférieurs: *mon, vous, lui, me, moi.*

La clausule des paragraphes est caractérisée par quelques mots relevant *a priori* d'un lexique philosophique (*théorie, réalité, monde, analytique*), mais très majoritairement par des mots relevant d'autres domaines (*psychanalyse, art*); il comprends des mots d'un registre populaire (*enculer, crotte*), des noms propres d'auteurs non philosophiques et de marque (*Caroll, Castro, Kodak*), des néologismes (*schizophrénisé*).

Sur le plan des coordinations, on peut relever l'opposition de *ni* et *si* respectivement première forme positive et première forme négative dans les derniers dixièmes de paragraphe. *Si* est d'ailleurs une forme sur-représentée dans les premiers dixièmes.

Bref, l'opposition entre les débuts et les fins de paragraphes semblent porter sur des éléments de thématiques et de registres. Une analyse plus détaillée de la disposition nous permettra de confirmer et d'affiner cette interprétation. L'alternance des variétés, du savant (philosophique) au populaire (et évaluatif), peut être observée plus finement en contrastant des sous-corpus regroupant chacun les n-ième de paragraphes contenant un mot-pôle donné. Je présente ci-dessous les résultats obtenus avec le mot *organe*, l'un des concepts élaborés par ce texte. Les sous-corpus sont donc d'un côté les premiers dixièmes de paragraphe contenant *organe*, de l'autre les derniers dixièmes de paragraphe contenant *organe*.

Les cooccurrents d'une part des premiers dixièmes et d'autre part des derniers dixièmes des paragraphes contenant *organe* offrent des oppositions similaires aux cooccurrents des débuts et des fins de paragraphes présentés ci-dessus. On observe dans l'opposition entre le début et la fin du paragraphe l'opposition entre registre technique et registre évaluatif. Les débuts de paragraphes contenant *organe* sélectionnent, en premier lieu, davantage de mots de haute fréquence relevant du lexique technique de *l'Anti-Œdipe*, comme *machine, socius, flux, produire, quantité, modèle ou code*, tandis que, du vocabulaire technique, seul *machine* est présent dans les fins de paragraphes. Plus généralement, un mot comme *acception*, ou des mots comme *biochimie* ou *biologique*, témoignent du soin portée à la terminologie et indexent le thème dans un registre philosophique classique. Dans le thème des premiers dixièmes de paragraphes contenant *organe*, on relève encore *sociologique* ou *phénoménologique*. Il s'agit donc de la thématisation de l'« emprunt » d'organe à une discipline externe et de sa constitution en concept abstrait, de très grande généralité.

À l'inverse, le sous-corpus des derniers dixièmes de paragraphes contenant *organe* est caractérisé par un vocabulaire peu technique: on note non seulement le pronom *te*, ce qui est conforme à l'opposition déjà observée des pronoms personnels entre les débuts et les fins de l'ensemble des paragraphes, mais également des mots d'un registre concret, voire trivial: *anus, latrine, crotte, cul, excrément*, ainsi que la forme *schizo* (*schizophrénie* est également relevé). Ces mots sont tous de faible fréquence dans l'ensemble du corpus.

Cette opposition des registres est une opposition de fonds sémantiques : au « ton élevé » des débuts de paragraphes s'oppose le «ton bas» des fins de paragraphes. On note aussi qu'à l'abstraction généralisante du début correspond le particularisant et le péjoré de la fin. À l'organe généralisé jusqu'à être un modèle abstrait de toute machine dans le premier cas répond dans le second un organe à l'acception particulièrement restreinte et antithétique: réduit à l'humain, et à un organe concret particulier, celui chargé des connotations les plus négatives. Ce réalisme est fortement évaluatif : l'insistance sur les fonctions (ou l'organe) du corps les moins positivement connotées est une sorte de renversement, où le quotidien et la trivialité sont accueillis dans le discours philosophique. Il s'agit donc non plus de l'organe comme modèle général, mais d'un organe particulier, traditionnellement le plus déprécié : un renversement des valeurs – projet philosophique dont relève ce texte – est observable sur l'axe tactique des paragraphes.

## 4.3 Variabilités lexicales dans un genre journalistique

Si la méthode utilisée s'avère utile pour caractériser une unité comme le paragraphe dans un texte, on peut s'interroger sur son intérêt pour caractériser le palier du texte dans une tradition discursive donnée. Dans le corpus de la rubrique d'opinions du quotidien *Le Monde*, des régularités peuvent en effet être mises au jour au palier du texte. Ici, la norme prépondérante dans le corpus n'est plus idiolectale mais générique.

Le contraste des cooccurrents de *démocratie* en début ou fin de texte permet d'observer des régularités d'ordre rhétorique. Les cooccurrents, dans le premier cas, sont *élection, parti, alternance, président, candidat, droit, ACLU, gauche, parlementaire* et *partisan*. Dans le second cas, ce sont *respect, pouvoir, participatif, autocratie, social, cause, peuple, fondement, impérialistes* et *minorité*. Les deux contextes se distinguent sémantiquement à travers l'ensemble des mots: dans le premier cas, c'est le fonctionnement institutionnel ou la vie publique qui est dominant ; dans le second cas, ce sont des mots relevant davantage de la formulation d'un projet politique, des fins dernières de l'action politique, et du jugement moral. On retrouve entre ces deux dominantes sémantiques l'opposition entre la démocratie comme «procédure» ou comme «régime» selon les termes de Castoriadis. Cette opposition entre début et fin des textes, où le début est davantage factuel et institutionnel, et où la fin voit l'invocation de valeurs communes, n'est pas, rhétoriquement, excessivement étonnante ; elle peut néanmoins par ce biais être caractérisée précisément.

Cette opposition n'est pas limitée au seul mot *démocratie*, bien que, naturellement, le point d'entrée choisi détermine un point de vue sur cette opposition générale entre début et fin des textes. Ainsi, le mot *Europe* permet d'observer également une opposition entre une thématique institutionnelle et factuelle et une

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100332


thématique de discussion de la nature du projet politique. Parmi les premiers cooccurrents (*après, semaine, mardi, jeudi, premier, élection, an, jour, début, mai, Danone, depuis, octobre, novembre, mois, dimanche, union, venir, candidat*) les éléments essentiellement chronologiques sont sur-représentés (comme c'est le cas avec de nombreux autres points d'entrée); parmi les derniers (*devoir, que, mais, pouvoir, ne, défense, sans, encore, il, pas, Tocqueville, nous, défi, contribuable, peut, intérêt, bon*), la modélisation, la première personne, les grandes références comme *Tocqueville*, etc., témoigne d'un emploi davantage axé sur le projet politique.

La différenciation des contextes d'une unité lexicale sur l'axe de la linéarité textuelle n'est donc pas observable seulement dans le cadre d'un texte et d'un usage très spécialisé; on la retrouve également à l'échelle d'un genre.

## 5 Conclusion

La méthode proposée permet de montrer des phénomènes de variation du signifié lexical de certains mots, à travers des variations de ses cooccurrents, à différentes positions textuelles. Si l'on ne peut aller jusqu'à parler d'une variation d'acception – à aucun moment il n'a été possible d'observer une distribution complémentaire entre différentes acceptions – on peut parler d'une variation fortement réglée d'emploi. Ces variations fournissent un matériau pour caractériser le rôle sémantique de différentes unités, particulièrement des unités qui, comme le paragraphe, sont peu décrites ou bien ont un statut relativement faible. Cette méthode permet également de caractériser différentes normes textuelles (comme l'idiolecte ou le genre), en fonction du corpus considéré. Les découpages produits (en dix segments) sont cependant arbitraires et doivent être considérés comme des points d'entrée pour une interprétation: ils ne sont pas articulés à une catégorie descriptive.

Cette variation des contextes d'emploi n'est pas moins forte que celle que l'on observe, d'un point de vue également distributionnel, en distinguant différents usages en fonction de différents types de cooccurrents. De la même façon que les occurrences d'un mot se distinguent en fonction de certains cooccurrents immédiats importants, de même les occurrences d'un mots se distinguent, dans le cadre d'un genre, en fonction de leur position. Le contexte sémantique n'est donc pas constitué par le seul environnement immédiat, mais également par la position textuelle, dans le cadre de certains genres.

De nombreux phénomènes de régularité textuelle influençant directement l'interprétabilité des données quantitatives sont encore peu explorés. Ces résultats plaident, avec d'autres, pour une prise en compte des phénomènes textuels dans l'utilisation de données quantitatives pour la description sémantique. Ils plaident également pour une certaine prudence dans la tentation de considérer la fréquence comme un phénomène interprétable en termes causalistes et univoques plutôt que relevant de procédures interprétatives.

## Références bibliographiques

Aschenberg, H. (2003). *Diskurstraditionen - Orientation und Fragestellungen*. In Aschenberg, H. & Wilhelm, R. (eds) *Romanische Sprachgeschichte und Diskurstraditionen*. Tübingen : Gunter Narr Verlag, 1–18.

Baayen, R. H. (2009). Corpus linguistics in morphology: morphological productivity. In Lüdeling, A. & Kyto, M./ McEnery, T. (eds), *Corpus Linguistic, An International Handbook*. New York, Berlin: Walter de Gruyter, 899–919.

Bybee, J. (1985). *Morphology: A Study of the Relationship between Meaning and Form.* Philadelphia: John Benjamins.

— (2003). Mechanisms of change in grammaticalization : The role of frequency. In Janda, B. J. R. (eds), *The handbook of historical linguistics*. Malden, MA: Blackwell, 602-623.

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

Bybee, J. & Hopper, P., (2001). Introduction. In Bybee, J. & Hopper, P. (eds), *Frequency and the Emergence of Linguistic structure.* Amsterdam : John Benjamins, 1-24.

Bybee, J. & Thompson, S. (2000). Three frequency effect in syntax. *Berkeley Linguistic Society, 23,* 378-388.

Cohen, M. (1949). Sur la statistique linguistique. *Conférences de l'institut de linguistique de l'université de Paris, 9.* Paris : Klincksieck, 7-16.

Croft, W. (2000). *Explaining language change. An evolutionary approach*. Essex : Pearson Ed. Ltd.

— (2008). On iconicity of distance. *Cognitive Linguistics, 19/1,* 49-58.

Coseriu, E. (1982 [1952]). Sistema, norma y habla. In *Teoría del lenguaje y lingüística general.* Madrid : Gredos, 11-117.

Du Bois, J. (1985). Competing motivations. In Haiman, J. (ed), *Iconicity in Syntax.* Amsterdam : Benjamins, 343-365.

Geeraerts, D., Grondelaers, S. & Bakema, P. (1994). *The structure of lexical variation: Meaning, naming, and context*. Berlin: Mouton de Gruyter.

Geeraerts, D. (2010). Recontextualizing Grammar: Underlying Trends in Thirty Years of Cognitive Linguistics. In Tabakowska, E., Choinski, M. & Wiraszka, L. (eds), *Cognitive Linguistics in Action, From Theory to Application and Back.* Berlin: de Gruyter/Mouton, 71-102.

Glessgen, M.-D., (2007). *Linguistique romane.* Paris : Armand Colin.

Glynn, D. & Fischer, K. (ed). *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*. Berlin : Mouton de Gruyter.

Goetz, B. (1998). Compte-rendu de Jean Pierre Faye, Qu'est-ce que la philosophie ?. Paris, Armand Colin, 1997. *Le portique, 1*, non paginé.

Greenberg, J. H. (1966). *Language Universals*. The Hague : Mouton.

Haiman, J. (1994). Ritualization and the Development of Language. In Pagliuca, W. (ed), *Perspectives on Grammaticalisation.* Amsterdam/Philadelphia : Benjamins.

Haspelmath, M. (2008). Frequency *vs.* iconicity in explaining grammatical asymmetries. *Cognitive Linguistics, 19 (1),* 1-33.

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

Hoey, M. (2005). *Lexical Priming*. Oxon : Routledge.

Hoffmann, S. (2004). Are low-frequency complex prepositions grammaticalized. On the limits of corpus data – and the importance of intuition. In Lindquist, H. & Mair, C. (eds), *Corpus Approaches to Grammaticalization in English.* Amsterdam : John Benjamin, 171-210.

Hopper, P. (1997). When 'Grammar' and Discourse Clash. In Bybee J., Haiman, J. & Thompson, S. A. (eds), *Essays on language function and language type.* Amsterdam/Philadelphia : Benjamins, 231-247.

Hopper, P. & Traugott, E. (1993). *Grammaticalization*. Cambridge: Cambridge University Press.

Käding, F. W. (1897). *Häufigkeitswörterbuch der deutschen Sprache*, Steglitz.

Koch, P. & Oesterreicher, W. (2001). Gesprochene Sprache und geschriebene Sprache. In Holtus, G., Metzeltin, M. & Schmitt, C. (eds), *Lexikon der Romanistischen Linguistik.* Tübingen : Max Niemeyer Verlag, 584-627.

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

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

Langacker, R. (1987). *Foundations of cognitive grammar, I.* Stanford : Stanford University Press.

Legallois, D. & François, J. (2011). La Linguistique fondée sur L'usage : parcours critique. *Travaux de linguistique, 62*, 7-33.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100332


Mair, C. (2004). Corpus linguistics and grammaticalisation theory: Statistics, frequencies, and beyond. In Lindquist, H. & Mair, C. (eds), *Corpus Approaches to Grammaticalization in English.* Amsterdam : John Benjamin, 121-150.

Mańczak, W. (1969). Quelques réflexions sur la doctrine de Noam Chomsky. *Linguistics, 49*, 18-27.

Martinet, A. (1955). *Économie des changements phonétiques.* Berne : Francke.

Muller, C. (1992 [1973]). *Initiation aux méthodes de la statistique linguistique.* Paris : Champion.

Rastier, F. (1989). *Sens et textualité*. Paris : Hachette.

Rogers, E. M. (1962). *Diffusion of Innovations.* New York/London: The Free Press.

Saussure, F. de (1995 [1916]). *Cours de linguistique générale.* Paris : Payot.

Schegloff, E. (1993). Reflections on Quantification in the Study of Conversation. *Research on Language and Social Interaction, 1, 26*, 99-128.

Schmid, H.-J., (2007). Entrenchment, salience, and basic levels. In Geeraerts, D. & Cuyckens, H. (eds), *The Oxford handbook of cognitive linguistics.* Oxford : Oxford University Press, 117-138.

Schmid, H.-J. (2010). Does frequency in text really instantiate entrenchment in the cognitive system? And do we have a quantitative grip on either of them?. In Glynn, D. & Fischer, K. (eds), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches.* Berlin : Mouton de Gruyter, 101-133.

Stefanowitsch, A. & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8/2*, 209-243.

Troubetzkoy, N. S. (1986 [1939]). *Principes de phonologie.* Paris : Klincksieck.

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

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

— (1949) *Human behavior and the principle of least effort*, Cambridge (MA) (Reprint 1965, New York / London).

---

[1] 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)

[2] Les remises en cause plus systématiques de ces distinctions et les propositions de parler plutôt d'un continuum, permettant d'intégrer davantage les faits de fréquence, sont particulièrement vives dans les descriptions basées sur l'usage ; cf. la remise en cause de la dichotomie grammaire/lexique dans les grammaires basées sur l'usage (Bybee et Hopper, 2001 : 2) ; la remise en cause des parties du discours par Sinclair (Bybee et Hopper, 2001 : 4-5 ; Hopper, 1997 : 235). Cette insistance sur le continuum se retrouve jusque dans la distinction entre unité linguistique et unité non linguistique (Langacker, 1987 : 59) : « the linguistic character of a unit is sometimes a matter of degree. »

[3] L'une des difficultés de la fréquence dans les perspectives de l'usage est justement qu'elle oscille entre un statut abstrait, théorique, et un statut opératoire, descriptif : « Indeed, there seems to be a general consensus that a relatively high discourse frequency is a prerequisite for a particular form to grammaticalize in the first place [...] although specific "threshold" frequencies have, to the best of my knowledge, never been suggested in the literature. » (Hoffmann, 2004 : 172)

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

[5] 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). Ces éléments sont particulièrement repris chez les saussuriens (cf. Sechehaye, 1908 : 184, ou la recension proposée par Coseriu dans 1982 : 13-24, notamment les définitions de Wartburg, Pederson, Porzig et Gardiner). La définition par Coseriu (1982 : *passim*) de la langue (comme de tout niveau d'abstraction) comme un faisceau d'isoglosses, une coïncidence dans la variété des parlers, relève du même argument quantitatif.

[6] « 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 précédé d'un fait, ou plutôt d'une multitude de faits similaires dans la sphère de la parole » (1995 : 138-139).

[77] « The Saussurean dichotomy between language and parole creates an internally divided grammar, a conception of language with so to speak a hole in the middle. »

[8] Elle est présente également dans le couple saussurien « esprit de clocher » / « force d'intercourse » (1995 : 281).

[9] Cette fréquence intuitive est appelée « informal quantification » par Schegloff (1993) ; on trouve également « subjective frequency » (*vs* « objective frequency ») dans la littérature psycholinguistique.

[10] *Inter alia* Lafon (1980 : 127), Tournier (1980 : 194), Heger (1969 : 56) : une fréquence mesurée n'a de pertinence que relativement à un genre, à une pratique, à un usage, et il n'y a pas de corpus représentatif d'une langue. La prise en compte de cet aspect n'est pas encore très avancée dans les perspectives basées sur l'usage où l'on utilise encore des fréquences supposée valoir en langue ; Hoffmann (2004 : 196) dit ainsi que les méthodes quantitatives sont « based on the assumption that the corpus is a representative sample of language use ».

[11] La question de l'*entrenchment* n'est d'ailleurs pas étrangères aux positions de Zipf sur le principe du moindre effort. L'ouvrage de Bybee et Hopper (2010) s'ouvre ainsi sur un hommage à Zipf : « [Zipf] anticipated many of the themes of more recent investigations of the relationship between frequency and structure ».

[12] La fréquence est ainsi mobilisée comme « explication » d'un très grand nombre de phénomènes : outre tout ce qui relève du changement, elle vaudrait aussi pour la syntaxe : « [the data] support the view that what has been called 'syntactic cohesion' is frequency of occurrence, the fact which determines the strength of the association between the first element and the second one » (Bybee, 2001 : 355). À chaque fois, certes, un rapport existe ; mais est-il suffisamment précis pour en tirer des faits ?

[13] Hoffman (2004 : 189) souligne particulièrement la dimension mécaniste de la conception « théorique » de la fréquence, par exemple chez Bybee, qui explique (Bybee 1985: 117) : « If we metaphorically assume that a word can be written into the [mental] lexicon, then each time a word in processing is mapped onto its lexical representation it is through the representation was traced over again, etching it with deeper and darker lines each time. Each time a word is heard and produced it leaves a slight trace in the lexicon, it increases in lexical strength. ».