TYPOLOGIE DES CONCEPTS DE LINGUISTIQUE : ÉVALUATION ET ÉLABORATION EN CORPUS DE CRITÈRES DISCRIMINANTS

Céline POUDAT
CORAL, Université d'Orléans

SOMMAIRE

1. Introduction

Le présent article propose un ensemble de critères typologiques pour discriminer les concepts de linguistique, éprouvé en corpus à partir d'une collection homogène en genre de 224 articles extraits de 32 numéros de revues (soit 11 revues) francophones de sciences du langage essentiellement publiés autour de 2000.

Si les dictionnaires sémantiques, les ontologies et les bases terminologiques sont en pleine expansion, la description des concepts scientifiques en tant qu'unités textuelles et textualisées est encore peu développée ; le contexte des termes ou des concepts est certes pris en compte par les dictionnaires contextuels, et on recense bien des travaux dans lesquels la sélection des termes est objectivée en corpus (par une analyse des spécificités par exemple, L'homme 2004), mais peu d'entreprises cherchent à caractériser les concepts en tant qu'unités méso-sémantiques (palier intermédiaire entre la micro-sémantique du mot et la macro-sémantique du texte) ; dans cette perspective, qui a montré son intérêt et son efficacité dans plusieurs études récentes (Loiseau 2003, Valette 2003), les concepts, que l'on peut décrire comme des thèmes, sont potentiellement corrélés à des marqueurs ou à des formes expressives relevant de tous les niveaux de l'analyse linguistique (Rastier 2003).

L'analyse thématique des concepts scientifiques étant encore à son stade exploratoire, on manque encore de critères pour discriminer les objets et les formes sémantiques. La présente étude vise ainsi à éprouver différents critères typologiques en corpus : fréquence et répartition des concepts dans les textes du corpus, corrélations morphosyntaxiques, co-occurrents lexicaux, incidence du style d'auteur et du numéro thématique de revue et configurations tactiques. On insistera sur le caractère non isolé des critères proposés, qui discriminent d'ailleurs souvent les mêmes phénomènes, mais sur des plans distincts.

Bien que le présent article porte sur un corpus génériquement et domanialement homogène, on peut penser que certains des critères discriminants établis sont généralisables à d'autres disciplines des Sciences Humaines.

2. Fréquence et répartition

Supposés constituer un mode d'accès privilégié aux thèmes scientifiques linguistiques, ce sont les substantifs les plus représentés que nous avons choisi d'extraire. Étant donné le peu de corpus de comparaison disponibles, le recours à une analyse des spécificités a dû être écarté[1].

[1] Mentionnons toutefois que nous avons par ailleurs mis à jour, au sein d'une étude comparative des discours linguistique, philosophique et critique (Loiseau, Poudat et Ablali 2006) certains items spécifiques à la linguistique, qui renvoient la discipline à ses observables (*verbe, phrase, énoncé, mot*, etc.), en excluant la *langue* ou le *sens,* que s'approprient également la philosophie et la critique ; il nous semblerait peu pertinent d'exclure certains objets parce que d'autres disciplines les empruntent.

Les variations flexionnelles (singulier/pluriel) ont été prises en compte, dans la mesure où le trait « nombre » n'indique pas seulement la pluralité : *langue* et *langues* sont ainsi deux concepts linguistiques distincts.
Bien que le *texte* l'emporte sur la *phrase* en termes de fréquences absolues (1313 *vs*. 1237), on observe qu'il apparaît pourtant dans un nombre plus restreint d'articles (121 *vs*. 143), et que c'est finalement *l'énoncé* qui domine selon ce dernier critère (149 textes) : on voit là l'intérêt de prendre la fréquence *et* le nombre de textes d'apparition de l'occurrence.
Nous avons donc ordonné l'ensemble des substantifs en prenant en compte les deux paramètres, ce qui nous donne le classement suivant (seuls les 20 premiers noms communs au singulier sont présentés dans le tableau qui suit) :

Rang	Substantif	Fréquence absolue	Textes
1	sens	2136	200
2	forme	1840	208
3	cas	1693	214
4	langue	2037	176
5	type	1650	205
6	relation	1654	183
7	objet	1500	191
8	point	1297	210
9	discours	1687	155
10	contexte	1568	157
11	rapport	1223	196
12	analyse	1263	185
13	verbe	1632	142
14	sujet	1351	165
15	fait	1057	194
16	fonction	1031	187
17	question	986	191
18	énoncé	1206	149
19	phrase	1237	143
20	partie	917	190

Graphique : Substantifs au singulier ordonnés par fréquence et par textes

On distingue globalement deux types de substantifs : les candidats concepts de linguistique (en gris), et les substantifs relevant visiblement de la méthodologie scientifique à l'œuvre dans les articles. On observe ainsi un intérêt prononcé de la discipline pour le *sens*, qui détrône la *langue*, objet pourtant intuitivement premier de la linguistique. Le *discours* est également très honorablement représenté, tandis qu'on observe un intérêt particulier pour le *verbe*, dont la forme fléchie plurielle détient le second rang des substantifs au pluriel (1225 occ. pour 117 textes).
Les substantifs restants relèvent de la logique (*relation, rapport*), de la typologie (*cas, type*) ou sont trop ambigus pour renvoyer directement à un objet linguistique (*sujet, objet, fonction, point*).
Les substantifs relevés au pluriel corroborent généralement le classement précédent, bien qu'on observe certains substantifs résolument déterminés en nombre : ainsi, « sens » est le substantif singulier le plus relevé dans le corpus (2186 occ. au singulier) et il est notable qu'il soit globalement peu employé au pluriel (179 occ. / 83^{e} rang). Il en va de même pour *discours* (9ème substantif), qui est relevé 1732 fois au singulier et seulement 146 fois au pluriel (118^{e} rang), ou encore pour *langage* (1208 occ. au singulier *vs*. 30 au singulier). S'il est intuitif que *le* discours et *les* discours, ou *le* langage et *les* langages ne renvoient pas aux mêmes concepts linguistiques, de tels écarts demeurent surprenants.
On ne note pas de différences aussi importantes à l'inverse, mais plusieurs substantifs qui relèvent davantage de la méthodologie que d'une thématique strictement linguistique, sont essentiellement pluriels : *données, conditions, caractéristiques, phénomènes, traits, critères, résultats* ou *contraintes*.

Soulignons que ce bref panorama des candidats concepts linguistiques ne porte que sur les *hautes fréquences*. Malgré leur intérêt descriptif, les hapax et les éléments moins – ou plus inégalement – représentés ont été globalement écartés des analyses qui suivent, dans la mesure où ils se prêtent difficilement à l'analyse statistique.

3. Co-occurrents et corrélats lexicaux : exploration de trois paliers de régulation linguistique

L'examen des co-occurrents lexicaux est particulièrement crucial dans le processus de qualification thématique des candidats concepts : si leur nombre implique une plus ou moins grande stabilisation de la forme et si les écarts observés mesurent le degré de corrélation contextuelle entre les mots, leurs éventuelles intersections sémiques – qui indique la présence d'isotopies – les font accéder au statut de *corrélats sémantiques* (Rastier).

Cette méthode, qui a montré sa pertinence dans l'examen des textes littéraires, linguistiques et philosophiques (*e.g.* Valette 2003, Bourion 2001, Loiseau 2003), permet de faire émerger des phénomènes linguistiques non-, voire contre-intuitifs, qui sauraient difficilement être appréhendés par d'autres biais. On soulignera qu'en matière de textes linguistiques ou philosophiques, elle a surtout été éprouvée au palier individuel de l'auteur (Deleuze chez Loiseau et Guillaume chez Valette).

Bien que le niveau de normalisation du *style d'auteur* soit naturellement pertinent pour évaluer l'appropriation singulière des concepts d'un domaine ou d'une discipline scientifique, deux paliers de niveaux supérieurs nous semblent également significatifs : celui du *genre*, et nous limiterons nos investigations à l'*article de revue*, et celui du *thème*, ou du *numéro thématique de revue* si l'on s'intéresse au domaine linguistique – la plupart des numéros de revue étant organisés autour d'une thématique ou d'une problématique fédératrice.

Nous adopterons ainsi une démarche progressive et descendante (de la généralité du genre à la singularité du style), en parcourant ces trois paliers de normalisation du discours scientifique linguistique. Dans la mesure où il n'est pas envisageable ici d'analyser les co-occurrents de l'ensemble des candidats concepts du corpus, nous avons choisi de nous concentrer sur les objets *sens* et *langue,* choix motivé par leurs hautes fréquences et leur qualité d'objets / objectifs de descriptions linguistiques privilégiés.

3.1. Le palier générique (corpus ASLF)

On notera d'abord qu'en dépit de leurs fréquences distinctes (2471 occ. de *sens vs*. 1798 occ. de *langue*), *sens* et *langue* ont quasi le même nombre de co-occurrents : 209 pour *sens vs*. 207 pour *langue*. On relève des écarts de corrélation plus importants pour *langue* que pour *sens* : *langue* draine ainsi plus de lieutenants stabilisés que *sens* (*e.g. langue des signes, parlée, française, maternelle, naturelle, étrangère, usuelle, de spécialité, courante*, etc.), eux-mêmes corrélés à des co-occurrents renvoyant à des contextes spécifiques.

Ce phénomène entraîne des difficultés d'évaluation dans la mesure où les premières corrélations obtenues ne sont pas nécessairement caractéristiques du corpus : elles peuvent facilement être liées à l'un de ses sous-ensembles singuliers (un numéro thématique le plus souvent). Par exemple, *langue* apparaissait d'abord corrélée à *signes, sourds, LSF, entendants*, etc., éléments bien spécifiques à *un* numéro thématique singulier du corpus, dédié à la *langue des signes*.

On voit là toute la difficulté que pose l'observation des formes de haute fréquence, qui drainent de nombreux figements qui s'autonomisent de la notion première.

On observe le même phénomène pour *sens*, qui n'a qu'un rôle de formant dans de nombreux figements : *sens littéral / sens figuré, sens strict / sens large, sens commun…* auxquels on peut adjoindre *immanence du sens, sens distributionnel, donation du sens,* [unité] *porteuse de sens*, *compositionnalité du sens…* si l'on restreint le contexte pris en compte du paragraphe aux 50 caractères avoisinant la notion.

Notons que les deux objets sont corrélés à *Saussure*, et que la *langue* est corrélée à la dichotomie saussurienne *langue / parole* (*parole* et *Saussure*) tandis que *sens* s'inscrit dans la problématique du *signe* (*signifié, signe, signifiant, Saussure*).

3.2. Le niveau du numéro thématique

Comme nous avons déjà pu l'observer avec *langue / langue des signes*, le numéro thématique de revue participe substantiellement à la régulation thématique du genre et du domaine linguistique.

Le différentiel entre le niveau du genre et celui du numéro thématique permet ainsi de mettre à jour les dominantes thématiques spécifiques des numéros et à terme, d'évaluer ce qui constituerait un noyau dur conceptuel de la discipline linguistique.
Si l'on prend par exemple un numéro thématique spécifique, la revue *Contexte(s)*[1] ici, on observe que la présence du lexème *contexte* dans les premiers co-occurrents de *sens* s'avère en grande partie liée au numéro *Contexte(s)* : au sein du corpus *Contexte(s)*, l'item *contexte* est le cinquième co-occurrent de *sens*, tandis qu'il n'est que le 63e lorsqu'on considère le reste du corpus d'articles.

Premiers co-occurrents de *sens* : Sens (34.66), classique (6.06), établissement (5.60), le (5.50), contenu (5.50), littéral (5.35), contexte (5.33), au (5.19), virtuel (4.80), commun (4.77), hors (4.77), notion (4.64), message (4.51), mot (4.46), approche (4.43)

Les co-occurrents de *sens* diffèrent d'ailleurs fortement aux niveaux de la revue et du reste du corpus : on ne retrouve pas la plupart des items obtenus d'un corpus à l'autre. Le concept de *sens* dans la revue observée ne semble pas débattu en tant que tel, mais relativement à la notion de contexte ; en d'autres termes, on ne s'intéresse au *sens* que par rapport au *contexte*, d'où le figement *sens littéral* (en contexte zéro) et la récurrence de la notion de *contenu* (topos : le sens n'a pas de contenu sans prise en compte du contexte) :

> Lakoff et Johnson (La métaphore dans la vie quotidienne 1981, p. 21) montrent, à travers la métaphore du conduit, que l'idée d'un **sens contenu** dans une forme linguistique est inhérente à notre perception de la langue : on parle de « faire passer, donner une idée, introduire une idée dans une phrase, d'une phrase vide de **sens**. » (Les contextes de contexte La notion de contexte dans les Page: 24 c (41ème occ.))
>
> […] il importe de dégager les processus qui permettent l'établissement d'un **sens** en soi, **contenu** dans la langue, indépendamment du contexte. (Production et interprétation du sens : la notion de contexte Page: 279 c (179ème occ.))

On observe également une co-occurrence de *sens* et de la préposition *hors*, qui renvoie encore une fois au contexte, sous des formes différentes (*hors contexte, hors circonstance, hors langue, etc.*).
Contrairement à *sens*, qui semble être un concept finalement caméléon, dont les co-occurrents – et les acceptions – varient selon l'objet observé, *langue* est un concept relativement plus stabilisé, qui semble invariablement porter les traits [Saussure], [système] [française] et [code]. On observera toutefois que l'objet *contexte* ne semble pas spécifiquement recourir à la dichotomie *langue / parole*, la *parole* étant absente des co-occurrents de *langue* dans l'ensemble du numéro.

3.3. Le palier du style

Afin de parfaire et d'approfondir notre description de la stabilisation des deux concepts et de leurs co-occurrents par paliers, nous avons cherché à observer leur stabilité d'un auteur à l'autre, à partir d'un corpus de 118 articles de 12 linguistes français accrédités dans le champ[2].
Les deux concepts sont d'abord très inégalement employés d'un auteur à l'autre : Kleiber recourt à *sens* 13 fois plus que Combettes, tandis que Bergounioux emploie *langue* 25 fois plus que Rabatel.
Le concept de *langue* semble encore une fois plus stabilisé que celui de *sens*, dans la mesure où ses co-occurrents varient peu d'un corpus à l'autre – les deux corpus d'étude étant pourtant bien distincts : la *langue* est ainsi invariablement associée à *linguistique, Saussure* et *système* et aux figements *langue française, langue maternelle* et *langue parlée*, qui apparaissent d'ailleurs plus corrélés à certains auteurs (respectivement Bergounioux, Neveu et Authier).
Outre ces figements, on observe des dominantes thématiques chez certains auteurs, qui renvoient à leurs intérêts et cadres de recherche généraux : *langue* s'inscrit dans une perspective historique chez Bergounioux (*histoire, était, diachronique, etc.*), diachronique chez Combettes (*état(s), ancienne, anciens, catégories, textes, système, trace, surviennent, permettent, moderne,*

[1] Schmoll (*éd.*), Scolia vol. 6, Strasbourg, 1996.
[2] Authier, Barbéris, Bergounioux, Combettes, François, Kerbrat, Kleiber, Moirand, Neveu, Rabatel, Rastier et Siblot. Soulignons que nous avons mis à jour dans des études précédentes (Poudat et Rinck 2005, 2006) le caractère significatif des styles en linguistique aux niveaux morphosyntaxique et lexical.

existence, faits, phénomènes, corpus, vestiges, etc.) ou didactique chez Rabatel (*élèves, maîtres, initiation, grammaire, appris, école, aider, etc.*).
Certains cadres semblent d'ailleurs plus définis et plus rigides que d'autres : il en va ainsi du cadre praxématique dans lequel s'inscrivent Barbéris et Siblot : chez Barbéris par exemple, *sens* est le formant du concept *production de sens.* Le *sens* s'inscrit ici dans une conception praxématique et il articule les objets linguistiques (*préposition, complément, verbes, lexical*) au praxématique (*production, programmes, état, producteur, gestalt, praxémique etc.*). La *langue* (*parlée*) s'inscrit également dans un cadre praxématique, d'où les co-occurrents linguistiques et marxistes *classes, discours, légitime, dominante, souterraine, registres, populaire, etc.* Bien que la *langue* soit discutée en termes sémiotiques saussuriens (*signes, Saussure, nomenclature, système, etc.*), et aux côtés du *langage* (*langage, langagières*) chez Siblot, on observe une conception praxématique du *sens* (co-occurrents *production, praxème, récepteur, procédures, praxis, sociales, programmes, produire*) appliquée cette fois à l'objet *texte* (*texte(s), clôture*).
Le concept de *sens* varie ainsi de manière plus significative et apparaît comme plus discuté que celui de *langue*, qui est généralement corrélé à l'ensemble des items tissant le système conceptuel général de l'auteur : on le trouve en effet associé à des éléments plus spécifiques participant à sa définition ou à sa discussion. Ainsi, il s'inscrit dans la problématique du *signe* chez Bergounioux, qui le distingue du *signifié* (8.98, premier co-occurrent observé) ; on observe ainsi un réseau conceptuel de corrélats articulés autour de cette question : *union* (signifiant/signifié), *unité*, *signe*, *substitution* (de sens à signifié), *signifiant*, *Saussure*, etc. Outre cette isotopie, *sens* est significativement corrélé à *homme* (6.52, rang 2). *Sens* semble ainsi permettre d'articuler l'homme au signe, et par extension, à la langue et au langage.
Si *sens* et *signifié* s'interdéfinissent chez Bergounioux, on le trouve aux côtés de *signification* chez Rastier (premier co-occurrent relevé, écart de 15.59) – *sens* et *signification* partagent d'ailleurs les mêmes corrélats, ce qui indique bien un emploi concomitant des deux notions. *Sens* s'inscrit ainsi dans un environnement sémiotique (*contenu, carré (sémiotique), signifié, signe*) et sémantique – on relève ainsi des éléments qui renvoient à différents modèles du sens (*contenu, mimesis, immanent, texte*…) ; bien qu'originellement distincts et susceptibles de polémiques, ces éléments sont conciliés dans le cadre sémantique rastiérien.
Enfin, on trouve *sens* associé à un sème [+ instable] chez Authier, lié à la perspective énonciative de l'auteur, d'où *risque, (référence) actuelle, ponctuellement, fixité (du signal), équivoque, paraphrasable, polysémie, effets (de sens),* etc., tandis qu'il apparaît particulièrement stabilisé chez Kleiber, et quasi exclusivement associé à des éléments partageant le sème [+stable] ou [+déterminé] : (*sens) descriptif* (8.50) – qui est d'ailleurs quasi figé chez l'auteur, *détermine* (7.19), *conventionnel* (6.57), *dénominatif* (6.49), *stable* (6.16), *représentationnel* (6.02), *déterminé* (6.02), *conditions* (5.96), *psychologique* (5.96), *codé* (5.62), *stables* (5.36), *instructionnel* (5.36).
On observe ainsi que *sens* constitue un meilleur point d'entrée dans les systèmes conceptuels développés par les auteurs observés : cette distinction pourrait éventuellement participer à celle de concepts de fond / concepts de forme proposée par (Rastier, 2003). *Langue* serait ainsi un concept de fond disciplinaire peu débattu tandis que *sens*, qui paraît moins stabilisé, serait une forme plus discutée.
De manière générale, on peut opposer les concepts dont les co-occurrents sont d'autres concepts des entrées dont les co-occurrents sont des exemples ou des objets de description linguistique : dans le premier cas, le concept semble (inter)défini, voire peut-être débattu, tandis que dans le second, il semble essentiellement instrumental ou méthodologique.
Si on ne considère que les concepts interdéfinis, on peut *a fortiori* opposer les concepts inscrits dans des cadres théoriques et méthodologiques (*e.g.* Praxématique, SI, Sémantique de la référence klébérienne, et cadre énonciatif d'Authier) des concepts ponctuellement débattus où le concept reste de faible fréquence, ou n'est pas pivot dans un système conceptuel – *e.g.* cadre historique de Bergounioux (sens/signe).

4. Corrélations morphosyntaxiques

Nous avons ensuite retenu une sélection de substantifs parmi les hautes fréquences relevées, auxquelles ont été par hypothèse adjoints *sémantique*, *énonciation* et *cotexte* :

SENS : SENS :Nsg, SENS :Npl
SEMANTIQUE : SEMA :Nsg, Npl, SEMA :Asg, Apl

LANGUE : LGUE :Nsg, Npl
DISCOURS : DISC :Nsg, Npl, Asg, Apl (discursif (ve) (s))
PAROLE : PAR :Nsg
LANGAGE : LANG Nsg, Npl, Asg, Apl (langagier, langagière (s))
TEXTE : TXT : Nsg, Npl, Asg, Apl (textuel(le)(s))
CORPUS : CORP Nsg, Npl
INTERPRETATION : INTER Nsg, Npl, Asg, Apl (interprétatif, interprétative)
CONTEXTE : CONT Nsg, Npl, Asg, Apl (contextuel(le)(s))
COTEXTE : COT Nsg, Npl
ENONCIATION : ENONC Nsg, Npl, Asg, Apl (énonciatif(ve)(s))

Dans la mesure où elles sont fondées sur le même morphème, les formes adjectivales des entrées ont également été prises en compte.
Les corrélations des entrées entre elles et avec le jeu de descripteurs morphosyntaxiques ont été observées et c'est sur les textes entiers (avec exemples et citations) qu'ont été effectuées les analyses. L'entreprise peut paraître discutable, mais il s'avère que si l'on extrait les exemples du corpus, il n'est plus possible de voir si les formes sont employées dans des textes contenant des exemples, ceux-ci étant particulièrement identifiables (*v.* Poudat 2006).
Étant donné les différences d'échelle des deux types de données[1], les substantifs sont globalement corrélés entre eux et peu de corrélations négatives significatives (seuil > -0.2) ont pu être relevées. À l'inverse, plusieurs corrélations positives très élevées (> +0.7) ont pu être observées.
De manière générale, les corrélats des candidats observés diffèrent au singulier et au pluriel, ce qui suppose qu'ils renvoient à des objets distincts. La différence est particulièrement visible si l'on s'intéresse à *sens* : si les deux formes sont corrélées entre elles (+0.33), leurs corrélats diffèrent d'abord en nombre, eu égard à la représentation dix fois plus élevée de *sens* au singulier (14 corrélations significatives au singulier *vs*. 6 au pluriel).
Sens au singulier est fortement corrélé à *contexte* et à *interprétation* (respectivement +0.47 et +0.41), et ses corrélats sont d'ailleurs essentiellement des candidats concepts au singulier : l'entrée est ainsi corrélée au *contexte* et non aux *contextes*, au *texte* et non aux *textes.* En outre, *sens* au singulier est négativement corrélé aux marques de formalisation (symboles, abréviations linguistiques, etc.), aux numéraux et aux parenthèses, caractéristiques des textes plus appliqués : le concept serait ainsi plus représenté dans les textes à dominante théorique.
Ces données conduisent à penser que la forme *sens* au singulier renvoie bien à une lexicalisation privilégiée de concept ; au contraire, la forme plurielle de *sens* est corrélée aux marqueurs caractéristiques des textes plus appliqués – et plus exemplifiés (pronom personnel *tu*, symboles linguistiques ?, *, ! et # et connecteurs d'exemplification). En ce sens, *sens* au pluriel serait une forme peu discutée, voire instrumentale.
On observe un phénomène similaire pour les deux formes singulier et pluriel de *corpus* – qui partagent d'ailleurs les mêmes corrélats : si on les observe aux côtés des *textes* et des *discours*, les deux formes sont corrélées à différentes caractéristiques des textes plus exemplifiés (interjections, connecteurs d'exemplification, etc.). *Corpus* serait ainsi un objet instrumental, au même titre que *sens* au pluriel.
De manière non surprenante, la *langue* et les *langues* ne renvoient pas aux mêmes objets : corrélée à la *parole* et au *langage* (substantifs et adjectifs), la *langue* semble discutée dans des textes à dominante historique, dans la mesure où elle est corrélée aux temps de l'imparfait 0.23 (et du plus-que-parfait), eux-mêmes fortement corrélés au passé simple et aux dates 0.2. *A fortiori*, elle s'oppose à de nombreuses caractéristiques des textes scientifiques (symboles, numéraux, marqueurs de structuration des textes, impératif et présent). Il en va fort différemment des *langues*, corrélées aux *langages* et au *sémantique* (substantifs et adjectifs), pour lesquelles on ne repère pas cette dimension historique.
Les corrélats morphosyntaxiques semblent ainsi représenter un critère efficace pour discriminer les candidats et les dimensions textuelles, voire même les pôles génériques (*v.* Poudat 2006) auxquels ils sont associés (textes historiques/exemplifiés/formels, etc.) ; notons toutefois que les

[1] Fréquences absolues des substantifs *vs.* fréquences relatives des variables morphosyntaxiques.

résultats obtenus ne sont pas tous aussi probants, les candidats observés ayant des fréquences inégales. *Langue(s)* et *sens* sont ainsi particulièrement représentés et stabilisés dans le corpus.

5. Tactique

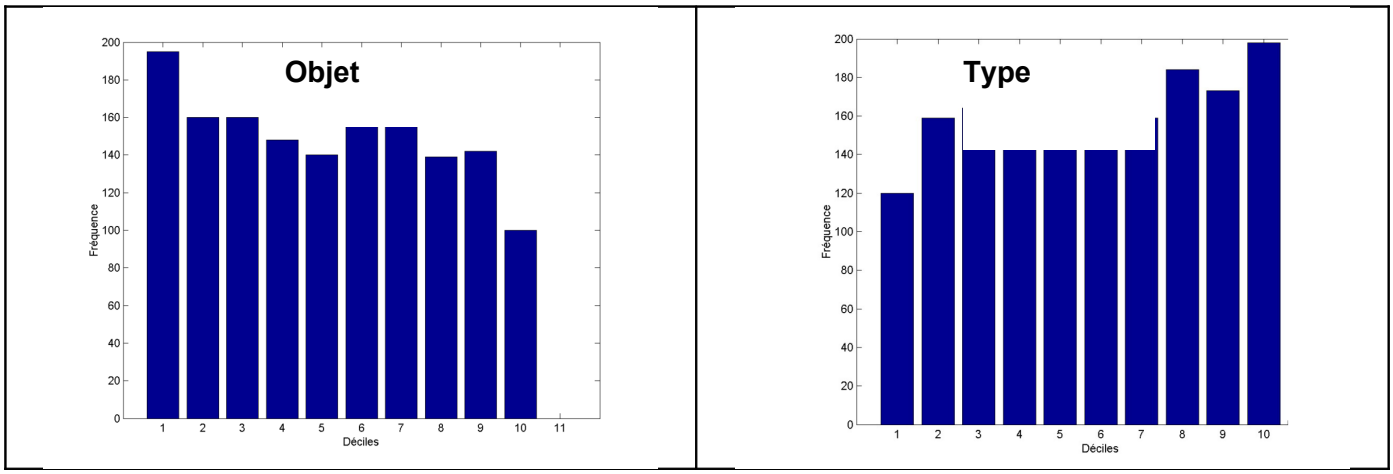
Parmi les composantes sémantiques proposées par F. Rastier, la *tactique*, qui renvoie à la *position* des unités sémantiques, intéresse particulièrement notre entreprise typologique, eu égard à la structure très normée du genre de l'article. On a ainsi apprécié la répartition des concepts dans les textes, fractionnés en dix sections de taille égale au moyen du logiciel CorpusReader développé par S. Loiseau[1], que nous appellerons *déciles de rang d'occurrences de mots par texte*. Chaque *décile* est la fréquence cumulée de l'ensemble des occurrences de l'item à cette position ; ce choix peut paraître singulier, mais Loiseau (2006) a montré que la prise en compte de la moyenne par texte (ou par unité) des occurrences à l'intérieur de chaque dixième ne modifiait pas significativement les résultats obtenus.

Cette représentation des concepts en déciles nous semble finalement plus adaptée que l'observation des candidats au sein de leurs sections textuelles, globalement très inégales[2].

Les profils obtenus sont particulièrement discriminants : certains items, comme *objet* ou *question* sont plus concentrés en début d'article et on observe une décroissance des deux entrées de l'introduction à la fin du texte. On peut légitimement penser qu'ils participent à la problématisation / exposition de la recherche présentée ; en ce sens, *objet* et *question* seraient des concepts instrumentaux plutôt que discutés.

Il en va de même des concepts de corps d'article comme *cas* ou *exemple*, de forme graduelle avec un double mouvement croissance / décroissance et un maximum obtenu en milieu d'article : peu problématisés et peu discutés en fin d'article, ces éléments sont essentiellement corrélés au développement et aux analyses menées. Ce sont donc également des concepts instrumentaux, ou méthodologiques.

Enfin, on observe des items comme *type* ou *construction*, plus denses en fin d'article : leur forme tactique est croissante, et ils semblent ainsi renvoyer aux objectifs généraux de la démarche scientifique linguistique – ici classificatoires et typologiques.



Graphique : Configurations tactiques de OBJET et TYPE (partitionnements en déciles)

Notons d'ailleurs qu'il serait intéressant de déterminer de manière plus précise et plus exhaustive les objets de début et de fin d'article, qui semblent plus méthodologiques que véritablement discutés. On pourrait ainsi les contraster d'une discipline scientifique à l'autre, afin de comparer les démarches et les présupposés méthodologiques adoptés.

Si l'on s'intéresse aux concepts discutés de l'article, la configuration tactique qui a particulièrement retenu notre attention a une forme précisément inverse de celle des concepts de corps d'article :

[1] http://panini.u-paris10.fr/~sloiseau/CR/

[2] Si l'article de revue linguistique contient en moyenne 3,46 sections de niveau 1 par texte, il n'est clairement pas soumis à la structure IMRAD (Introduction, Materials and methods, Results, Analysis, Discussion) ; en d'autres termes, son organisation est sémantiquement hétérogène (une section 2 ne renvoie à aucun objet spécifique).

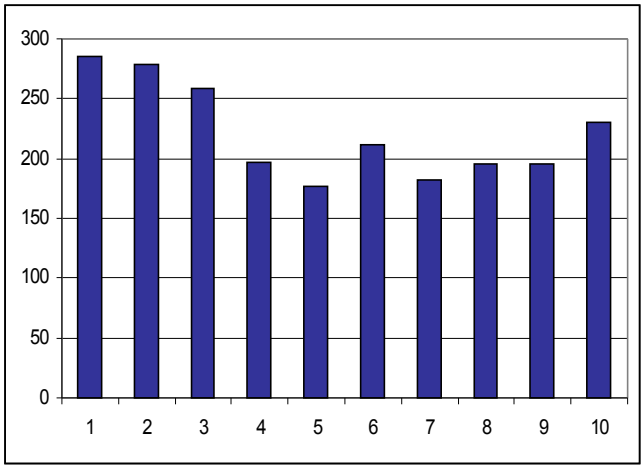


Fig. 1 : Configuration tactique de SENS

Sens est ainsi plus concentré en début et en fin d'article : on note une décroissance régulière du concept dans les trois premiers déciles, qui évoque un passage du général au spécifique (à partir du décile 4). La tendance s'inverse au-delà du décile 6, jusqu'au dixième décile – qui correspond globalement à la conclusion de l'article, où le concept revient brusquement, de manière vraisemblablement rhétorique.
Cette forme tactique incurvée semble spécifique aux concepts débattus, qui seraient ainsi davantage représentés en début et en fin d'article qu'en son corps :

Généralisation/problématisation (maximum atteint) ➔ spécification ➔ retour conclusif

Reconsidérons la liste des 20 substantifs de haute fréquence (et les plus également répartis dans le corpus) mise au jour précédemment : *sens, forme, cas, langue, type, relation, objet, point, discours, contexte, rapport, analyse, verbe, sujet, fait, fonction, question, énoncé, phrase* et *partie*.
Outre *sens*, seuls deux candidats satisfont le critère tactique qui nous intéresse : *discours* et *langue* :

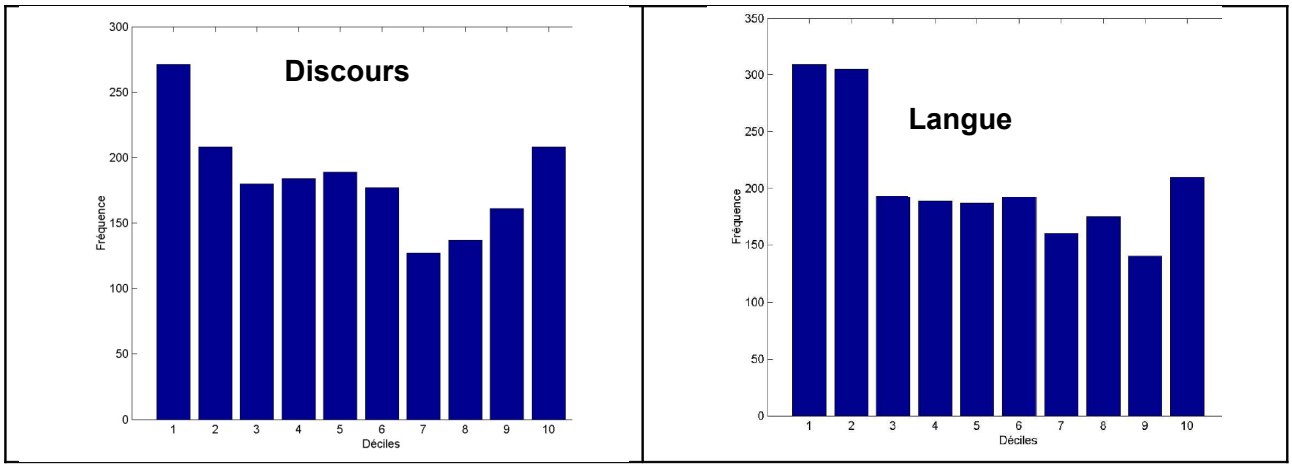


Fig. 2 : Configurations tactiques de DISCOURS et LANGUE

Malgré des différences de répartition des deux concepts dans le développement de l'article, on observe bien un retour sur le concept en fin d'article et un pic de représentation maximal en début d'article : *discours* et *langue* seraient bien des concepts de fond disciplinaire, et il en va d'ailleurs de même pour *langues* et *langage* :

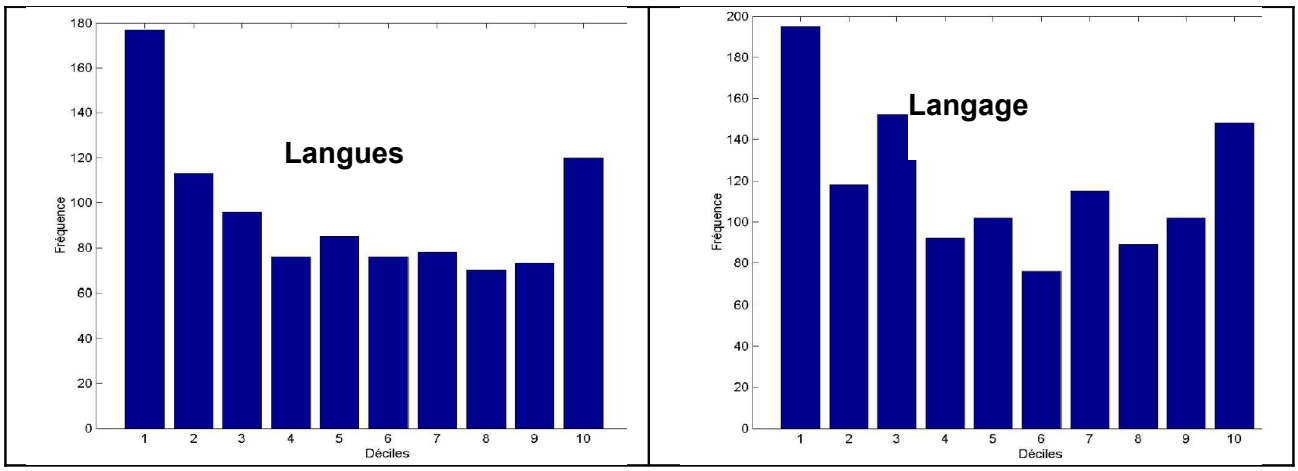


Fig. 3 : Configurations tactiques de LANGUES et LANGAGE

Si cette organisation tactique combinée au critère de fréquence nous semble ainsi permettre de discriminer les concepts discutés des autres, mentionnons une irrégularité, figurée par le substantif de haute fréquence *analyse*[1] qui manifeste également cette configuration.
Les résultats obtenus figurant davantage des formes textuelles génériques que des formes textuelles individuelles – les décomptes étant effectués au niveau du corpus –, nous nous sommes dans un deuxième temps attachée à vérifier l'existence de ces formes et leur statut d'objet discuté dans les textes du corpus.
Les configurations tactiques de chaque texte ont été observées manuellement : les résultats obtenus n'ont qu'une valeur d'approximation, l'identification des formes étant souvent délicate – il serait à terme pertinent de développer un module d'identification automatique des configurations tactiques.
Aux six concepts mis au jour avons-nous adjoint à titre illustratif quatre concepts de configuration tactique globale distincte : la forme plurielle de *discours*, *énonciation*, qui s'avère un concept de fin plutôt que de début d'article, *contexte* et *verbe*, pour lesquels nous n'avons détecté aucune forme tactique particulière.
Les résultats obtenus montrent que *langue* est le concept le plus débattu : 20% des textes qui contiennent l'entrée ont la configuration tactique qui nous intéresse, contrairement à *sens*, qui s'avère comparativement plus employé en début de texte :

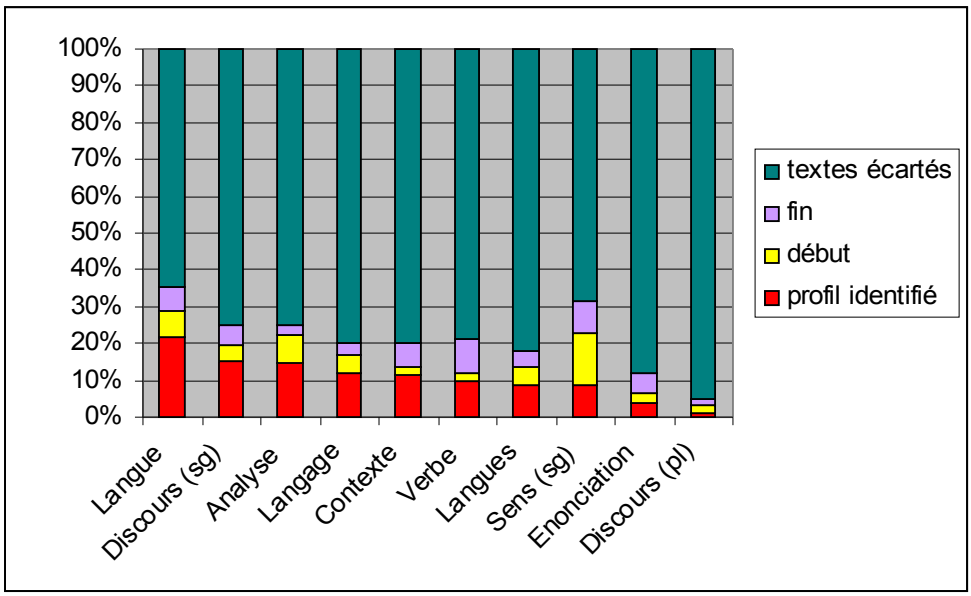


Fig. 4 : Configurations tactiques des formes les plus représentées du corpus

A fortiori, les articles ayant la configuration tactique précédemment observée semblent effectivement discuter la notion : le critère appliqué à l'objet *langage* permet par exemple

[1] Qui est néanmoins aussi un concept hjelmslevien.

d'identifier tous les textes contenant le concept dans leur titre et, de manière plus intéressante, les textes qui discutent la notion sans qu'elle soit nécessairement annoncée. Par exemple, on relève la forme tactique pour le concept de *langue* dans un article de D. Leeman[1] ; le concept est en effet discuté, ce qui n'aurait pas nécessairement été mis au jour sans ce critère.
Ce paramètre tactique, de mise en œuvre plus aisée que d'autres critères, semble ainsi particulièrement discriminant, et pourrait intéresser certaines applications de recherche d'information, en facilitant le repérage et la localisation des thèmes textuels.

6. Conclusion

Le présent article a tenté d'évaluer l'intérêt et la pertinence descriptives de plusieurs critères typologiques. L'entreprise mériterait naturellement d'être approfondie et étendue à l'ensemble des concepts du corpus d'étude, dans la mesure où nous nous sommes particulièrement intéressée aux concepts de haute fréquence *sens* et *langue*.
Si les substantifs de haute fréquence ébauchent le fond disciplinaire de la linguistique, leur seule prise en compte est insuffisante, dans la mesure où leur statut thématique demeure peu défini : s'agit-il de concepts discutés, instrumentaux ou encore des lexicalisations lieutenantes d'une forme ?
L'examen des corrélats morphosyntaxiques et lexicaux des formes permet en partie de répondre à cette interrogation en discriminant différentes acceptions du concept. Si les co-occurrents lexicaux permettent de distinguer, et d'isoler les formants, les corrélats morphosyntaxiques semblent particulièrement discriminants – à condition d'être comme nous au fait de la morphosyntaxe du corpus : ainsi, *sens* au singulier renverrait bien à une lexicalisation privilégiée de concept, tandis que *les sens* seraient des formes peu discutées, voire instrumentales car corrélées aux marqueurs de l'exemple.
La disposition tactique du candidat dans l'article permet de mettre au jour de manière très claire les formes discutées des formes non débattues, ce qui est particulièrement intéressant, dans la mesure où le critère est généralement peu pris en compte : le logiciel CR de S. Loiseau qui permet de telles manipulations de corpus nous semble ainsi particulièrement intéressant.
Si les critères pris en compte se sont avérés présenter un intérêt descriptif et discriminant, la liste est bien entendu ouverte : nous avons par exemple écarté le critère syntaxique. De surcroît, si l'on admet que les concepts les plus fréquents et les plus stabilisés font partie du fonds disciplinaire (Rastier 2005), les concepts émergeants devraient logiquement être plus discutés, donc moins stabilisés ; il en va ainsi par exemple du concept d'*intertexte*, inconnu du glossaire bibliographique Gobert, mais de fréquence suffisante pour se prêter à l'analyse statistique – l'un des numéros thématiques du corpus[2] lui est en effet dédié :

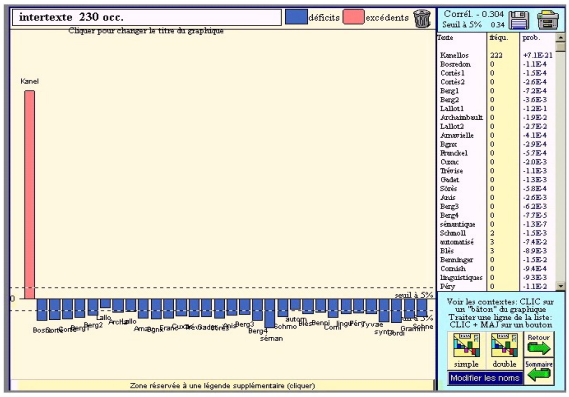


Fig. 5 : Représentation du concept d'intertexte dans les 32 numéros de revue

[1] Leeman, D. Dans un juron, il sauta sur ses pistolets. Aspects de la polysémie de la préposition, in *G.* Bergounioux (éd.), *Approches Sémantiques des prépositions*, RSP vol. 6, Orléans, 1999.
[2] Kanellos (éd.), Sémantique de l'intertexte, *Cahiers de Praxématique*, vol. 33, Montpellier, 1999.

BIBLIOGRAPHIE

BOURION, É. 2001. *L'aide à l'interprétation des textes électronique,* Thèse, Université de Nancy II.
GOBERT, F. 2001. *Glossaire bibliographique des sciences du langage*, Parnormitis.
L'HOMME, M.-C. 2004. Sélection de termes dans un dictionnaire d'informatique : comparaison de corpus et critères lexico-sémantiques, in Actes Euralex 2004, Lorient (France), 6 au 10 juillet 2004, pp. 583-593.
LOISEAU, S. 2003. Philosophical discourse from autonomy to engagement: Deleuze commentator of Spinoza, in K. Fløttum et F. Rastier (éds.), *Academic discourse — Multidisciplinary Approaches*, Oslo, Novus, pp. 36-54.
LOISEAU, S. 2005. Thématique et sémantique conceptuelle d'un concept philosophique, in G. Williams (dir.), *La linguistique de corpus*, Rennes, Presses Universitaires de Rennes.
LOISEAU, S., POUDAT, C. et ABLALI, D. 2006. Exploration contrastive de trois corpus de sciences humaines, in *Actes des 8e JADT*, 19-21 avril 2006, Besançon, pp. 631-642.
POUDAT, C. 2004. Une annotation de corpus dédiée à la caractérisation du genre de l'article scientifique, in *Workshop TCAN Construction du Savoir Scientifique dans la Langue*, Maison Alpes des Sciences Humaines, 20-21 octobre 2004.
POUDAT, C. 2006. *Étude contrastive de l'article scientifique de revue linguistique*, Thèse, Université d'Orléans.
POUDAT, C. et RINCK, F. 2006. Contrastes internes et variations stylistiques du genre de l'article scientifique en linguistique, in *Actes des 8e JADT*, 19-21 avril 2006, Besançon, pp. 785-796.
POUDAT, C. et RINCK, F. 2005. Genres scientifiques et style d'auteur : des variations stylistiques de l'article de revue linguistique, *4e Journées Internationales de la Linguistique de Corpus*, Lorient, 15-17 septembre 2005.
RASTIER, F. 2001. *Arts et sciences du texte*, Paris, PUF.
RASTIER, F. 2003. Semantics of theoretical texts in K. Fløttum et F. Rastier (éds.), *Academic Discourse — Multidisciplinary Approaches*, Oslo, Novus, pp. 15-35.
VALETTE, M. 2003. Conceptualisation and Evolution of Concepts. The example of French Linguist Gustave Guillaume, in K. Fløttum et F. Rastier (éds.), *Academic Discourse — Multidisciplinary Approaches*, Oslo, Novus, pp. 55-74.