Chapitre 3 : Morphosyntaxe du genre de l'article

Toute caractérisation étant différentielle, décrire le genre de l'article ne saurait se faire sans champ(s) de contraste. Une présentation quantitative brute des résultats obtenus serait naturellement peu significative : une proportion de 80% de verbes conjugués au présent dans le genre de l'article est en effet difficile à apprécier sans élément de comparaison. En revanche, si l'on prend en considération que les genres de l'essai et du roman n'en contiennent respectivement que 59,93% et 37,27% (*v.* 3.4.4.1.), les chiffres obtenus sur le genre de l'article deviennent éloquents : la proportion de verbes conjugués au présent dans le genre se fait ainsi *massive*. Caractériser le genre de l'article nécessite donc l'utilisation d'un corpus de référence qui permettra d'apprécier les résultats obtenus sur le corpus observé.

Il va par ailleurs de soi que l'analyse des 145 descripteurs pris en compte n'est pas envisageable isolément. Chaque descripteur doit d'abord être appréhendé par rapport à sa (ses) classe(s) linguistique(s) d'appartenance : l'analyse du présent sera ainsi fondée sur les proportions obtenues par le temps verbal relativement à l'ensemble des temps conjugués, voire des formes verbales, selon les objectifs descriptifs privilégiés. On rendra compte des caractéristiques et de l'organisation des différents systèmes de description à partir des statistiques descriptives calculées sur les sorties (révisées) de l'étiqueteur TnT.

Les descripteurs seront dans un deuxième temps examinés au regard de leurs corrélations textuelles, qui permettront de mettre en évidence les attractions et les tensions des catégories au sein du genre. Si l'analyse factorielle est elle-même fondée sur ces systèmes de corrélations, il nous a semblé pertinent d'intégrer l'étude des corrélations à celle des systèmes de descripteurs, tant pour préciser la description que pour faciliter l'interprétation des axes de la factorielle menée dans une étape ultérieure.

Etant donné les particularités attendues des exemples des textes, nous avons enfin procédé à une mise en contraste systématique des résultats statistiques obtenus sur les textes entiers (FT pour *Full Text*[1]) et dont les exemples ont été extraits[2] (TE pour *Textes extraits*), au sein de l'analyse de chaque système de description. Ce choix méthodologique présente un double intérêt : il permet d'une part de préciser les caractéristiques du genre en spécifiant celles du corps seul de l'article et d'autre part d'amorcer celles des exemples, qui seront détaillées dans un chapitre ultérieur (*v.* chapitre 5).

Les régularités statistiques générales de l'article établies, on procèdera à une analyse factorielle, et plus précisément à une Analyse en Composantes Principales (ACP), qui permettra de faire émerger les dimensions principales du genre : on s'intéressera ainsi aux principales modalités de variation du genre de l'article de revue linguistique.

Cette ACP sera complétée d'une Classification Ascendante Hiérarchique (CAH) qui nous permettra de mettre au jour les proximités et les écarts entre les textes du corpus, de même que leurs positionnements sur les axes factoriels obtenus.

N.B. : Nous avons bien entendu conscience des limites d'une démarche descriptive essentiellement quantitative : l'ensemble des résultats statistiques obtenus devraient être

---

[1] Texte entier hors bibliographie bien entendu.

[2] Précisons que nous n'avons globalement extrait que les exemples *décrochés* des textes, c'est-à-dire délimités par un saut de ligne – et éventuellement une indentation – dans le corps de l'article.

affinés, voire dans certains cas, validés en contexte et en corpus. Cette opération n'est néanmoins pas envisageable pour chaque descripteur, d'une part parce que les modalités d'analyse d'un marqueur varient considérablement en fonction des objectifs et des présupposés théoriques de l'étude, et d'autre part parce que notre objectif est autre : nous visons en effet principalement à décrire les lieux de stabilité du genre, et à en dresser un *profil général*, avec toutes les limites que la démarche implique dans son fondement. En revanche, nous nous sommes efforcée de ne pas sur-interpréter les phénomènes statistiques obtenus qui nécessiteraient un approfondissement en corpus ; certains phénomènes ont ainsi été examinés de manière plus qualitative, ce qui nous a permis d'affiner la description du genre tout en présentant différentes méthodologies possibles d'analyses en corpus.

3.1. Méthodologie générale

Après avoir présenté le corpus de référence auquel sera contrasté le corpus observé, nous exposerons l'ensemble des tests statistiques exploités dans le présent chapitre.

3.1.1. Corpus de référence

Force est d'abord de constater que la communauté linguistique française accuse un retard sérieux sur les anglo-saxons en matière de description des usages linguistiques. Il y a ainsi peu de corpus et de statistiques disponibles, ce qui est particulièrement problématique pour notre entreprise.

Nous avons toutefois pu récupérer un ensemble de données statistiques obtenues à partir des sorties de l'analyseur Cordial® sur trois grands corpus de genre et de type de discours différents : un corpus de romans sérieux[3], un corpus de textes juridiques et un corpus d'essais caractérisés à partir de 180 descripteurs[4].

Malgré la nature hétérogène de la catégorie « textes juridiques » et la pertinence relative de la comparaison, la mise en contraste de notre corpus avec ces trois sous-ensembles nous semble permettre une première caractérisation du genre, qu'il serait naturellement pertinent et intéressant de valider et de préciser avec d'autres grands corpus dans les années qui viennent.

Le corpus d'articles, d'ailleurs augmenté de 23 textes, a ainsi été soumis à Cordial® et les statistiques obtenues ont servi de fondement à une première caractérisation du genre (Poudat, 2003). Soulignons bien entendu que le jeu d'étiquettes du logiciel est moins adapté au discours scientifique que celui que nous avons élaboré, ce qui limite parfois la comparaison – notamment en ce qui concerne les marques de formalisation, certains numéraux, les *il* impersonnels, etc.

3.1.2. Statistiques descriptives

3.1.2.1. Médiane

La valeur médiane d'un ensemble de données renvoie à l'observation qui se situe au point milieu de cette liste ordonnée. 50% des valeurs sont donc supérieures à la médiane, tandis que 50% lui sont inférieures.

---

[3] Roman de début du XX^{e} incluant plusieurs sous-genres communiquant entre eux (romans de formation, romans par lettres, romans psychologiques, *etc.*) à l'exclusion du roman policier, selon la typologie proposée par Rastier (2001).

[4] Je remercie d'ailleurs vivement D. Malrieu de m'avoir communiqué ces données.

3.1.2.2. Variance et écart-type

La variance et l'écart-type sont des indicateurs de dispersion communément utilisés en analyse de données.

La variance est la moyenne des carrés des écarts à la moyenne. Elle mesure donc la dispersion des valeurs autour de la moyenne. On dit que la variance traduit la notion d'incertitude : plus la variance est faible, moins le résultat de l'expérience est incertain.

Pour les nombres 1, 2 et 3, par exemple, la moyenne est 2 et la variance, 0,667.

$[(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2] \div 3 = 0{,}667$

[écart au carré moyen] ÷ nombre d'observations = variance

Comme le calcul de la variance se fait à partir des carrés des écarts, les unités de mesure ne sont pas les mêmes que celles des observations originales. Par exemple, les longueurs mesurées en mètres (m) ont une variance mesurée en mètres carrés (m^2).

L'écart-type a l'avantage de s'exprimer dans la même unité que le caractère :

Écart-type (S) = Racine carrée de la variance

L'écart-type est la mesure de dispersion la plus couramment utilisée en statistique lorsqu'on emploie la moyenne pour calculer une tendance centrale. Il mesure donc la dispersion autour de la moyenne. En raison de ses liens étroits avec la moyenne, l'écart-type peut être grandement influencé si cette dernière donne une mauvaise mesure de tendance centrale.

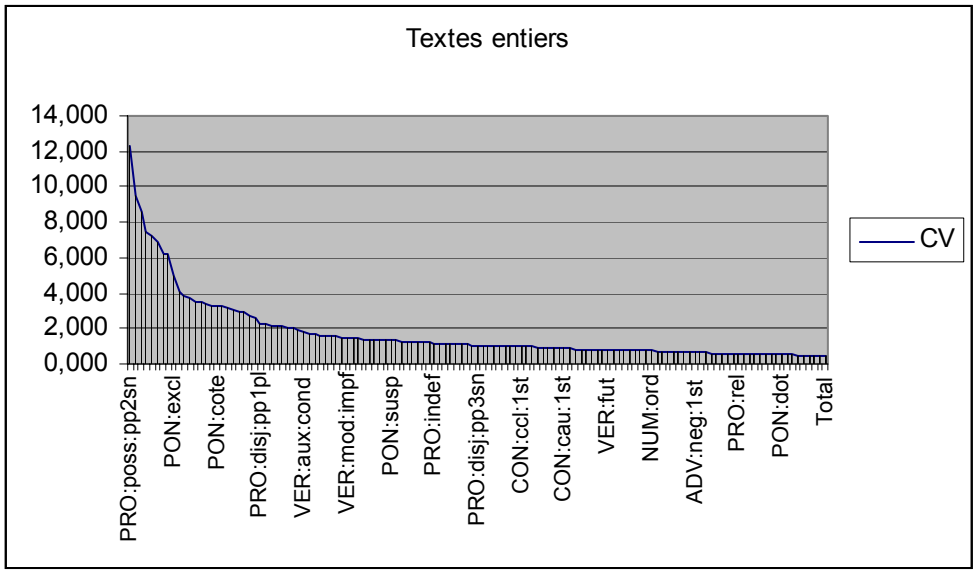
3.1.2.3. Coefficient de variation

Si des mesures statistiques comme la moyenne, les valeurs minimale et maximale et l'écart-type apportent de précieuses indications quant à la caractérisation de notre objet, elles ne permettent pas d'évaluer précisément la variabilité d'un ensemble de données.

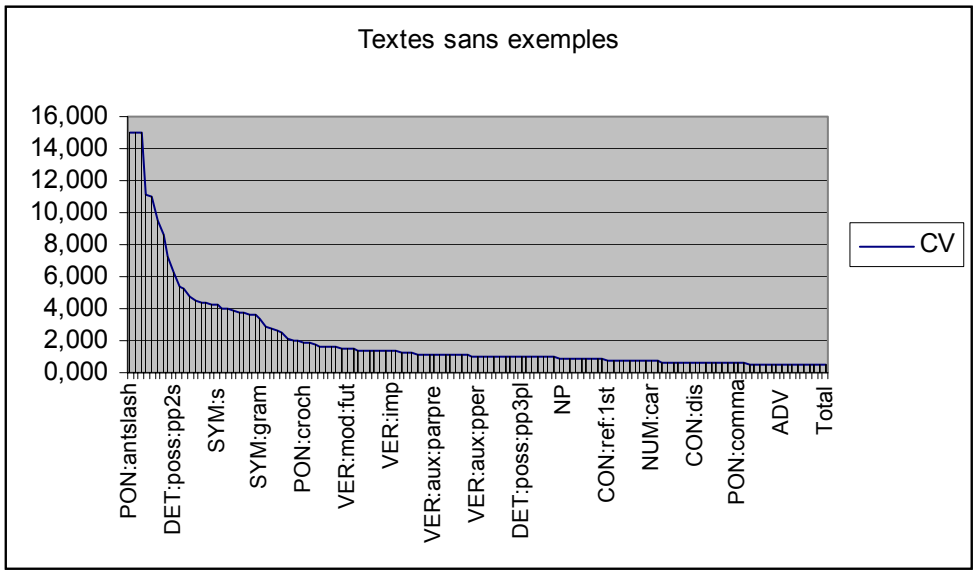
Nous avons donc utilisé le coefficient de variation (CV), souvent employé à cet effet : le CV est l'écart-type divisé par la moyenne, souvent exprimé comme un pourcentage à la moyenne :

$$CV = \frac{\sigma}{\mu}(100\%)$$

Comme l'illustrent les graphiques 17 et 18, plus des deux tiers des variables observées ont un CV inférieur à 2 (soit 200%), ce qui nous semble déjà constituer un premier « noyau dur » du genre.



Graphique : Répartition des variables selon leur coefficient de variation (FT)



Graphique : Répartition des variables selon leur coefficient de variation (TE)

Par conséquent, les éléments ayant un CV supérieur à 2 seront par la suite considérés comme plus aléatoires, et de fait moins caractéristiques du genre.

3.1.3. Analyse des corrélations

Une corrélation positive ou négative entre deux variables est considérée comme *significative* si elle est supérieure ou égale à :

$$\pm 2 / \sqrt{(n-1)}$$

n = nombre d'individus

On obtient donc un seuil de 0.13, qu'il est plus raisonnable d'arrondir à 0.2 en raison des comparaisons multiples[5]. On notera le petit seuil obtenu, qui reflète la très grande homogénéité du corpus étudié.

3.1.4. Tactique des variables

Parmi les composantes sémantiques proposées par F. Rastier, la *tactique*, qui renvoie à la *position* des unités sémantiques, intéresse particulièrement notre entreprise descriptive, 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 CR développé par S. Loiseau[6], que nous appellerons *déciles de rang d'occurrences de mots par texte*.

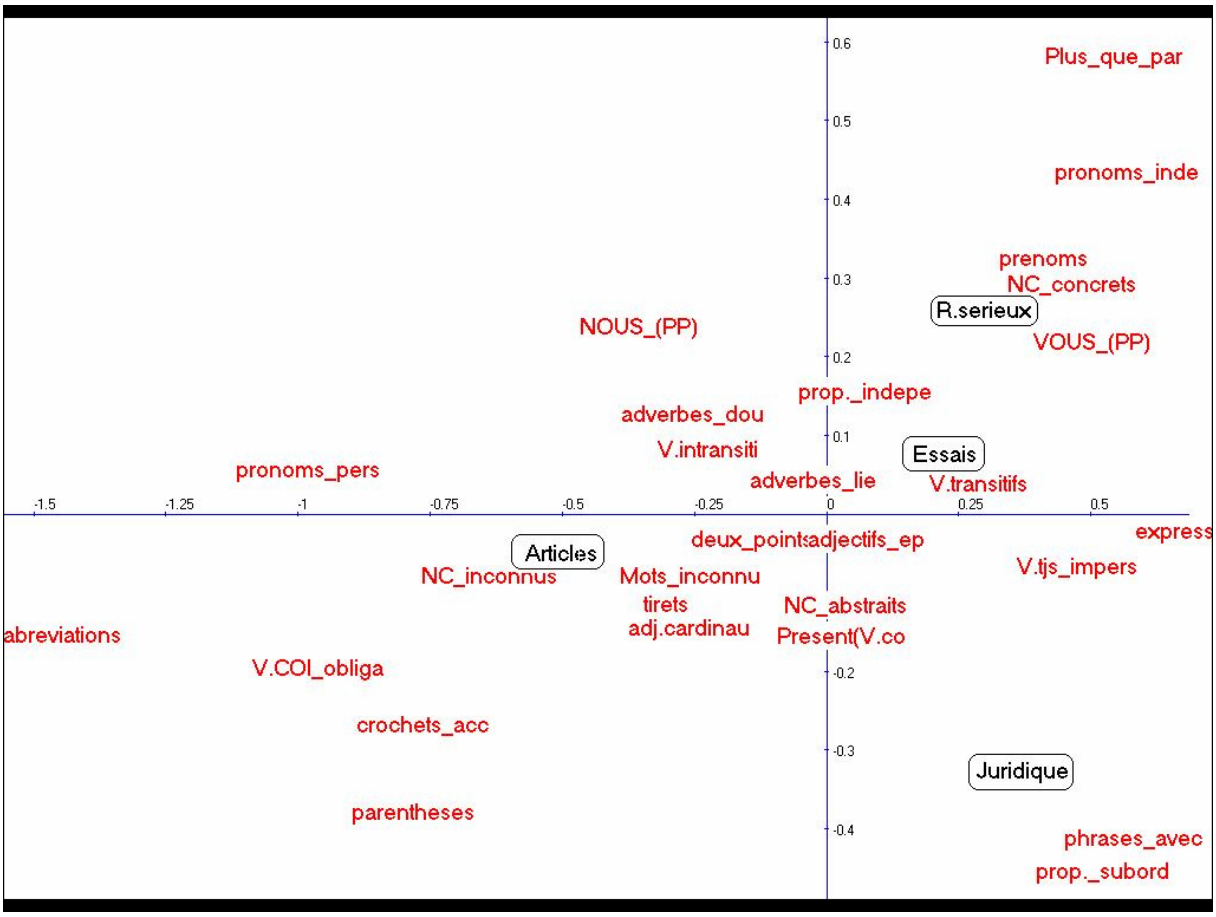
Chaque *décile* est la fréquence cumulée de l'ensemble des occurrences de l'item à cette position ; ce choix peut paraître singulier, mais Loiseau (2006) a montré que la prise en compte de la moyenne par texte (ou par unité) des occurrences à l'intérieur de chaque dixième ne modifiait pas significativement les résultats obtenus.

3.2. Caractéristiques générales du genre de l'article

Observons les proximités et les oppositions du genre de l'article avec les trois autres corpus considérés :

[5] Je remercie L. Lebart de m'avoir communiqué cette précieuse information.

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



Graphique : Caractéristiques du genre de l'article et des trois autres genres et discours observés (Analyse factorielle des correspondances)[7]

Si l'article possède des propriétés propres, il semble plus proche des textes juridiques que des essais et des romans sérieux. Considérons les caractéristiques propres et partagées qu'il semble posséder.

- **Caractéristiques communes de l'article et du roman sérieux** : les deux genres se rapprochent par un nombre plus important de mots par paragraphe, de noms propres 'géographiques' et de verbes conjugués à la première personne du singulier. Ils semblent contenir moins de verbes conjugués au passé composé que les deux autres ensembles
- **Caractéristiques communes de l'article et de l'essai :** les deux genres contiennent davantage de pronoms possessifs de seconde personne du singulier et de première et de troisième personne du pluriel que les textes juridiques et les romans sérieux. De surcroît, ils contiennent des phrases plus longues et un nombre d'adverbes par proposition plus important. Enfin, on relève un déficit de points de suspension.
- **Caractéristiques communes de l'article et des textes juridiques :** soulignons d'abord que les deux ensembles se rapprochent davantage par des déficits que par des excédents. Ainsi, ils contiennent moins de points, de virgules, de points d'interrogation et d'exclamation que les deux autres genres. De manière non surprenante, ils se caractérisent également par un déficit de verbes conjugués au passé simple, à l'imparfait et à l'impératif, et de verbes employés à la seconde personne du pluriel. On relève également moins de négations, un

[7] Je remercie vivement L. Lebart de cette analyse et de ce graphique.

nombre de phrases par paragraphe plus restreint et un déficit de propositions contenant un complément circonstanciel de temps.

Les deux formations emploient globalement plus de futurs que les deux autres genres, et davantage de 'noms propres inconnus'. Elles contiennent des propositions comprenant plus de mots en général, et globalement plus d'adjectifs. Enfin, on relève un nombre plus important de verbes conjugués à la troisième personne du singulier.

- **Caractéristiques propres au genre de l'article :** outre les éléments préalablement mentionnés, le genre de l'article comporte des caractéristiques spécifiques qui le distinguent des trois autres genres ; le tableau ci-dessous rassemble les différences les plus notables observées – en gris sont représentés les éléments sur-représentés dans le genre de l'article :

	Articles	Essais	Juridique	R. sérieux
% deux points	6,92	4,15	3,53	3,11
% parenthèses	19,22	2,03	5,85	0,50
% tirets	1,71	0,67	0,87	0,65
% crochets et accolades	1,11	0,18	0,27	0,06
% Mots inconnus	3,00	1,15	1,31	1,11
% NC inconnus	2,06	0,63	0,38	0,36
% NC abstraits	64,76	48,34	54,89	36,82
%NC concrets	9,29	14,78	14,15	25,15
% abréviations (NP)	6,14	0,03	0,01	0,07
% prénoms	7,37	12,67	14,70	26,49
% adj. cardinaux (dét.)	4,43	1,52	2,25	1,57
% pronoms indéfinis	4,90	31,39	12,41	31,10
% pronoms personnels	57,71	6,36	0,72	9,55
% NOUS (PP)	8,68	5,53	0,76	3,58
% VOUS (PP)	1,45	3,83	3,78	5,73
% adjectifs épithètes	96,58	75,99	82,11	73,02
% adverbes de lieu	11,71	6,93	6,48	7,88
% adverbes de doute	6,71	3,26	1,79	3,33
% Présent (V. conjugués)	80,43	59,93	66,88	37,72
% Plus-que-parfait (idem)	0,30	1,64	1,15	3,99
% V. intransitifs	7,71	3,68	2,91	3,92
% V. transitifs directs	54,25	79,96	77,66	77,77
% V. COI obligatoire	13,19	1,12	1,86	0,92
% V. tjs impersonnels	4,57	11,18	13,13	9,90
% prop. indépendantes	86,25	75,17	51,31	80,38
% prop. subordonnées	9,74	19,06	45,83	14,83
% phrases avec 1 subordonnée	7,66	22,62	53,26	18,18
% expressions usuelles (mots)	0,15	2,17	2,83	2,40

Tableau : Première caractérisation du genre de l'article[8]

Le genre de l'article se démarque d'abord nettement au niveau des ponctuations : on relève un emploi très important des parenthèses (qui représentent près de 20% de l'ensemble des ponctuations), des deux points, tirets, crochets et accolades. L'usage massif de parenthèses dans l'article nous semble lié d'une part à la présence importante de références dans les corps d'articles, et d'autre part aux formalisations et digressions éventuelles, ce qui serait à valider en contexte, de même que la présence des deux points, qui pourrait

[8] Les pourcentages sont calculés à partir des classes construites par Cordial. Ainsi, le pourcentage de parenthèses est calculé en fonction du nombre total de ponctuations, et celui de noms communs abstraits par rapport à l'ensemble des noms. Lorsque la catégorie de référence n'est pas intuitive, elle est mentionnée entre parenthèses.

correspondre à une dimension plus démonstrative des textes scientifiques. L'article est en outre bien connu pour ses formalisations, qui requièrent l'utilisation d'une sémiotique textuelle particulière impliquant l'usage de ponctuations aussi spécifiques que les crochets et les accolades. La présence importante de tirets est difficile à interpréter, dans la mesure où le tiret est ici ambigu : il renvoie tant aux indices de liste qu'au symbole mathématique négatif ou de soustraction, éléments toutefois spécifiques au discours scientifique.

On relève également un emploi considérable du présent, qui représente 80% des temps conjugués, et un déficit de verbes conjugués au plus-que-parfait, lié nous semble-t-il à la sous représentation déjà relevée d'imparfaits, et probablement de passé simple, temps du récit en principe peu caractéristiques du genre de l'article.

Il semble par ailleurs que l'article emploie davantage de pronoms personnels *nous* et *vous*. *Nous* étant la personne privilégiée par l'auteur pour se manifester, ce phénomène est peu étonnant : la présence de *vous* est par contre plus inattendue, et pourrait être corrélée aux exemples des textes, qui relèvent de genres autres.

L'article semble de surcroît comporter des spécificités syntaxiques particulières : on y relève deux fois moins de propositions subordonnées, et trois fois moins de phrases qui en contiennent au moins une. Le genre de l'article privilégie en outre nettement les verbes intransitifs et à COI obligatoire aux verbes transitifs directs, phénomène que nous tenterons d'élucider *infra*.

Enfin, et avec toutes les réserves que les listes lexicales peu transparentes de Cordial® impliquent, il semble que le corpus contienne davantage de termes « abstraits » et d'adverbes de « doute » et de lieu.

3.3. Eléments marginaux du genre

Afin d'évaluer les éléments les plus marginaux des textes, considérons les éléments absents de la majorité des textes :

	Textes entiers		**Textes sans exemples**	
	Nb textes de valeur 0	**% valeur 0**	**Nb textes de valeur 0**	**% valeur 0**
Pronoms possessifs de 2nde pers. du sg.	222	99,10	223	99,55
Pronoms possessifs de 2nde pers. du pl.	221	98,66	223	99,55
Pronoms possessifs de 3e pers. du pl.	220	98,21	221	98,66
Pronoms possessifs de 1ère pers. du sg.	220	98,21	221	98,66
Modaux au passé simple	215	95,98	222	99,11
Antislashs	215	95,98	223	99,55
Pronoms possessifs de 1ère pers. du pl.	206	91,96	209	93,30
Accolades	201	89,73	208	92,86
Pronoms possessifs de 3e personne du sg.	198	88,39	210	93,75
Disjoints VOUS	195	87,05	214	95,54
Auxiliaires au passé simple	191	85,26	205	91,52
Subjonctif imparfait	185	82,58	195	87,05

Disjoints TOI	183	81,69	204	91,07
Dét. possessifs de 2nde pers. du pl.	180	80,35	204	91,07
Dét. possessifs de 2nde pers. du sg.	173	77,23	201	89,73
Clitiques VOUS	172	76,78	202	90,18
Clitiques TE	172	76,78	203	90,63
Marqueurs grammaticaux	172	76,78	181	80,80
Interjections	158	70,53	187	83,48
Disjoints NOUS	154	68,70	168	75,00
Disjoints MOI	144	64,28	179	79,91
PP VOUS	144	64,28	184	82,14
PP TU	142	63,39	182	81,25
Cotes	141	62,94	150	66,96
Passé simple	131	58,48	172	76,79
Symboles linguistiques grammaticaux	124	55,35	129	57,59
Points d'exclamation	115	51,33	156	69,64
Symboles linguistiques			148	66,07
Dét. possessifs de 1ère personne du sg.			123	54,91
Modaux à l'imparfait			121	54,02
Clitiques ME			114	50,89

Tableau : Eléments marginaux du genre de l'article

Les éléments les plus marginaux du genre de l'article sont d'abord les pronoms possessifs (le *tien,* le *nôtre*, le *mien*, etc.) : ces éléments sont absents de plus de 90% des textes (voire de 99% après extraction des exemples). Les relations marquées de possession personnelle semblent ainsi absentes du genre de l'article, ce qui n'est pas surprenant en soi.

On remarquera ensuite que les marqueurs de seconde personne du singulier et du pluriel sont très faiblement représentées : contrairement à d'autres genres comme la lettre ou parfois le roman, l'article, même *reader-friendly,* ne contient en principe pas d'adresse au lecteur. Ces marqueurs semblent d'ailleurs essentiellement relever de l'exemple, dans la mesure où leur absence est nettement plus prononcée dans les textes extraits : ainsi, les pronoms personnels *vous* et *tu*, absents des 6/10e des textes entiers, sont absents des 8/10e des textes après extraction.

L'auteur n'étant pas supposé se mettre en valeur de manière explicite, les pronoms disjoints – ou toniques – de première personne, qui renforcent et valorisent la première personne, sont bannis de plus des deux tiers des textes. Après extraction des exemples, les formes de première personne se marginalisent d'ailleurs de manière significative : ainsi, le déterminant possessif et le clitique ME s'avèrent absents de 54.91% et de la moitié des textes. On notera que le pronom *MOI* est très employé dans les exemples des textes : absent de 64,28% des textes entiers, il est significativement absent de 79,91% des textes sans exemples.

Les temps les plus marginaux des textes sont le subjonctif imparfait (absent de 82.58% FT/87,05% TE) et le passé simple (58,48% FT/76,79% TE), qui semble très présent dans les exemples du texte. Temps révolu du récit, le passé simple est caractéristique de nombreux genres (littéraires, journalistiques, etc.) : il conviendra ultérieurement de déterminer si les exemples du texte contiennent des fragments attestés de genres narratifs, ou si l'exemple

construit est facilement conjugué au passé simple. De manière générale, le passé simple s'oppose par son caractère révolu au discours universel et atemporel des textes scientifiques. Peu, voire non modalisé (abs. 96% FT/99,11 TE)[9], sa représentation dans plus de la moitié des textes du corpus semble indiquer la présence de fragments narratifs, caractéristiques des textes plus historiques, ou qui contiennent une dimension épistémologico-historique. Enfin, soulignons que si le passé simple est lui-même employé dans les 2/5^{e} du corpus, ce sont ses formes modales et composées (passé antérieur) qui sont les plus absentes (respectivement absentes de 96% et 85.26% des textes).

Parmi les ponctuations les plus absentes du genre observé, mentionnons d'abord la faible représentation des antislashs et des accolades (abs. des 9/10^{e} des textes). L'antislash n'est pas caractéristique du corps de l'article : après extraction des exemples, on n'en relève plus que dans un texte du corpus, ce qui est extrêmement marginal. Si nous avions d'emblée écarté les textes trop formalisés pour des raisons évidentes de facilités de traitement, cette sous-représentation demeure remarquable, dans la mesure où d'autres ponctuations candidates à l'expression de formalismes comme le slash ou le crochet sont largement plus représentées.

Le point d'exclamation est également marginal, mais de manière moindre. Il semble d'emblée peu représentatif du genre étudié et du discours scientifique, dans la mesure où il est généralement tenu pour subjectif (expression de sentiments, etc.) : il nuirait ainsi à l'objectivité et à la scientificité du texte, et devrait par conséquent être sinon proscrit, du moins évité. Après extraction des exemples, qui ne sont pas soumis à aux mêmes régulations, on n'en relève d'ailleurs que dans trois textes sur dix.

Concurrentes directes des guillemets en matière de (non) prise en charge textuelle, les guillemets simples sont enfin absents de 62.94% des textes : la préférence semble ainsi aller aux guillemets, ce qui n'est pas surprenant.

Les quatre derniers éléments non abordés du tableau 21 (marqueurs grammaticaux[10], interjections, symboles linguistiques et symboles linguistiques grammaticaux) sont naturellement non caractéristiques du genre : théoriquement absentes des corps d'articles, les interjections (abs. 70.53% FT/83.48% TE) relèvent en principe d'articles travaillant sur l'oral ou utilisant du matériel oral (ou mimant l'oral comme certains romans). De même, les symboles linguistiques grammaticaux et les symboles linguistiques sont généralement employés dans des articles de syntaxe (*, ?, NP, SN, SV, etc.) ou de morphologie/morphosyntaxe (le préfixe –RE, etc.). Si ces variables ont un rôle d'indicateur domanial non négligeable, leur sous-représentation est plutôt encourageante quant à la représentativité du corpus par rapport à l'ensemble du domaine linguistique.

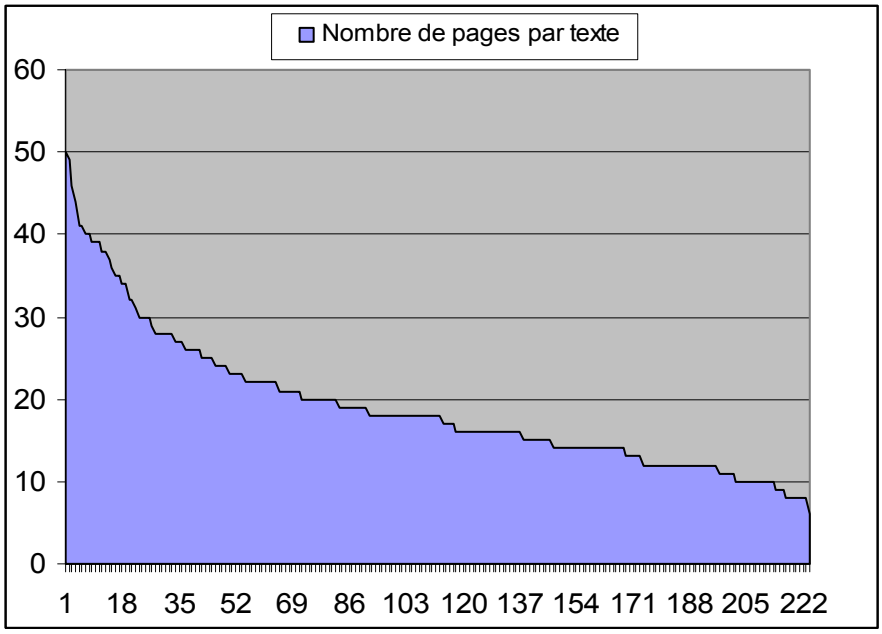
3.4. Eléments de description du genre

3.4.1. De la longueur des textes

L'article scientifique de linguistique aurait ainsi une longueur moyenne de 19.10 pages, avec un écart-type de 8.42 pages, soit une moyenne de 7333.04 tokens (écart type de 3144.5).

[9] On notera qu'à l'instar du passé simple, l'imparfait est également peu modalisé (abs. 54% TE), ce qui demeure à élucider.

[10] Catégories hybrides fonctionnelles, destinées à pallier certains problèmes d'étiquetage.



Graphique : Nombre de pages par texte

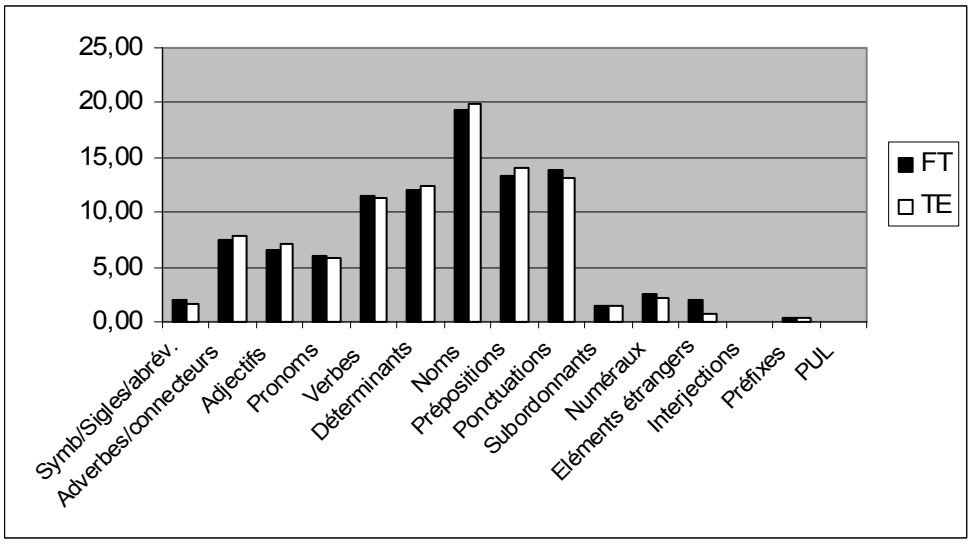
Plus un texte est long, plus il contient de virgules (+0.18), de relateurs prépositionnels (+0.17) et propositionnels (+0.15), et plus il est court, plus il contient de numéraux (-0.18) – et plus particulièrement de chiffres cardinaux (-0.17) – et de slashs : les textes reportant des résultats chiffrés et plus formalisés, seraient plus courts que les articles spéculatifs ou historiques.

Cette tendance se confirme nettement si l'on extrait les exemples des textes : les corrélations obtenues sont plus fortes et on observe une corrélation des textes plus longs au temps du subjonctif imparfait (+0.17), indice d'un style rédactionnel plus soutenu, et aux dates (+0.14). Les marques de troisième personne du singulier (PP3 sg anaphorique, déterminants possessifs et pronoms disjoints), caractéristiques des textes plus conceptuels qui décrivent souvent un objet (ou une personne) et son fonctionnement, leur sont également corrélés.

Enfin, les textes plus courts sont plus fortement corrélés aux numéraux (+0.22) et aux slashs, mais également aux deux points (+0.19), à valeur globalement démonstrative dans les textes scientifiques.

3.4.2. De la répartition des classes linguistiques au sein du genre

Malgré les difficultés d'interprétation qu'il pose étant donné que nous ne disposons pas de champ de comparaison pour ces données, le graphique suivant présente les répartitions des 15 classes descriptives élaborées chapitre 2.

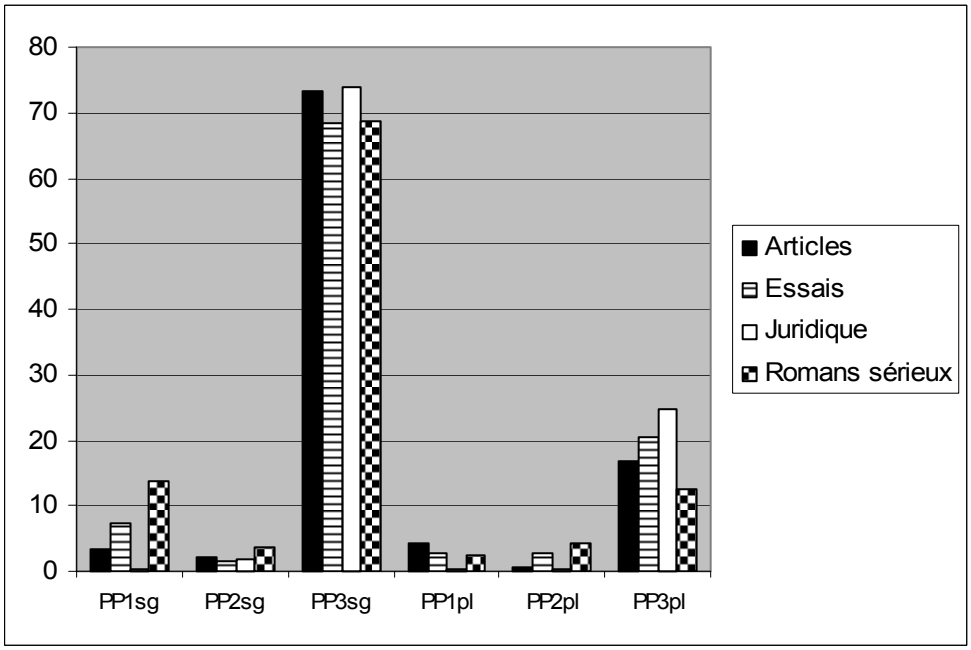


Graphique : Répartitions des classes linguistiques au sein du genre (en %)

3.4.3. Des personnes

3.4.3.1. Genre de l'article vs. genres autres

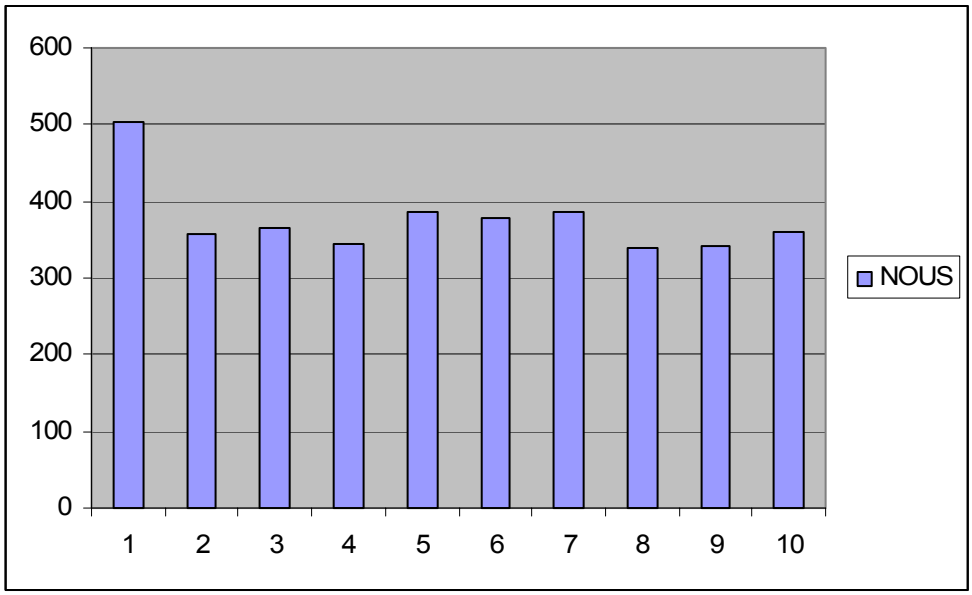
Il est d'abord nécessaire de caractériser le genre de l'article par rapport à d'autres genres et types de discours. Comme l'illustre le graphique suivant, qui rassemble les résultats obtenus à partir des sorties de l'analyseur Cordial®, les différences notables observées entre les quatre corpus concernent les pronoms de première et de seconde personne. Si l'on relève un écart important entre les pronoms de troisième personne du singulier et du pluriel, il ne semble pas spécifique au genre de l'article.



Graphique : Répartition en pourcentage des verbes conjugués par personne et par genre

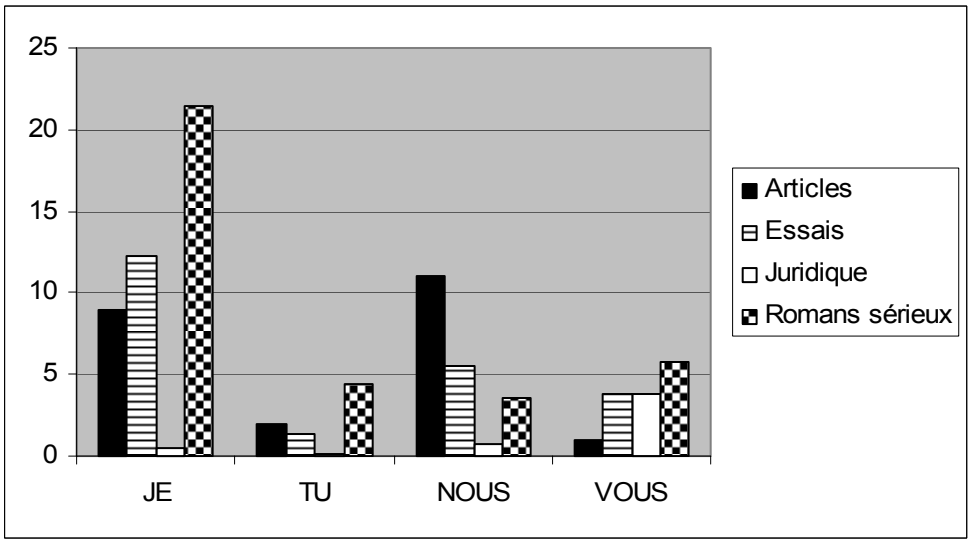
L'article se démarquerait ainsi par un nombre important de *nous* et un déficit de *vous* ; le pronom personnel de première personne du pluriel est en effet particulièrement usité dans les textes scientifiques, de manière tant exclusive (pour renvoyer à l'auteur) qu'inclusive (pour

inclure le lecteur ou une communauté de rattachement particulière). On observe d'ailleurs un emploi particulièrement intensif du pronom en introduction d'article, et il est vraisemblable, étant donné la fonction de la section introductive (détaillée chapitre 4) que le *nous* employé soit inclusif avec le lecteur, ce qui lui conférerait une fonction de guide dans l'annonce du développement à venir :



Graphique : Configuration tactique de NOUS (fréquences absolues / déciles)

L'article ne contenant normalement pas d'adresse au lecteur, on remarquera que *vous*, en principe marginal dans le corps de l'article, semble également peu employé dans les exemples des textes étant donné sa faible représentation. Si *je* est quasi absent des textes juridiques, et massivement représenté dans le roman sérieux, l'article et l'essai en contiennent une proportion comparable, bien que le pronom soit plus représenté dans le dernier genre.



Graphique : Répartition des pronoms personnels de première et de seconde personne par genre

Si l'auteur s'investit significativement plus dans le genre de l'essai que dans celui de l'article, plus bureaucratique et moins personnel, il serait cependant inapproprié d'interpréter les résultats obtenus sous cet éclairage, dans la mesure où les valeurs endossées par les

pronoms varient au sein de systèmes diversement équilibrés selon les discours, genres ou textes. Une étude comparative d'essais philosophiques et d'articles de linguistique (Poudat & Loiseau, 2002) a ainsi montré que *je* ne désignait jamais directement l'auteur dans le genre de l'essai. La première personne du singulier a ainsi toujours une acception universalisante reposant sur l'indétermination référentielle de la première personne (Benveniste). Cependant, si le *je* désigne « tout sujet », l'auteur comme le lecteur est inclus dans cette universalisation. Ainsi l'auteur est à la fois narrateur et « acteur » de ce qu'il décrit. De plus le *je* universel permet de représenter un univers commun au lecteur et à l'auteur, où l'adhésion du lecteur est facilitée et sollicitée. La faible valeur autoriale du *je* est attestée à travers les lemmes corrélés, qui concernent très peu le registre de la prise en charge discursive, mais plutôt des actions concrètes qui ne sont pas celle de l'auteur dans son texte, comme *dissoudre, refouler, mourir, voir*… Quand ces verbes ont un contenu gnoséologique, c'est toujours celui de l'expérience naïve, affectée d'une valeur négative d'un point de vue philosophique : *percevoir, imaginer sentir souvenir*… L'auteur se manifesterait donc indirectement dans le genre de l'essai philosophique, tandis que le *je* du genre de l'article de revue linguistique semble indiquer une implication forte de l'auteur dans ses propos, et aurait ainsi une valeur d'auto-référence très comparable au *nous* de l'essai :

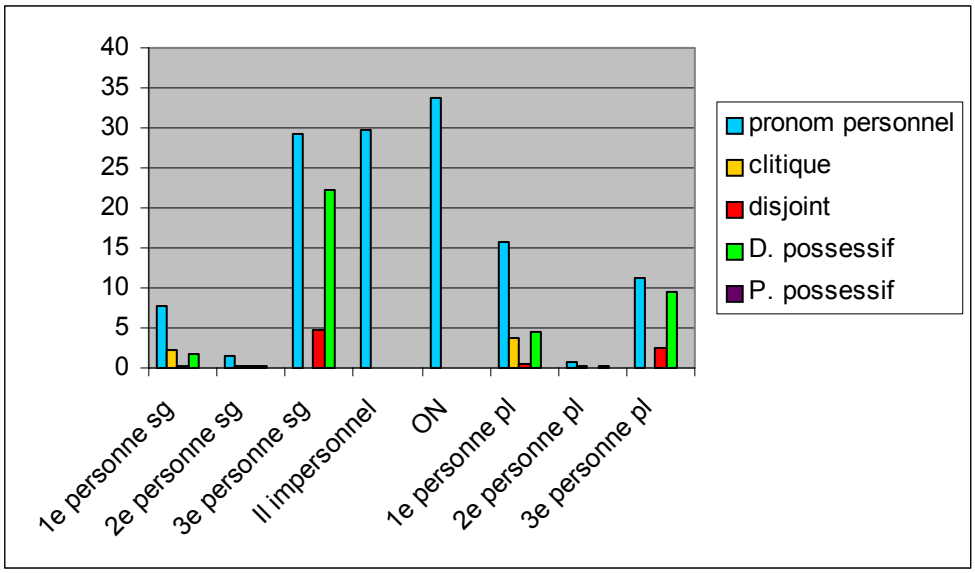
> Nous ne suggérons aucune ressemblance entre le Dionysos de Nietzsche et le Dieu de Kierkegaard. (Deleuze, *Différence et répétition*)
>
> Finalement, cette quête d'indices de contextualisation conduit, au plan de la méthode, à rassembler en premier lieu **ce que j'appelle** des sous-corpus d'énoncés (Moirand, PRAX 33)

Si les résultats présentés ci-dessus sont cruciaux pour décrire le genre de l'article, qui ne saurait être caractérisé sans champ de contraste, ils demeurent ainsi globalement indiciels et nécessitent d'être substantiellement approfondis sur les plans quantitatif et qualitatif.

3.4.3.2. Les personnes et leurs marqueurs

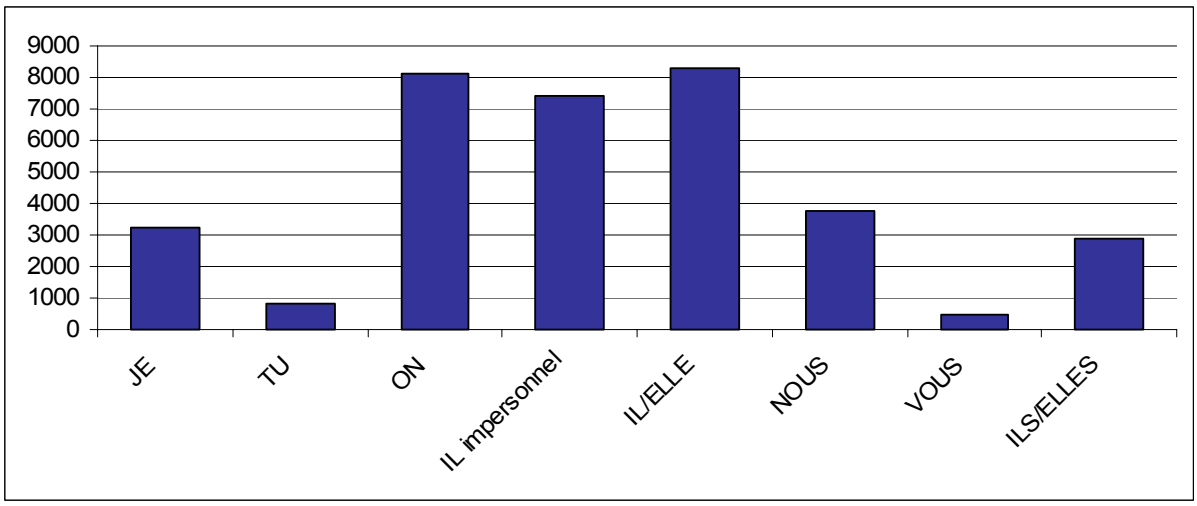
Dans la mesure où l'ensemble des marques de personne leur est globalement corrélé[11], on examinera d'abord les pronoms personnels et leur répartition avant d'analyser les marqueurs qui leur sont associés (clitiques, disjoints, déterminants et pronoms possessifs) ; comme l'illustre le graphique qui suit, les pronoms sont en effet les marqueurs personnels les plus représentés :

[11] L'ensemble des marques d'une personne est en effet d'abord inter-corrélé au niveau textuel avant d'être associé à d'autres variables.

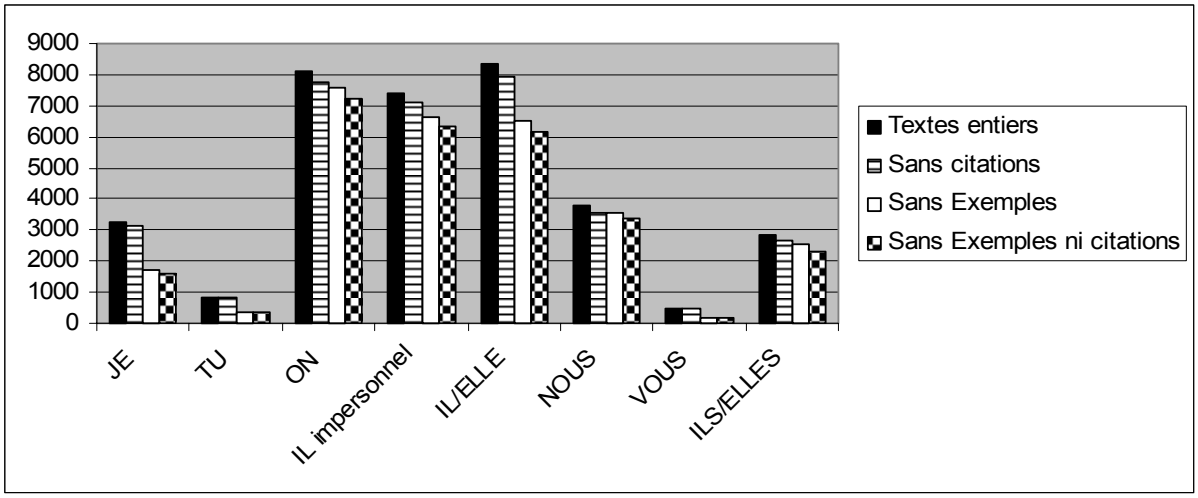


Graphique : Répartition des marques de personne dans les textes sans exemples (moyennes absolues par texte)

3.4.3.3. Répartition des pronoms personnels



Graphique : Répartition des pronoms personnels dans les textes entiers



Graphique : Répartition des pronoms personnels en chiffres absolus avec/sans exemples et citations

Les pronoms de deuxième personne sont d'abord les plus marginaux : contrairement à d'autres genres comme la lettre ou parfois le roman, l'article ne contient pas en principe d'adresse au lecteur. Comme l'illustre le graphique 29, les pronoms de deuxième personne relèvent globalement des exemples des textes.

Si les nombres de *je* et *nous* sont comparables dans les textes entiers (graphique 28), on notera que le nombre de *je* chute de manière importante après extraction des exemples : *je* est ainsi particulièrement présent dans les exemples des textes, contrairement à *nous*, qui demeure relativement stable, à l'instar des pronoms *on* et *il* impersonnel. *Nous* semble d'ailleurs particulièrement peu employé dans les exemples des articles

Les pronoms de troisième personne du singulier sont ainsi les plus employés des textes scientifiques. On observe un emploi important des pronoms impersonnel et indéfini *il* et *on*, souvent remarqué dans les textes scientifiques français[12], qui se conforment encore largement aux recommandations de la tradition.

Le nombre important d'anaphoriques singuliers par rapports aux anaphoriques pluriel serait enfin susceptible de refléter l'emploi privilégié d'éléments au singulier, candidats concepts ou personnes dans le genre de l'article.

3.4.3.4. Répartition de l'ensemble des marques de personne

Comme on l'a déjà fait observer, les pronoms possessifs et les marques de seconde personne (en jaune) sont nettement marginaux : leur coefficient de variation est en effet bien supérieur à 3, soit 300%.

On notera également la marginalité des pronoms disjoints *moi* et *nous*. Si les disjoints de troisième personne sont globalement peu représentés dans les textes, ils demeurent pourtant caractéristiques du genre, dans la mesure où leur coefficient de variation varie à moins de 2.

La première personne du singulier se manifesterait davantage avec les formes clitiques que possessives, au contraire de la première personne du pluriel, ce qui est dû au caractère inclusif de l'englobant NOTRE, contrairement au plus individuel MON.

[12] Ce qui a d'ailleurs conduit Fløttum (2003) à qualifier les français d'« indefinite French ».

Les chaînes de référence, et syntagmes nominaux de rattachement potentiels des possessifs de troisième personne du singulier et du pluriel n'ayant pas été pris en compte dans ce relevé, les possessifs de troisième personne du singulier et du pluriel, de même que le réflexif SE ne sont bien entendu présentés qu'à titre indicatif.

3.4.3.5. Approfondissement en corpus : analyse des pronoms impersonnel et indéfini IL et ON

Nous avons jusqu'à présent observé les marqueurs de personne de manière systémique et générale, en termes quantitatifs de répartition et de fréquence. Il serait évidemment souhaitable d'examiner chacun de ces marqueurs en corpus, afin de mettre au jour les phénomènes d'énonciation représentée qui caractérisent le genre de l'article (manifestations de l'auteur dans le texte, rôles et fonctions associées aux marqueurs de personnes dans le texte, etc.), mais ce type d'analyse n'est raisonnablement pas envisageable, d'une part parce que l'étude des personnes dans le genre de l'article pourrait elle-même constituer l'objet d'un travail de doctorat, et d'autre part parce que la présente étude vise essentiellement à caractériser l'ensemble du genre de manière quantitative, avec toutes les réserves que ce parti pris implique.

S'il n'est donc pas envisageable d'analyser l'ensemble de ces marques, il est en revanche raisonnable de mener une étude plus contextuelle sur un ensemble restreint de marqueurs : les deux pronoms les plus représentés du corpus – l'indéfini *on* et l'impersonnel *il* – ont ainsi été approfondis en contexte et en corpus, ce qui permettra de contribuer à décrire le genre de l'article, tout en présentant un type d'analyse envisageable conciliant quantitatif et qualitatif.

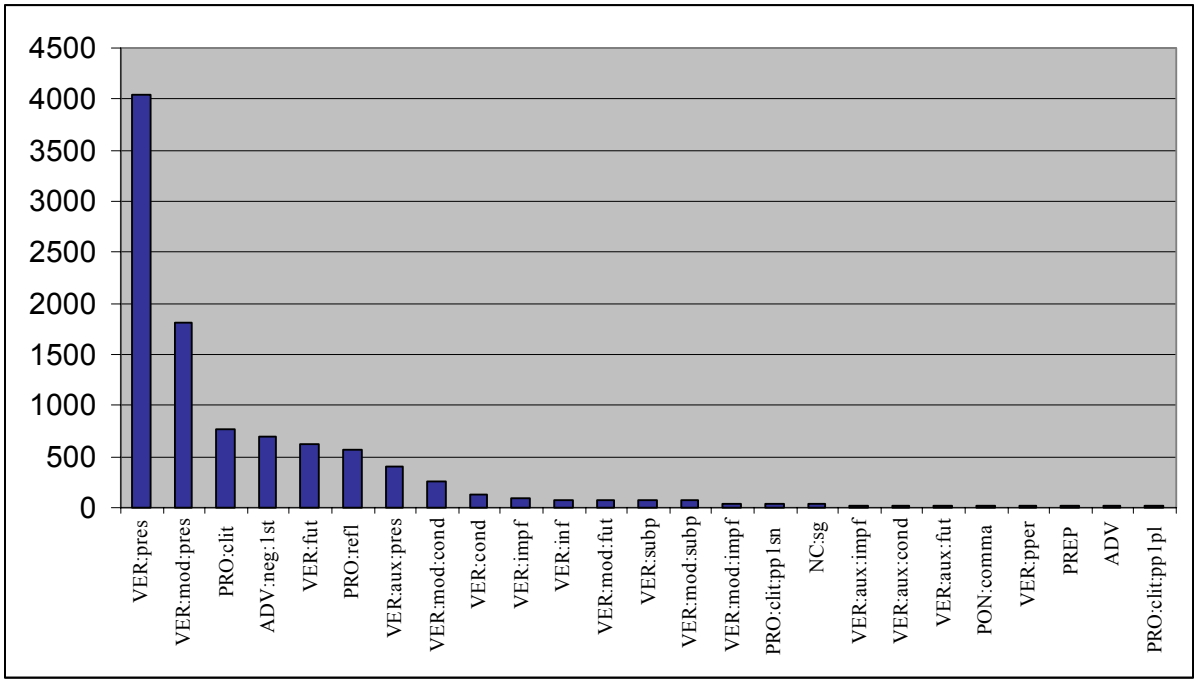
A. ON

1. Pronom le plus représenté et le plus indéfini

Le pronom *on* est particulièrement représenté dans le genre de l'article (36.29 occ. par FT / 33,82 occ. par TE en moyenne, soit 8.39% de l'ensemble PP employés par FT / 9% TE). Il est présent dans tous les textes, avec un minimum de deux occ. par texte, observé dans un article du corpus seulement.

On est le pronom personnel le plus indéfini et le plus ambigu, dans la mesure où il peut renvoyer successivement ou simultanément à l'auteur, à son lecteur ou à une communauté déterminée ou indéterminée. Cette indétermination – volontiers maintenue[13] - permet à l'auteur d'introduire un objet sans pour autant le prendre en charge de manière explicite. Plutôt que d'employer les PP *je* ou *nous*, qui le désignent ou l'incluent, l'auteur emploie plus volontiers le pronom *on*, individu type, acteur compétent du processus de recherche qui *observe, voit, remarque* ou *trouve*.

[13] Alors que le pronom ON est désambiguïsable en contexte.



Graphique : Répartition des co-occurrents droits de ON dans les textes entiers (fréquences absolues, seuil de 10)

Remarque : Par souci de représentativité, le graphique qui précède prend bien entendu en compte les temps verbaux associés aux formes clitiques et réflexives, de même que les temps verbaux suivant les négations, ou les clitiques associés aux formes modales (e.g. *on peut y trouver*…).

2. *ON* atemporel

Comme l'illustre le graphique 30, *on* est massivement employé au présent (4041 occ. de verbes conjugués au présent + 1813 occ. de modaux au présent). Le pronom paraît ainsi avoir une valeur *épistémique/actuelle/atemporelle* ; il est globalement peu employé aux temps du passé[14] et si l'on examine les emplois de *on* au futur, celui-ci n'a pas une valeur véritablement programmatique/prospective (*on reviendra ensuite sur*...) : la majorité des futurs employés est épistémique, et un impératif présent leur est substituable :

> Finalement, sans vouloir sur-évaluer les considérations formelles, **on soulignera [soulignons]** d'une manière générale le caractère facultatif du travail de figuration qui accompagne les actes de discours en question. De surcroît, l'atténuation de ces actes est sujette à des restrictions: Je vous prie de trouver ci-joint... ne saurait être remplacé, par exemple, par *Pourriez-vous trouver ci-joint...? (040)

On est d'ailleurs corrélé aux impératifs du texte, indices de l'utilisation d'un style dialogique (+0,18 dans les textes entiers, + 0,23 dans les textes sans exemples). Le caractère indéfini et donc potentiellement inclusif du pronom lui confère globalement un statut dialogique peut-être plus prononcé qu'avec *nous*, qui renvoie dans nombre de cas à l'auteur

[14] On relève d'ailleurs une corrélation négative de ON avec les temps de l'imparfait et du plus-que-parfait sur les textes entiers.

seul dans les textes scientifiques. *Nous* étant de surcroît de registre plus soutenu que *on*, ce dernier entraîne probablement une plus grande proximité avec le lecteur.

On relève d'ailleurs une corrélation négative intéressante qui oppose *on* à *nous* (-0,2) lorsqu'on examine les textes dont les exemples ont été extraits, ce qui souligne bien une concurrence potentielle des deux pronoms.

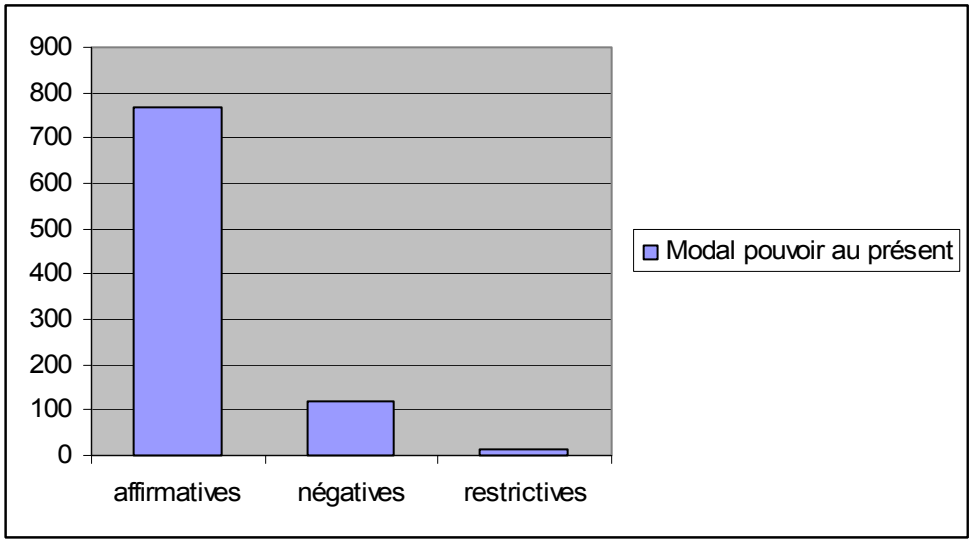
3. *ON* + modal pouvoir

On est globalement corrélé aux modaux au futur (+0,27/+0,28) et au présent (+0,23/+0,22) au niveau textuel, et plus spécifiquement au modal *pouvoir,* qui représente ainsi 88,72% des occ. de modaux au présent relevées, soit 1377 occ. sur les 1552 en chiffres absolus. En termes de co-occurrents, on relève également une proportion significative d'emploi du pronom avec le modal *pourrait* (250 occ. de modaux au conditionnel et 235 *pourrait*).

On + pouvoir renvoie plus directement à la notion de possibilité, le pronom pouvant aussi bien être glosé par « tout chercheur compétent » que par « quiconque », selon son contexte d'apparition. Ainsi, *on* ne semble renvoyer à aucune instance particulière dans l'énoncé qui suit :

> La conclusion que **l'on peut tirer** de ces faits est donc qu'il n'existe pas de phénomène d'accord entre un prédicat nominal et les groupes nominaux qui se trouvent dans son domaine (dans le même groupe nominal que lui)…(178)

Fait notable, *on + pouvoir* est massivement employé dans des tournures affirmatives (graphique 31). En ce sens, la séquence renvoie à la notion de *possibilité* plutôt qu'à celle d'*impossibilité*.



Graphique : Tournures d'emploi de ON + pouvoir

On pourra a également une valeur de potentialité/prédiction. Bien que le futur nous semble plus dialogique, il est aisé de lui substituer un présent :

> **On pourra (peut) énoncer** Dire qu'il se mariait aujourd'hui / la semaine prochaine (" il fut un temps où il était prévu qu'il se mariait aujourd'hui ") (083)

4. *ON* associé à/ inclusif avec l'auteur = objectivation de la démarche scientifique

Après observation des contextes, le pronom *on* est principalement employé dans les différents temps de la démarche scientifique (observation d'un phénomène scientifique,

analyse des données et formulation des régularités, validation ou infirmation d'une hypothèse, etc.) avec des verbes ayant trait au processus de recherche.

Le pronom permet de maintenir un ton dialogique en proposant au lecteur un pôle d'observation, tout en objectivant les opérations décrites, généralisables puisque leur source est indéfinie, et par conséquent non subjective :

- Objectivation du processus d'observation d'un phénomène linguistique : le phénomène considéré par l'auteur est observable par tous et par conséquent objectif + relation dialogique avec le lecteur, invité à observer le phénomène décrit.

On + (modal POUVOIR) + *observer, noter, constater,* etc.

> **On peut observer** d'une part que les formes visées n'ont pas toutes le même degré d'existence (038).
>
> Dans un deuxième temps, **on peut observer** que le choix du joncteur dépend aussi de la classe du lexème-racine (021)
>
> En ce qui concerne les expressions temporelles, **on observe** une claire dissymétrie entre les trois expressions temporelles (116)

- Dégagement des régularités et objectivation de l'analyse et du processus de validation des résultats (*on constate, obtient, voit, notera,* etc. / *on peut déduire, inférer, établir, conclure, affirmer,* etc.) : la déduction de l'auteur est « logique » et objectivée par le fait que *on* (= tout chercheur compétent) en tirerait des conclusions similaires + implication du lecteur dans les analyses menées.

> Outre les échos sémiotiques et textuels que l'on peut repérer d'un document à l'autre, ce sont les différentes classes de locuteurs convoqués **que l'on peut dégager** ici selon deux types de catégories (005)
>
> **On obtient** deux sortes de graphismes, des unités discrètes ou des unités continues disposées linéairement ou non.
>
> Quelle que soit la dénomination traditionnellement retenue (qui reflète des caractéristiques sociales ou démographiques des locuteurs, plutôt qu'une caractérisation par la situation), **on peut supposer** que c'est ici de fait le contexte d'usage (en particulier dans quelle mesure la possibilité d'interpréter prend appui sur le contexte immédiat) qui l'emporte sur les particularités sociales (087)
>
> **On peut représenter** ce jeu d'emboîtement des paramètres matériels de la communication en complétant le schéma (3) par (4) (126)

On notera que les verbes trop conclusifs comme *déduire, inférer, établir, affirmer, conclure, postuler, etc.* sont le plus souvent modérés par l'emploi du modal *pouvoir*, voire par un marqueur comme *raisonnablement* ou *légitimement* (*on peut raisonnablement affirmer, on peut postuler raisonnablement…*).

> **On peut en conclure** que *sanft* exprime une propriété intrinsèque du terme avec lequel il entre en connexion, alors que *mild* exprime une propriété relationnelle du dit terme de connexion. (015)

ce qui nous semble caractéristique des sciences de l'homme et de la société, et de la tradition scientifique européenne, moins positivistes : nombre des contextes observés se rapportent à la démarche de l'auteur (méthodologique et théorique), et seraient facilement omis ou rapportés avec le pronom *I* dans un contexte anglo-saxon. La possible omission du pronom est particulièrement visible lorsqu'il fait office de précaution oratoire, et plus spécifiquement lorsqu'il est employé avec le modal *pouvoir* suivi d'un verbe de parole (*on peut dire, énoncer, exprimer*, etc.) :

> A cet égard, **on peut dire** qu'un traducteur ayant une bonne formation universitaire doit pouvoir écrire de la même façon que lorsqu'il s'exprime à l'oral sous une forme non relâchée (023)

> **On peut parler** de formes anticausatives ou, pour reprendre la terminologie de Oesterreicher spécialement conçue pour les langues romanes, du sous-groupe du médium au sein de la catégorie de la pseudoréflexivité grammaticale. (029)

Le modal *pouvoir* permet en outre d'atténuer les verbes plus personnels et plus subjectifs comme *penser* ou *se demander* qui, directement rattachés à l'auteur du texte, compromettraient l'objectivité de son propos et son assurance, dans la mesure où il est censé détenir les réponses au(x) problème(s) posé(s).

> **On peut néanmoins penser** qu'elle aura un effet indirect ou différé en suscitant sur le tard chez l'auteur littéraire qu'est aussi Saint-Paul des " thèmes prosodiques " pour reprendre le titre d'une plaquette de 1929… (062)

> **On peut se demander** si (32) ne résulte pas d'une reformulation d'un tel énoncé, en vue d'éviter la répétition des négations ne. (135)

5. *ON* est un autre

De telles séquences sont d'ailleurs potentiellement rattachées à une source autre que l'auteur, et affectées d'une valeur négative : *on* peut ainsi matérialiser les critiques éventuelles qu'il serait possible d'opposer aux éléments soutenus par l'auteur :

> Avec cette forme ainsi orthographiée, **on peut penser que** l'on a affaire à la forme que, **et donc être tenté de** poser que le matériau morphologique des relatives, dans le français parlé, tend à se réduire à cette seule forme, puisque l'on a déjà des énoncés du type… (097)

> **On peut objecter** qu'il s'agit là peut-être d'une différence en termes de vitesse d'élocution, plutôt que d'un conditionnement sociolinguistique indépendant, mais en situation de lecture, nous avons curieusement observé des rapports inversés entre les généra (090)

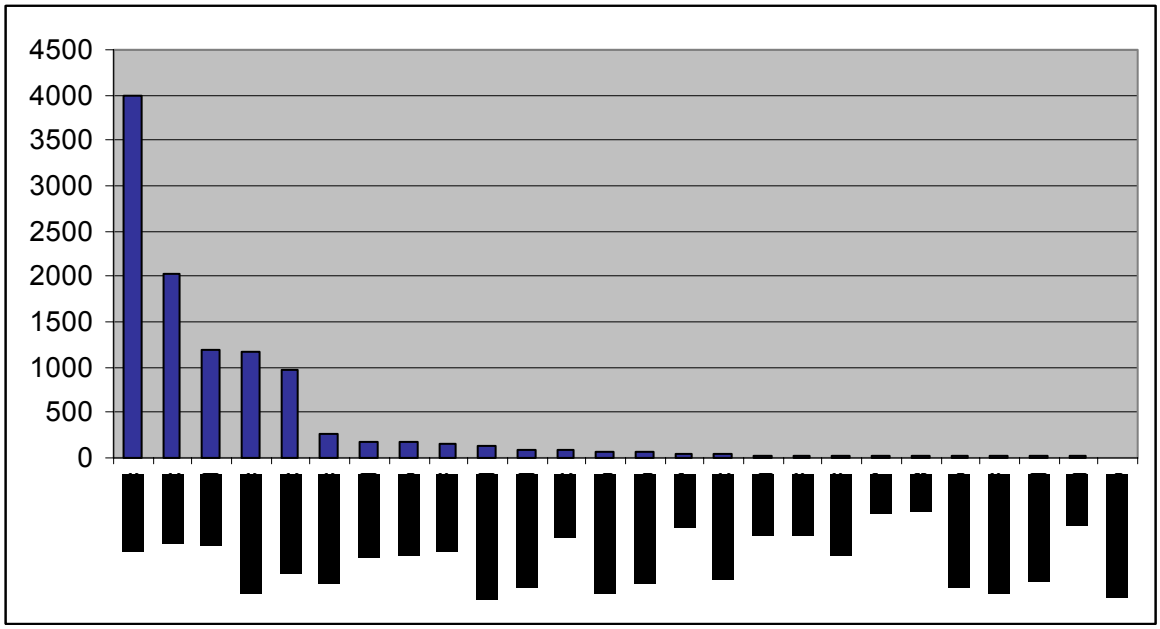
De manière générale, *on* est un non-référent bien utile dans la mesure où il permet d'introduire un objet sans en mentionner la source, ce qui permet à l'auteur d'évoquer des courants et des points de vue scientifiques sans les nommer. Cette indétermination est particulièrement économique dans le cas d'un objet d'étude largement analysé et/ou polémique : *on* limite ainsi les controverses trop frontales en référant sans les désigner à des auteurs ou courants scientifiques de point de vue divergent :

> La traduction en allemand des constructions réfléchies françaises a rarement fait l'objet d'une étude spécifique: **on considère généralement** qu'il n'existe pas de règles et que la traduction de chaque verbe fait l'objet d'un traitement lexical ad hoc.

> Il s'agit, cette fois encore, d'une recherche propre à un genus dicendi, la légende ou l'épopée en l'occurrence, genre qui relève de ce qu'**on appelle** aujourd'hui, à tort ou à raison, un type textuel ou discursif spécifique : la narration (Bouquet, PRAX 33)

B. *IL* impersonnel

Second PP le plus employé, le pronom impersonnel *il* est à l'instar de *on* traditionnellement associé à un effort d'objectivation et de dépersonnalisation / déresponsabilisation de l'auteur. Si l'auteur est moins visible, ses choix scientifiques ne le sont que plus. Il serait bien entendu souhaitable par la suite de déterminer précisément dans quelles configurations l'auteur choisit de s'effacer, par opposition aux séquences dans lesquelles il choisit d'apparaître.

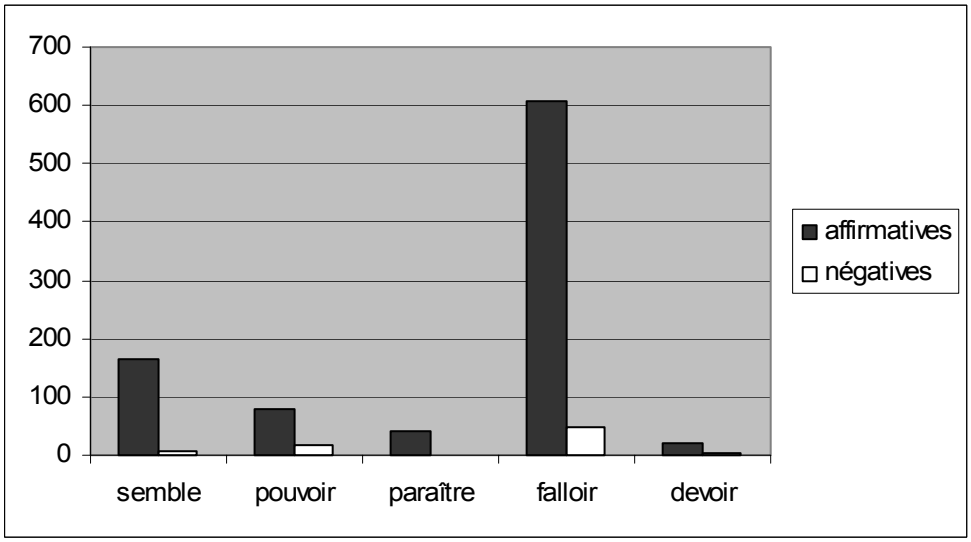


Graphique : Répartition des co-occurrents droits de il impersonnel dans les textes entiers (fréquences absolues, seuil de 10)

Malgré des régularités significatives, notamment en ce qui concerne l'emploi massif du présent (environ 40% des contextes), on relève toutefois des différences notables entre les deux pronoms (graphiques 30 et 32).

L'impersonnel est d'abord sensiblement moins employé avec des modaux au présent que le pronom *on* : 12,57% *vs.* 19% des co-occurrences droits sont des modaux au présent[15], sachant que les modaux associés à l'impersonnel sont plus diversifiés.

Si *on* était corrélé avec la notion de possibilité, l'impersonnel semble plus déontique, comme l'illustre le graphique 33 :



Graphique : Tournures d'emploi de il impersonnel + MODAL au présent

[15] On remarquera que nous avons pour ce calcul comptabilisé les formes modalisées réfléchies comme *il se peut*.

De la même manière que *on* + *pouvoir* exprimait plus volontiers la possibilité que l'impossibilité, *il* + *falloir* ou *devoir* expriment l'obligation plutôt que l'interdiction. Il en va également de même pour l'expression *il convient de*, toutefois non traitée comme modale.

Cependant, l'obligation exprimée par *il impersonnel* est à nuancer : après observation des contextes, peu d'emplois expriment l'obligation en tant qu'engagement ou nécessité, comme dans :

> Pour faire partie de la mémoire discursive, c'est-à-dire être identifiable par l'interlocuteur, **il faut que** le référent figure dans une première mention ou soit accessible par inférence à partir du contexte. (220)
>
> Dans le titre en revanche du film de Peter Greenaway, **il faut nécessairement choisir** entre les antécédents (144)

il faut est ainsi, à l'instar de *on peut*, essentiellement rhétorique, comme l'illustrent les exemples qui suivent :

> Cette distinction une fois posée, **il faut ajouter que** la frontière entre morpho-syntaxe et pragma-syntaxe ne passe pas toujours au même en-droit pour tous les locuteurs. (191)
>
> **Il faut alors s'interroger** sur la spécificité linguistique des processus déviants. (163)
>
> En plus de la notion de point de vue, **il faut considérer** la manière dont l'artefact est envisagé, appelons cela provisoirement sa " facette " (12)

Les pronoms *on* et *il* impersonnels partagent ainsi une fonction d'objectivation. Si le pronom *on* permet par son indéfinitude de généraliser et d'objectiver la recherche rapportée et ses processus, l'impersonnel l'objectivise en la dépersonnalisant plus complètement : les observations et interprétations émises sont ainsi présentées avec le modal *falloir* comme des contraintes externes imposées par la science ou par la logique, que le lecteur est donc tenu d'accepter.

L'effet d'objectivité est particulièrement visible avec *il* + clitique, association d'ailleurs remarquée au niveau textuel, avec une corrélation de +0,37 entre les deux variables ; contrairement à *on*, qui est indifféremment employé avec les clitiques y, en, la, le ou les, l'impersonnel est systématiquement associé aux clitiques EN et Y. La proportion de structures présentatives en *il y a* est éloquente : on en relève pas moins de 1200, soient les 3/5 des séquences *il* + clitiques.

Les constructions en *il y a* génèrent ainsi un effet de réel remarquable :

> Dans cet exemple, **il y a** de multiples déplacements de Cr entre énoncés, aucun topique ne réussissant à s'établir pour la durée du segment maximal (171)
>
> **Il y a** des strates dans la diachronie et, en cela, la synchronie garde bien la trace de la diachronie. (102)

Bien qu'elles n'aient pas les mêmes emplois et valeurs, d'autres types de structures également très représentées dans le corpus participent à la mise en place de cette objectivité : *il (ne) s'agit (pas) de* (810 emplois observés, soit 20,25% des emplois de l'impersonnel au présent) et *il (n') existe* (366, soit 9%) sont ainsi les expressions relevées les plus fréquentes.

Les tournures impersonnelles sont particulièrement employées lorsqu'il s'agit de dégager des conclusions :

> **Il en résulte** que la relation entre le sens de l'adjectif et celui du syntagme entier peut alors pleinement tirer parti de la multitude de rapports de contiguïté possibles (217)
>
> **Il s'ensuit** qu'un énoncé peut avoir plusieurs centres qui sont appelés centres anticipateurs (Ca). (114)

Des exemples qui précèdent **il ressort** clairement que tel anaphorisant a vocation à récupérer des éléments notionnels destinés à caractériser le constituant nominal auquel il se rapporte. (134)

Si l'autonomisation des phénomènes et des objets observés est particulièrement fréquente dans les textes scientifiques – un objet *s'analyse, se définit, se comporte* ou encore *se manifeste* -, les tournures impersonnelles renforcent cette représentation bien ancrée dans la tradition occidentale européenne d'une science qui se construirait d'elle-même avec la seule logique.

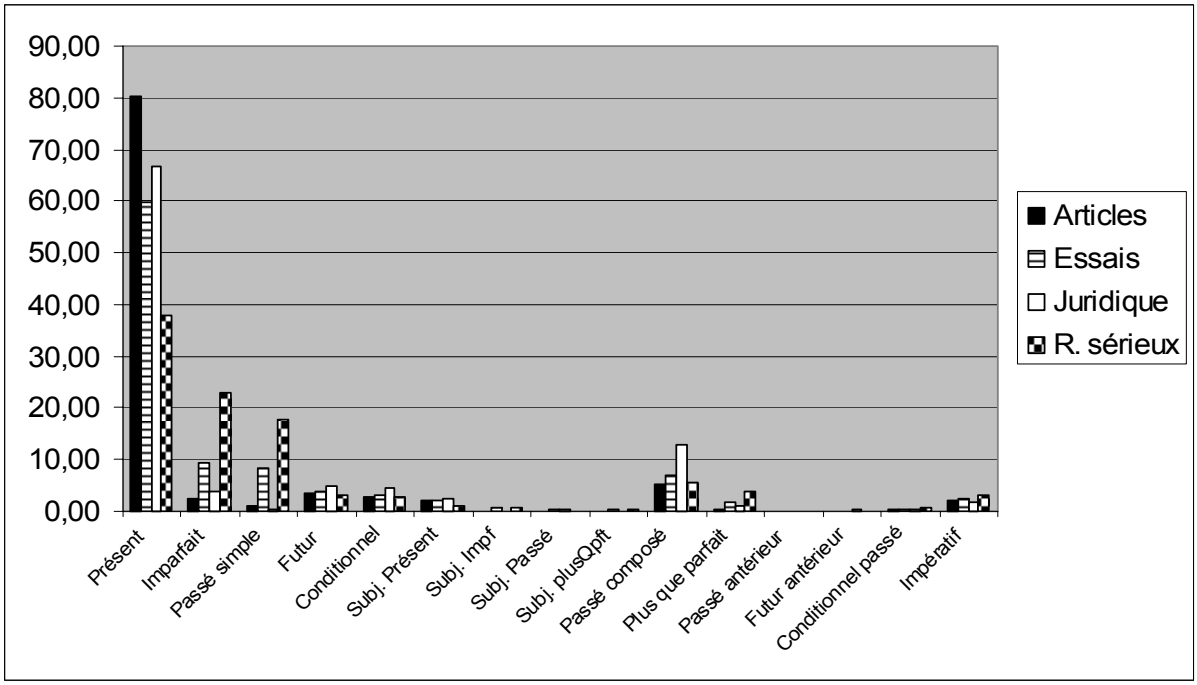
Au niveau textuel, l'impersonnel entre de manière non surprenante en concurrence avec les pronoms démonstratifs (-0.24 TE), et l'on relève une corrélation du pronom au mode du subjonctif (passé, imparfait et présent). Plus d'un dixième des subjonctifs présents est en effet conjugué avec *il* (177 sur les 1139 recensés).

Il impersonnel est significativement corrélé au conditionnel (+0.21 FT/+0.14 TE) et aux modaux au conditionnel (+0.14 TE), bien que le pronom ne soit pas particulièrement conjugué avec ce temps. A fortiori, *on* était comparativement plus employé au conditionnel sans que l'on relève de corrélation significative entre les deux variables au niveau textuel. La corrélation observée semble ainsi indiquer une dimension spéculative des textes dans lequel apparaît l'impersonnel. Contrairement à *on*, l'impersonnel n'est pas corrélé aux indices d'un style dialogique, et il est négativement corrélé à l'ensemble des marques de seconde personne du singulier (-0.19 *tu*, -0.15 *ton*, *ta*, *tes*, -0.14 *toi* et -0.13 *te* clitique), qui apparaissent quasi exclusivement dans les exemples des textes.

3.4.4. Des temps verbaux

3.4.4.1. Genre de l'article vs. genres autres

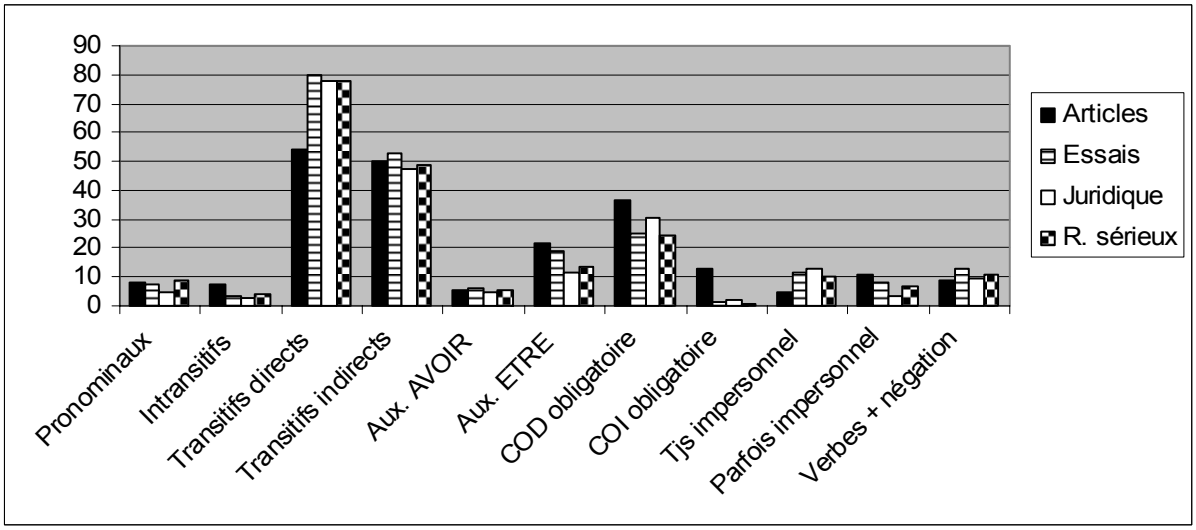
L'article se démarque d'abord par un usage massif du présent. Dans le graphique ci-dessous, c'est en effet le seul temps qui émerge distinctement. En revanche, il en va différemment des trois autres genres, qui semblent caractérisés par un panel plus important de temps conjugués : présent, imparfait, passé simple et passé composé pour le roman sérieux et l'essai de manière moindre, et présent et passé composé pour les textes juridiques.



Graphique : Répartition en pourcentage des verbes par temps conjugué et par genre

En ce qui concerne les caractéristiques des formes verbales du genre de l'article, on relève un emploi plus important des verbes intransitifs que dans les autres genres. Si les verbes transitifs sont de loin plus représentés dans les quatre corpus observés, le genre de l'article se démarque nettement des autres par un usage indifférencié des formes transitives directe et indirecte : les verbes transitifs directs sont considérablement plus représentés dans les trois autres genres.

On observe par ailleurs un emploi plus important de verbes « parfois impersonnels » que « toujours impersonnels » : si ces éléments ne peuvent tenir lieu que d'indices, dans la mesure où nous ne disposons pas des listes des verbes « toujours » ou « parfois » impersonnels élaborées par Cordial, le genre de l'article pourrait se caractériser par un emploi impersonnel de verbes qui ne le sont pas au départ.

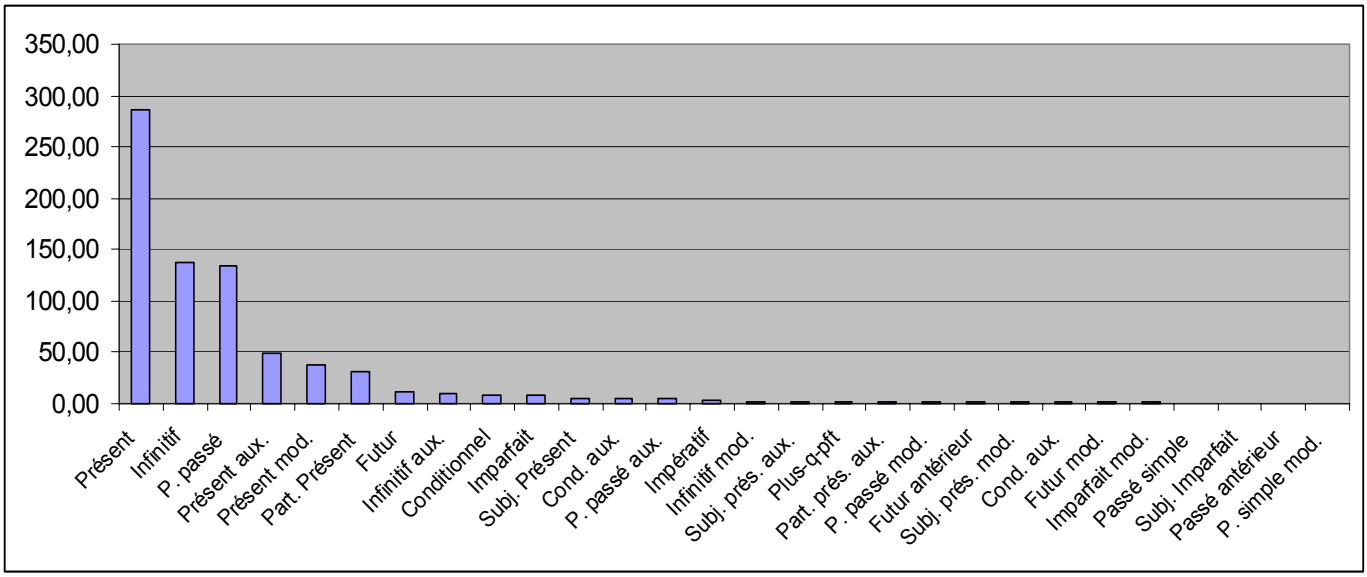


Graphique : Caractéristiques des formes verbales par genre

3.4.4.2. Répartition des formes verbales et conjuguées au sein du genre

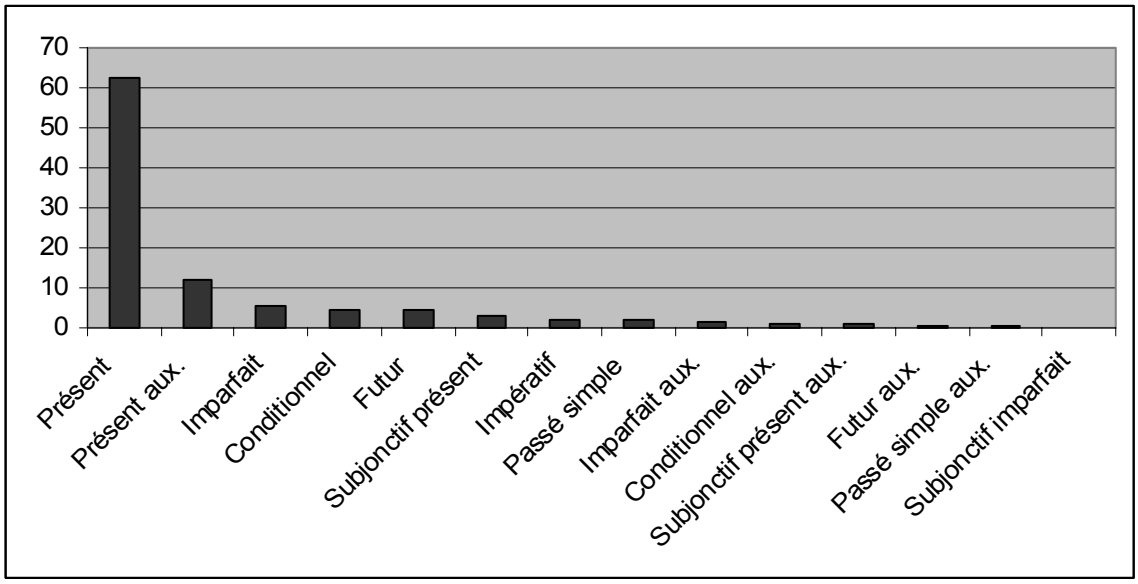
Remarque : les statistiques descriptives obtenues sur textes entiers et textes sans exemples sont présentées en annexe 10.

Examinons d'abord les résultats obtenus sur l'ensemble des formes verbales :



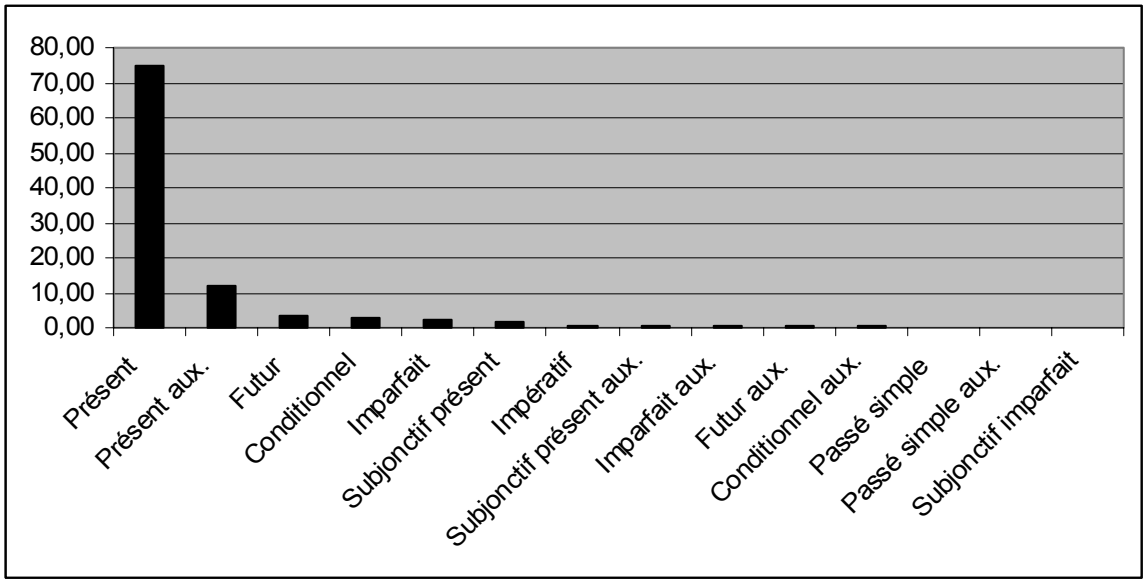
Graphique : Répartition de l'ensemble des formes verbales dans les dont les exemples ont été extraits (moy. absolues par texte)

Si l'on s'intéresse aux formes non conjuguées, on relève un grand nombre de formes infinitives et participiales ; notons que les participes passés, indices de formes passives et de verbes conjugués au participe passé[16] sont 4 à 5 fois plus représentés que les participes présent.



Graphique : Répartition des temps conjugués dans les textes entiers (% par texte)

[16] Que nous n'avons malheureusement pas pu désambiguïser.



Graphique : Répartition des temps conjugués dans les textes sans exemple (% par texte)

Comme l'illustrent les graphiques 37 et 38, on relève des variations significatives avec/sans extraction des exemples, malgré des régularités notables. Ainsi, le présent est massivement représenté[17] : avec une moyenne de 357,22 verbes conjugués au présent par FT[18] / 324,08 par TE, il représente 62% des formes conjuguées FT / 75,07% TE. Les modaux pris en compte y tiennent 13% FT / 7,5% TE, ce qui est considérable.

Les variations importantes observées avec/sans exemples indiquent que les exemples contiennent nettement moins de verbes conjugués au présent que le corps de l'article n'en contient, ce qui laisse penser qu'ils ne relèvent pas du même genre. On relève ainsi des différences significatives en ce qui concerne les temps de l'imparfait et du passé simple, temps narratifs par excellence, et leurs homologues composés : leurs proportions chutent considérablement après extraction, tandis que la plupart des autres temps conservent des rapports similaires. On observe également des différences notables pour l'impératif : si certaines formes impératives conjuguées à la première personne du pluriel sont caractéristiques d'un style dialogique (*convenons, soulignons, mentionnons,* etc.), la plupart des impératifs semblent contenus dans les exemples des textes, et très probablement dans les articles traitant de l'oral.

3.4.4.3. Analyse des corrélations verbales

On retrouve ces différences lorsqu'on examine les corrélations textuelles des formes verbales dans les textes avec/sans exemples : si les temps de l'imparfait et du plus-que-parfait, et du passé simple et du passé antérieur sont dans les deux cas les plus corrélés en positif, on observe des ensembles de corrélations plus stables et plus interprétables dans les textes dont les exemples ont été extraits :

Imparfait	**Aux. imparfait**

[17] On notera qu'il est sensiblement moins représenté qu'il ne l'était avec Cordial, ce qui est nettement dû à l'annotation des formes passives en tant qu'auxiliaire présent.

[18] En cumulant l'ensemble des formes conjuguées au présent, modaux compris.

FT	TE	FT	TE
Aux. imparfait **0,635** Passé simple **0,459** Aux. passé simple **0,458** Noms propres **0,293** ME **0,262** Antislashs **0,241** JE **0,232** Modaux imparfait **0,229** D. poss. pp1sg **0,217** Participes passé **0,200** Modaux part. passé **0,198**	Aux. imparfait **0,777** Passé simple **0,586** Aux. passé simple **0,570** Modaux imparfait **0,359** Noms propres **0,316** C. addition **0,253** D. poss. pp3sg **0,233** Mod. part. passé **0,229** Subj. imparfait **0,222** Dates **0,208** Abréviations **0,208**	Imparfait **0,635** Aux. conditionnel **0,462** Aux. passé simple **0,380** Conditionnel **0,333** Passé simple **0,303** VOUS **0,288** Auxiliaires ETE **0,283** Participes passé **0,281** Modaux passé simple **0,244** Modaux imparfait **0,214**	Imparfait **0,777** Aux. passé simple **0,539** Passé simple **0,416** Modaux imparfait **0,319** Auxiliaires ETE **0,277** Noms propres **0,256** Participes passé **0,251** D. poss. pp3sg **0,241** VOUS **0,240** Modaux part. passé **0,233** Sigles **0,218** Dates **0,216**
Présent **-0,346** Modaux présent **-0,266**	Présent **-0,310** Modaux présent **-0,270** Renvois **-0,228** Deux points **-0,186**	Présent **-0,367** Modaux présent **-0,334**	Présent **-0,406** Modaux présent **-0,330** Renvois **-0,216** Deux points **-0,184**

Tableau : Corrélations positives et négatives de l'imparfait et de l'auxiliaire imparfait dans les textes avec/sans exemples

Ainsi, les imparfaits et plus-que-parfait semblent inclus dans une dimension historico-narrative (surligné), caractérisée par l'emploi des temps narratifs, des dates, noms propres et marqueurs de troisième personne (thématisation d'un objet singulier en particulier, candidat concept ou personne). Cette dimension est toutefois moins marquée dans les textes entiers : l'imparfait y est également corrélé à plusieurs marques de première personne, visiblement contenues dans les exemples des textes, tandis que le plus-que-parfait est également associé au conditionnel et au pronom personnel *vous*, corrélation que l'on retrouve après extraction.

Cette dimension historico-narrative semble s'opposer à des caractéristiques avérées du discours scientifique : présent, modaux au présent, deux points et indices de renvois, particulièrement présents dans les textes exemplifiés.

Si nous n'analyserons pas précisément chacune des corrélations verbales observées, examinons les résultats obtenus sur les quatre temps les plus représentés du corpus[19]. Etant donné l'hétérogénéité des corrélations observées sur textes entiers, nous n'analyserons que les résultats sur TE, plus stables et plus interprétables.

Présent	Aux. présent	Futur	Conditionnel
Ensemble des pronoms 0,41	VER:aux:pper 0,62	VER:mod:fut 0,46	VER:mod:cond 0,45
LS 0,24	VER:pper 0,51	VER:aux:fut 0,24	VER:aux:cond 0,35
NC:sg 0,24	Ensemble des noms 0,31	VER:imp 0,19	CON:dou:1st 0,31
Adverbes et connecteurs 0,24	CON:add:1st 0,31	DET:indef 0,17	Ensemble subordonnants 0,30
ADJ:refl 0,21	Ens. des numéraux 0,26	PON:colon 0,16	PUL 0,27
%classe2 0,21	NC:pl 0,24	SYM:gram 0,16	%classe4 0,21
ADV:neg:1st 0,18	PON:dot 0,23	PRO:pp1pl 0,16	DET:poss:pp2pl 0,21
ADJ:sg 0,18	DET:def 0,22	SIG:ling 0,16	PRO:pp3isn 0,21
PON:colon 0,16	ADJ:pl 0,21	CON:ref:1st 0,16	PRO:clit:pp2pl 0,19
VER:imp 0,16	DTC:pl 0,21	NUM:ord 0,15	CON:opp:1st 0,19
INT 0,15	PRO:pp3pl 0,18	PRO:clit:pp1pl 0,14	VER:subp 0,19
FGW 0,15	PON:par 0,17		PRO:clit:pp2sn 0,18

[19] En excluant leurs versants modaux.

CON:cau:1st 0,14 SYM 0,13	PRO:pp1pl 0,16 VER:impf 0,16 NP 0,15 VER:aux:impf 0,15		PRO:pp2pl 0,18 PON:int 0,17
VER:pper -0,50 VER:aux:pper -0,41 VER:aux:impf -0,40 VER:aux:pres -0,33 VER:impf -0,31 Ens. prépositions -0,27 Ens. noms -0,26 VER:parpres -0,24 NC:pl -0,24 VER:inf -0,23 DTC:pl -0,22 VER:aux:inf -0,21 SIG -0,21 NUM:dat -0,19 PRO:pp3pl -0,19 ADJ:pl -0,19 VER:mod:impf -0,19 ADV -0,18 VER:simp -0,15 VER:mod:fut -0,14 VER:aux:simp -0,14 DET:poss:pp3pl -0,14 PRO:disj:pp3pl -0,13	VER:inf -0,40 VER:pres -0,33 VER:mod:pres -0,33 %classe4 -0,32 NC:sg -0,28 %classe2 -0,24 CON:dou:1st -0,24 PON:colon -0,23 DET:indef -0,23 CON:jus:1st -0,22 ADJ:sg -0,21 VER:fut -0,20	PON:par -0,22 CON:add:1st -0,21 VER:aux:pres -0,20 DET:def -0,20 VER:aux:pper -0,17 NUM:dat -0,16 NP -0,16	LS -0,21 PRO:disj:pp3pl -0,19 VER:aux:pres -0,18 CON:add:1st -0,18 DTC:pl -0,18 SUB:1st -0,16 SIG -0,16 %classe7 -0,15 PON:dot -0,15 NC:pl -0,14 PON:par -0,14 VER:pper -0,13 %classe1 -0,13

Tableau : corrélations des temps verbaux les plus représentés (TE)

Force est de constater que l'on relève peu de corrélations textuelles véritablement significatives si l'on considère le temps du présent : en raison de son importante représentation dans l'ensemble du corpus, ce dernier est en effet associé aux éléments les plus représentés (e.g. éléments au singulier, cf. *supra*) ou aux catégories grammaticalement associées aux formes verbales en général (cf. corrélation du présent avec l'ensemble des pronoms). Les corrélations négatives obtenues indiquent une opposition du présent avec les participes passés et les formes auxiliaires les plus représentées (aux. présent et imparfait). On observe enfin une corrélation négative du présent avec les indices de la dimension historico-narrative déjà relevée.

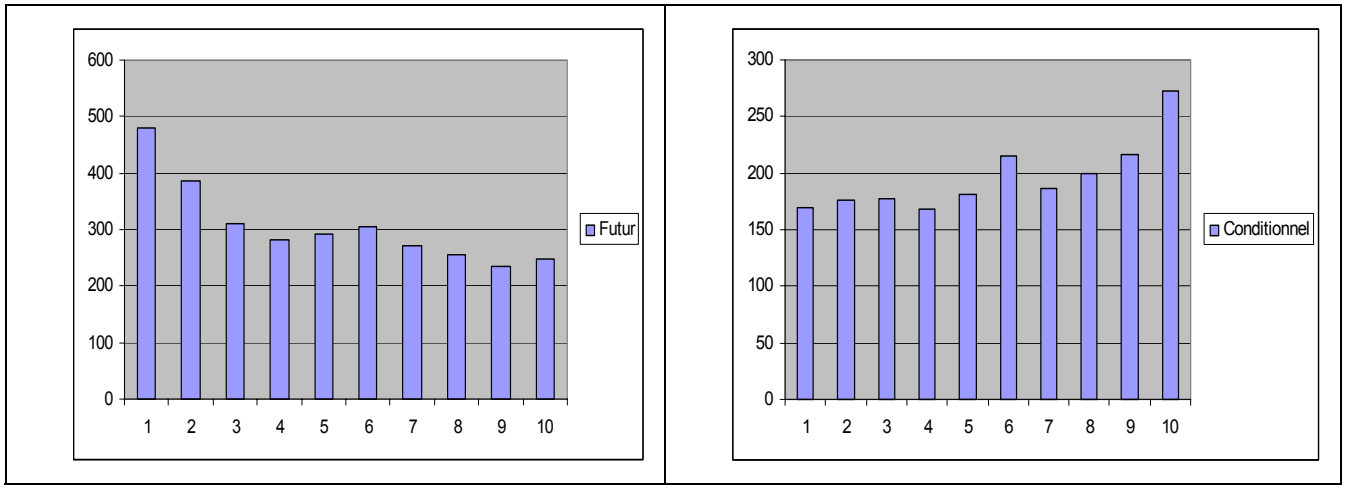
Les corrélations associées aux auxiliaires présent sont également difficilement interprétables, ce qui est très probablement lié à l'ambiguïté de la variable, qui renvoie aussi bien au temps du passé composé qu'aux passifs. On relève par exemple une association du descripteur aux numéraux et au pluriel, de même qu'au pronom *nous*, ce qui est globalement peu significatif.

Les corrélations des temps du futur et du conditionnel sont par contraste beaucoup plus claires : le futur semble être corrélé à une dimension plus inclusive, peut-être dialogique : il est en effet associé aux marques de première personne du pluriel et aux impératifs. Il serait ainsi employé dans des textes plus attentifs au lecteur et à son bon repérage dans le texte (*nous verrons ultérieurement,* etc.), ce qui semble confirmé par la présence d'une part des numéraux ordinaux (*dans un premier temps, deuxième temps,* etc.) et des deux points, et d'autre part des connecteurs de reformulation, qui dénotent un souci de clarté manifeste. On notera que le futur ne s'oppose ni au présent, ni à l'ensemble des temps du passé : on relève

par contre une tension négative avec les auxiliaires présent, associés aux passifs et au passé composé, et les dates, noms propres et parenthèses[20], naturellement corrélés au révolu.

Le conditionnel est quant à lui positivement corrélé au spéculatif (connecteurs de doute et d'opposition, *il* impersonnel), et semble employé dans des textes stylistiquement plus travaillés au regard de ses corrélations avec le subjonctif présent et l'ensemble des subordonnants, indices d'une syntaxe plus complexe. On relève également une corrélation du conditionnel avec les marqueurs de seconde personne du pluriel et du singulier, globalement employés dans les exemples du corpus. Il s'agit vraisemblablement d'un biais du corpus : les exemples imbriqués dans les corps de texte n'ayant pas pu être tous écartés, cette corrélation semble liée à un usage important du conditionnel dans les exemples restant.

Soulignons enfin que les deux temps verbaux manifestent des répartitions tactiques opposées : si l'on observe une décroissance du futur au fil du texte, le conditionnel atteint son maximum en fin d'article :



Graphique : Configurations tactiques du futur et du conditionnel (fréquences absolues par décile de rang)

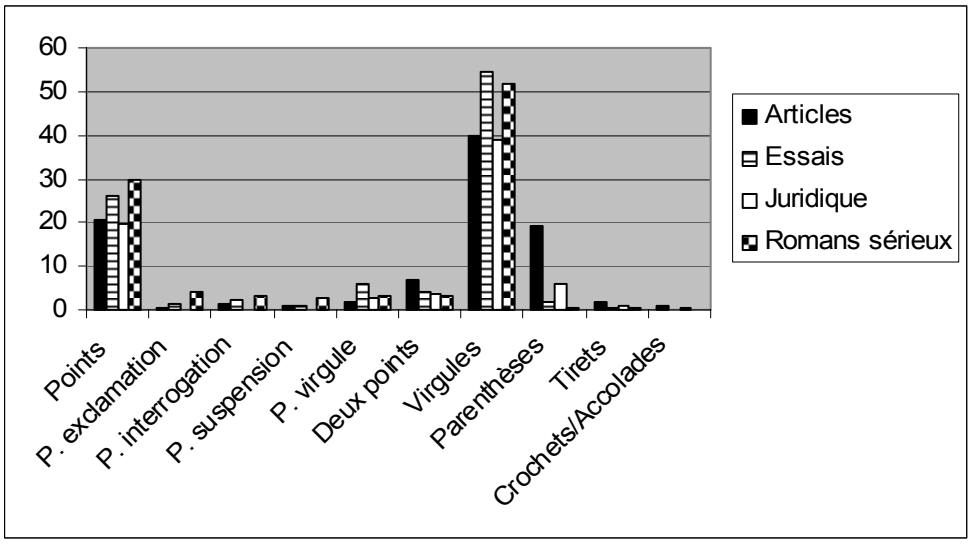
Le futur remplirait donc une fonction de *guide* et serait vraisemblablement associé à l'hypothèse développée au sein de l'article, tandis que le conditionnel serait associé à la conclusion (aux conclusions) de l'article (v. chapitre 4), et ouvrirait le champ de la recherche à de nouvelles hypothèses.

3.4.5. Des ponctuations

3.4.5.1. Genre de l'article vs. genres autres

Considérons d'abord les résultats comparatifs obtenus à partir des sorties de Cordial®, bien que les ponctuations qui nous intéressent ne soient pas toutes prises en compte :

[20] Les parenthèses sont nettement associées aux dates et noms propres, eu égard aux pratiques de citation des textes scientifiques généralement soumises au modèle (Nom_propre, date).



Graphique : Répartition en pourcentage des ponctuations par genre

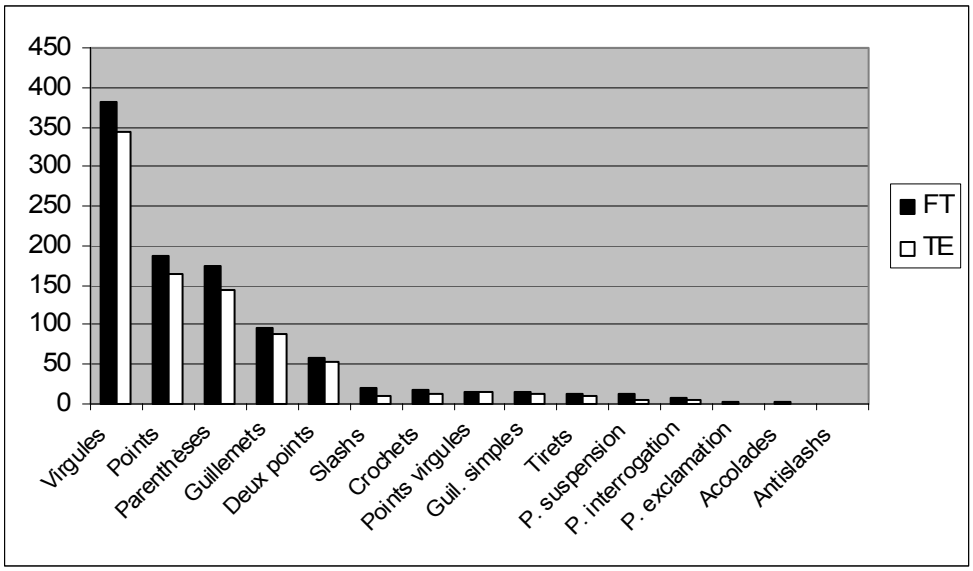
Si les points et les virgules sont les ponctuations les plus représentées dans les quatre genres observés, le genre de l'article semble se caractériser par une proportion de parenthèses très importante (presque équivalente à celle des points). On remarquera également la représentation significative des deux points, liée à la fonction démonstrative de l'article, et celles, moins visibles sur le graphique, des tirets, crochets et accolades, indices de la présence d'un métalangage et de marques de formalisation.

On relève à l'inverse un nombre peu élevé de points d'exclamation : si les textes juridiques n'en contiennent aucun, l'essai, plus individuel, en contient trois fois plus et le roman sérieux six à sept fois plus que l'article. Les points de suspension semblent enfin relativement marginaux dans les genres de l'article et de l'essai (0.7% de l'ensemble des ponctuations), tandis que les textes juridiques en contiennent sept fois moins, et le roman quatre fois plus.

3.4.5.2. Représentation des ponctuations

Remarque : les statistiques descriptives obtenues sur textes entiers et textes sans exemples sont présentées en annexe 10.

Comme cela a déjà été souligné, les ponctuations les plus marginales sont de loin les antislashs, accolades, points d'exclamation et guillemets simples. Les résultats obtenus sur les sorties Cordial® se trouvent dans l'ensemble confirmés : outre une représentation importante de virgules et de points, on remarque une proportion importante de parenthèses (moyenne de 175,35 par FT / 143,54 par TE), de guillemets (moy. de 96,63 par FT / 87,51 par TE) et de deux points (moy. de 59,30 par FT / 53,09 par TE), éléments fréquemment posés comme caractéristiques du discours scientifique.



Graphique : Répartition des ponctuations par texte avec/sans exemples (moyennes absolues par texte)

Si les ponctuations les plus représentées diminuent à proportion égale après extraction des exemples, on observe que les ponctuations plus marginales connaissent d'importantes fluctuations avec/sans extraction : les très marginaux antislashs relèvent ainsi exclusivement des exemples, tandis que la moitié, voire la majorité des accolades, slashs, crochets, points d'exclamation et de suspension y sont employés.

Les slashs et les points de suspension (qui ne sont d'ailleurs pas corrélés) sont souvent mobilisés par les exemples oraux des textes, comme l'illustrent les deux extraits suivants :

(A) j'en ai vu un / de film

lequel / de film ? (/ pour pause)

(B) c'en est une vraiment belle / histoire

laquelle histoire ? (texte 092)

(2) B : Comme quelqu'un... admettons y aurait une gang de filles avec une gang de gars, les gars i **vont dire** "ça **sera** pas chaud". ça veut dire... ça **sera** pas chaud (rire léger) I **va se passer** presqu'un orgie, comme i pourraient vouloir dire, t'sé...

A : hum

(Deshaies, *Corpus de la Ville de Québec*, F02-M, 001080) (texte 169)

tandis que les crochets servent souvent à contextualiser l'exemple :

(18) [Devant une chambre dans le plus grand désordre] *Jamais vu un tel fouillis!*

(18) a. [Devant une mer d'huile] ?? *Jamais vu un tel fouillis!* (texte 134)

Au contraire, les points virgules semblent caractéristiques du genre de l'article, dans la mesure où leurs proportions demeurent quasi strictement identiques.

3.4.5.3. Analyse des corrélations

Les corrélations textuelles reflètent bien évidemment ces différences : les ponctuations qui varient le moins après extraction conservent les mêmes corrélats. Par exemple, les virgules s'opposent fortement aux parenthèses (-0.4) dans les deux cas. Les deux ponctuations semblent ainsi particulièrement opposées : comme l'illustre le tableau suivant, on retrouve la majorité des corrélats négatifs chez l'une en positif chez l'autre, et vice-versa.

Parenthèses			Virgules		
Ensemble symb, sig, abrév.	0,386		Pronoms relatifs		0,283
Numéraux	0,367		Ensemble ADV et connecteurs	0,257	
Déterminants définis	0,310		Amalgames pluriel	0,246	
Dates		0,277	Ensemble prépositions		0,238
Noms propres		0,273	Numéraux ordinaux	0,222	
Symboles	0,211		Connecteurs concession		0,209
Accolades	0,210		Total Formes		0,198
Pronoms réflexifs	0,178		Ensemble déterminants		0,167
Auxiliaires présent	0,177		Connecteurs doute	0,153	
Connecteurs disjonction		0,168	Noms communs pluriel		0,153
Connecteurs addition	0,161		Ensemble pronoms	0,149	
Participes passés		0,156	Sigles		0,145
Virgules		-0,444	Parenthèses	-0,444	
Guillemets	-0,349		Ensemble symb, sig, abrév.	-0,301	
Ensemble pronoms	-0,313		Slashs		-0,283
Dét. poss. PP3SG	-0,236		Guillemets	-0,282	
Futur		-0,224	Ensemble ponctuations		-0,254
Ensemble ADV et connecteurs	-0,221		Deux points	-0,244	
Connecteurs concession		-0,219	Ensemble numéraux	-0,236	
Connecteurs justification		-0,203	Accolades	-0,230	
Déterminants démonstratifs	-0,199		Crochets	-0,193	
Indices strucutration	-0,196		Prépositions	-0,190	
Sigles		-0,195	Cotes		-0,160
Numéraux ordinaux	-0,190		Disjoint TOI	-0,151	
Connecteurs doute	-0,183		Noms propres		-0,146
Déterminants indéfinis		-0,182	Préfixes		-0,137
Points		-0,182	Déterminants définis	-0,136	
Ensemble déterminants		-0,180	Points de suspension	-0,133	
Noms communs singulier		-0,167			
Modaux présent		-0,161			
Ensemble verbes		-0,161			
Ensemble prépositions		-0,160			

Tableau : corrélats des parenthèses et des virgules (TE)

Les parenthèses renvoient ainsi à un changement de registre (parallèle), tandis que les virgules sont associées à une opération de succession (série), et on observe que les parenthèses sont associées à des descripteurs formels. En ce sens, la parenthèse supplanterait la virgule dans des cadres textuels plus formalisés.

Etant donné qu'il est peu envisageable d'examiner le détail des corrélats de chacune des ponctuations, nous ne mentionnerons que quelques résultats parmi les plus immédiatement interprétables : les guillemets simples sont ainsi corrélés avec les éléments de langue étrangère (+0.17), et semblent donc avoir une fonction de démarcation de ces éléments ou de leurs traductions françaises dans le texte[21] :

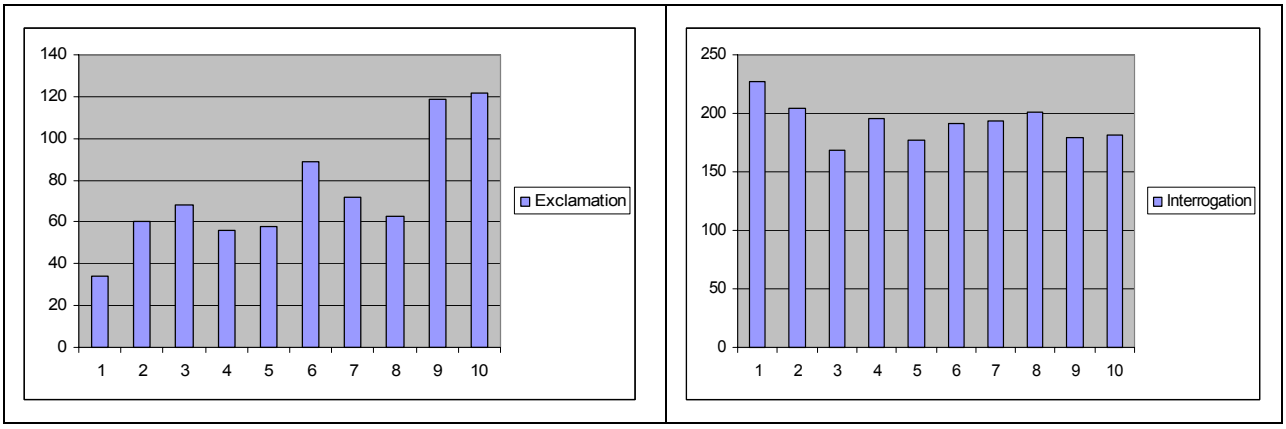
> *Sladkij* (variante dialectale : *solodkij* 'doux' et 'bon, goûteux, savoureux') vient du slave commun *soldŭ-kŭ correspondant à la base i.-e. *sal-d- 'salé', qui est représentée notamment dans l'angl. *salt* et l'all. *Salz* 'sel'. La racine i.-e. est bien entendu *sal- 'sel', cf. le grec *hals* 'sel' et 'mer' (> fr. *halogène*), le lat. *sal* (> fr. *sel, salé)*, le russe *sol'*. (100)

[21] Fonction partagée avec les italiques, gras et éventuellement soulignés, pour lesquels nous ne disposons malheureusement pas de chiffres aussi précis.

Les points d'exclamation semblent associés à une dimension polémique, dans la mesure où ils sont fortement corrélés aux connecteurs d'opposition (0.38). On remarque également une corrélation significative des ponctuations observées avec le pronom *on* (0.29), pronom dialogique de l'objectivation (cf. *supra*). Ce phénomène ne signifie évidemment pas que le point d'exclamation est systématiquement employé dans des phrases polémiques contenant le pronom *on*, comme dans l'exemple qui suit, mais on soulignera que l'association des trois descripteurs est fortement intuitive :

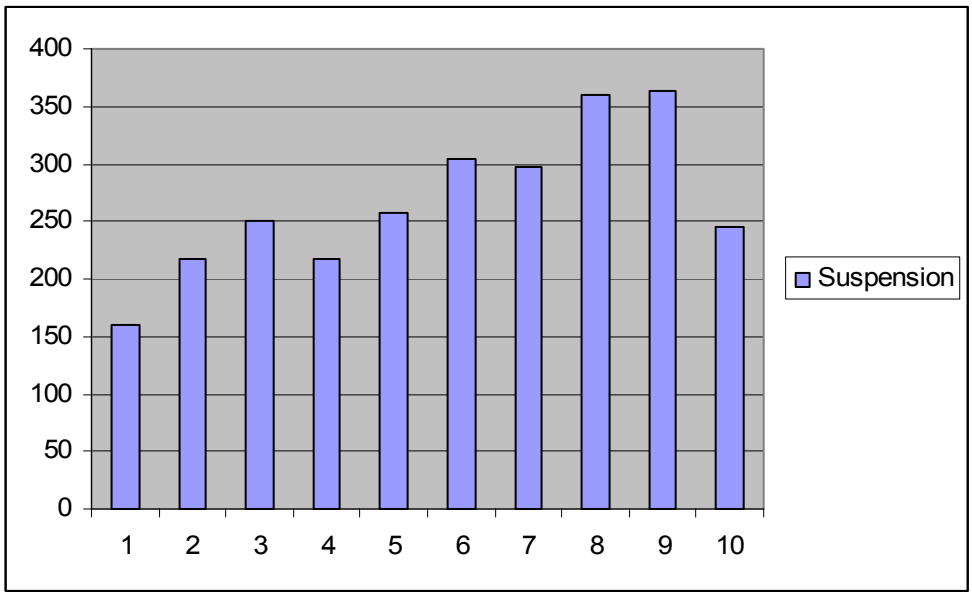
> (…) cette petite étude révèle combien la syntaxe dépend également de la sémantique. En effet, plus **on** s'éloigne de la phrase simple, plus étroite est la dépendance entre sémantique et syntaxe **! Or** au niveau phrastique même, on voit (déjà) l'influence de considérations textuelles, comme nous l'avons constaté en ce qui concerne l'interdépendance entre la focalisation et l'ordre des mots. (200)

On mentionnera enfin que les points d'exclamation, potentiellement polémiques, sont plus employés en fin d'article, contrairement aux points d'interrogation, qui décroissent significativement au fil du texte :



Graphique : Profils tactiques des points d'exclamation et d'interrogation

Ces phénomènes font écho aux configurations du conditionnel et du futur : le début d'article et l'hypothèse de départ s'inscrivent dans une dimension plus interrogative, tandis que la conclusion est plus spéculative, exclamative et suspensive :



Graphique : Profils tactiques des points de suspension

3.4.6. Des éléments étrangers

3.4.6.1. Répartition et corrélations des éléments de langue étrangère

	Min.	% min..	Max.	Total	Moyenne par texte	Ecart type	Coeff. variation	Variance
FT	0	17,41	2457	31016	138,46	297,10	2,15	88269,74
TE	0	20,98	848	12711	56,75	116,56	2,05	13585,23

Tableau : Statistiques descriptives des éléments de langue étrangère

Les éléments de langue étrangère semblent bien caractéristiques du genre de l'article : à peine les 4/5e du corpus n'en contiennent aucun, et le CV obtenu par la variable est globalement peu élevé. Bien que l'on en relève quasiment trois fois moins après extraction des exemples, on en observe pas moins de 56,75 occurrences par texte en moyenne, ce qui est honorable.

On notera toutefois que la médiane du descripteur se situe à 16.5, ce qui signifie que la moitié des textes se situe en dessous de ce chiffre ; un quart du corpus contient entre 0 et 2 éléments de langue étrangère, un second quart entre 2 et 16.5, un troisième quart entre 16.5 et 135 et un dernier quart un nombre supérieur à 135 éléments. Il semble donc possible d'en inférer qu'un quart du corpus décrit un objet linguistique de langue étrangère.

On voit quoi qu'il en soit la nécessité d'étiqueter les éléments de langue étrangère dans les textes scientifiques et *a fortiori* de linguistique, afin de limiter les erreurs d'étiquetage[22].

Corrélés comme il l'a été vu *supra* avec les guillemets simples, les éléments étrangers s'opposent d'abord et de manière logique aux déterminants et aux noms, avec lesquels ils entrent en concurrence claire :

> En effet, *up* n'est pas une préposition dans l'énoncé *He looked up the word in the dictionary*, alors que *at* appartient à cette catégorie syntaxique aussi bien dans *He was looking at the painting* que dans *He stayed at the hotel*. (187)

On remarque en effet que les éléments de langue étrangère ont un fonctionnement proche du nom propre : ils ne sont jamais précédés d'un déterminant bien qu'ils aient un fonctionnement de nom. Il n'est donc pas surprenant qu'ils s'opposent aux déterminants et à l'ensemble des noms.

3.4.6.2. Eléments les plus représentés

La langue anglaise est de loin la plus représentée : si l'on extrait les éléments étiquetés 'FGW' par ordre de fréquence, on obtient d'abord les mots grammaticaux de l'anglais : *the* (642 occ.), *a* (545), *of* (371), *to* (325), *in* (292), *and* (250), *is* (199), etc.

Hormis l'anglais, les langues espagnole et allemande et l'ancien français sont également très présentes, bien que de manière moindre : on relève ainsi 134 *der*, 134 *die* et 101 *ein*. On

[22] Dans le meilleur des cas sont-ils étiquetés <élément_inconnu>, catégorie composite et naturellement non interprétable.

relève en outre une représentation significative du chinois, liée aux textes 86 et 103 dont il constitue l'objet et au numéro sur les universaux linguistiques, de même qu'une proportion non négligeable d'éléments de langue italienne. Les éléments relevant d'autres langues sont globalement des hapax (une à deux apparition max. dans l'ensemble du corpus).

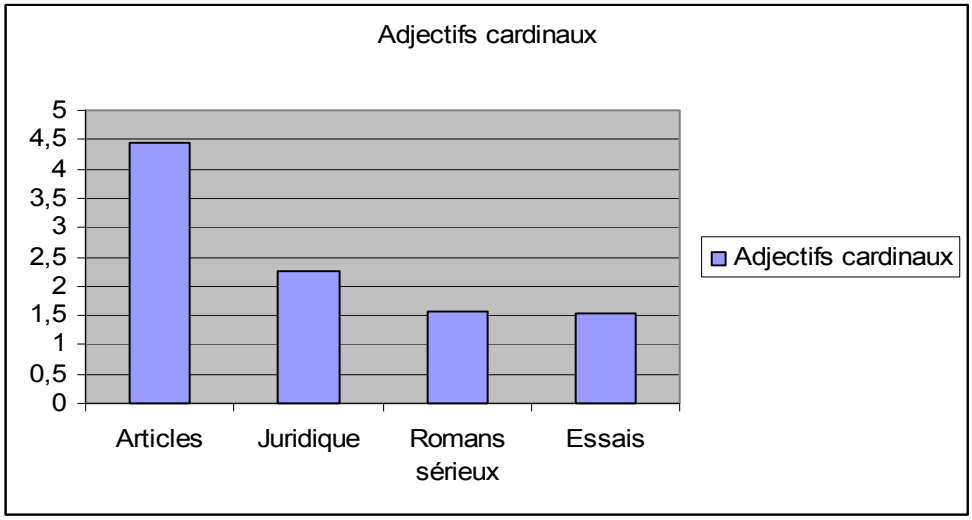
De manière générale, on a pu remarquer que l'étiqueteur entraîné obtenait de très bons résultats sur l'anglais et l'allemand, tandis que l'ancien français et l'espagnol occasionnaient de nombreuses erreurs étant donné les nombreuses formes grammaticales partagées avec le français actuel. Les autres langues sont nettement moins bien traitées, notamment lorsqu'elles ne sont pas prises en charge par la table de codage Latin 1.

Ces résultats intéressent particulièrement le champ de la linguistique française, qui semble plus intéressé par la langue (française) que par les langues, qui se limitent d'ailleurs bien souvent aux langues occidentales européennes.

3.4.7. Des numéraux

3.4.7.1. Genre de l'article vs. genres autres

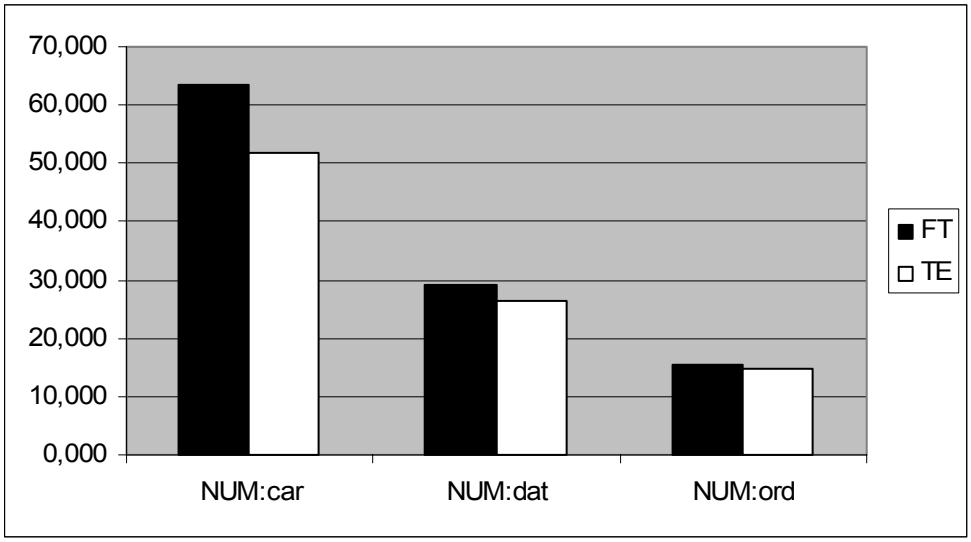
De manière non surprenante, le genre de l'article contient un nombre beaucoup plus important de chiffres que les trois autres genres considérés :



Graphique : Répartition en pourcentage[23] des cardinaux par genre

[23] Relativement à l'ensemble des adjectifs.

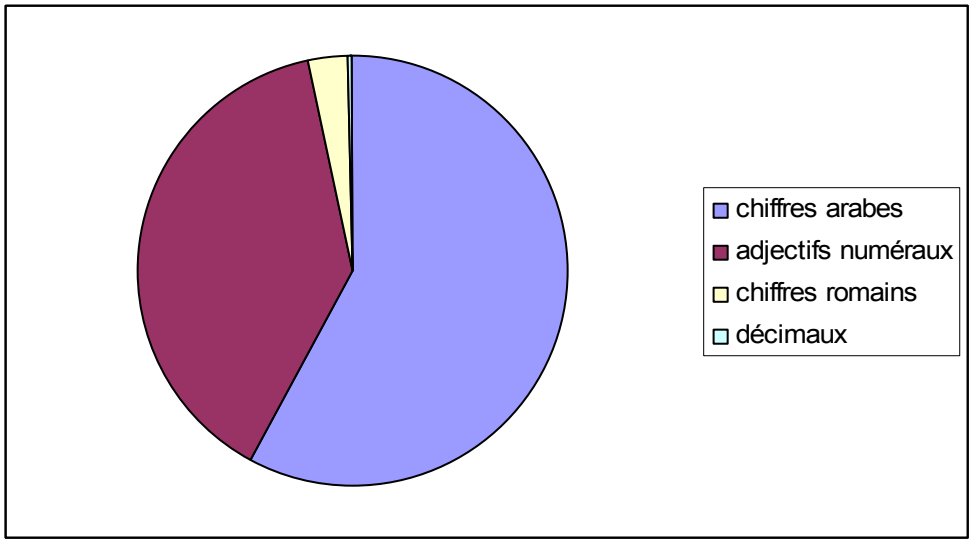
3.4.7.2. Répartition et corrélations des numéraux



Graphique : Répartition des numéraux avec/sans exemples (moyennes absolues par texte)

A. Numéraux cardinaux

Les cardinaux sont d'abord les numéraux les plus représentés dans l'ensemble du corpus. Rappelons que nous n'avions pas différencié les adjectifs numéraux des chiffres arabes ou romains. Le graphique suivant propose donc une quantification plus fine des éléments constitutifs de la catégorie cardinaux, effectuée après extraction de l'ensemble de ses éléments :

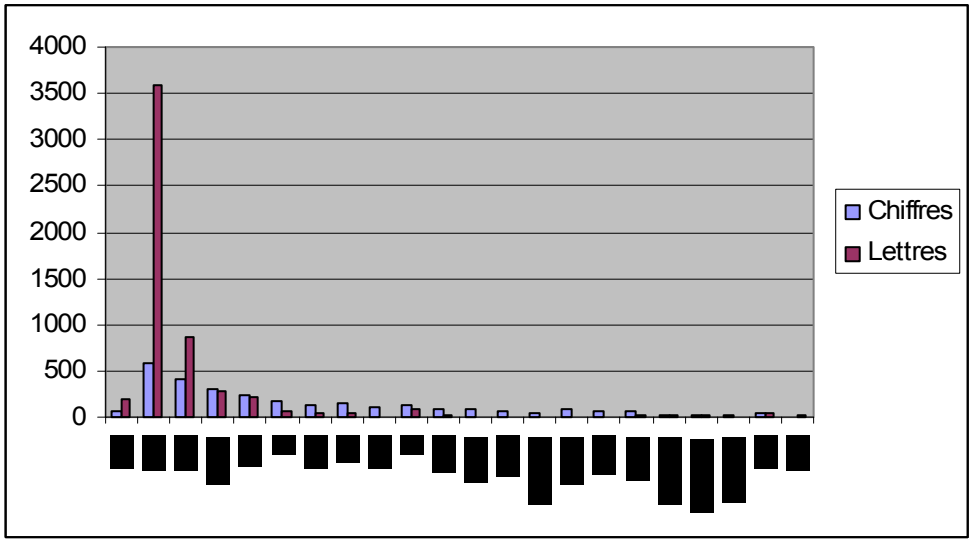


Graphique : Représentation en pourcentage des catégories de numéraux cardinaux

On notera d'abord la proportion extrêmement faible des numéraux décimaux, qui nous semble particulièrement illustrer la non affiliation de la discipline linguistique aux sciences exactes, les chiffres décimaux résultant systématiquement d'un calcul mathématique.

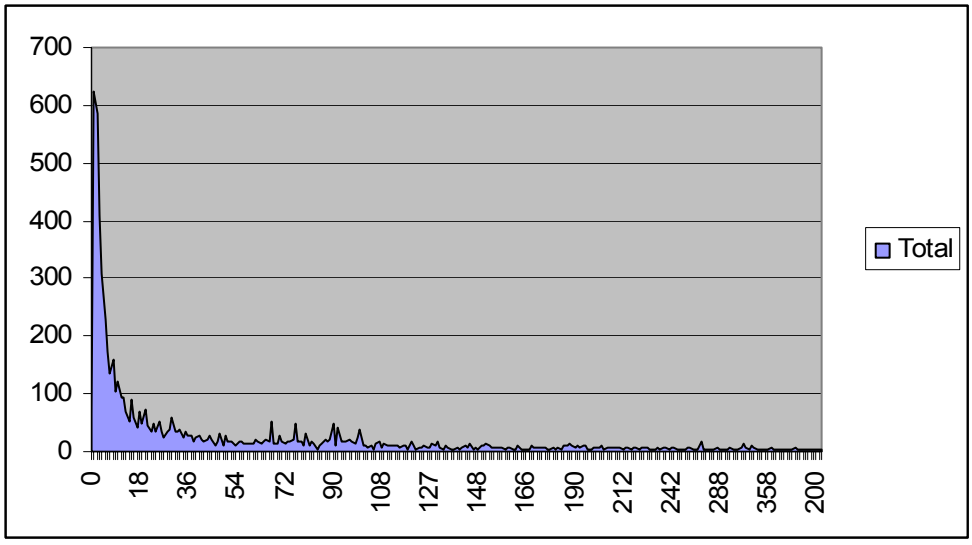
En ce qui concerne les formes lettres et chiffres des numéraux, seuls 22 nombres sont représentés en lettres, ce qui est notable si l'on considère que les lettres représentent 40% des cardinaux – et encore doit-on souligner que le nombre *un* n'a pas pu être pris en compte, la distinction ne pouvant être effectuée entre les déterminants et les formes chiffrées.

Comme l'illustre le graphique suivant, c'est le nombre *deux* qui est le plus représenté : il représente 64.58 des formes orthographiées des chiffres, ce qui est considérable. *Trois* représente 15.53% des formes et avec *zéro, deux* et *trois* sont les seuls nombres prioritairement représentés en lettres.



Graphique : Formes lettres et chiffres des nombres (chiffres absolus)

Si l'on examine maintenant les formes chiffrées les plus représentées, on s'aperçoit que les chiffres inférieurs à 100 représentent 74.22% de l'ensemble des cardinaux et que les chiffres inférieurs à 10 en représentent 39.65%, ce qui est considérable :



Graphique : Représentation des chiffres de fréquence supérieure à 4 dans l'ensemble du corpus (chiffres absolus)

La linguistique française semble ainsi faiblement intéressée par les données chiffrées ; on peut supposer que les chiffres inférieurs à cinq ne relèvent pas du calcul mathématique, ni d'un travail quelconque de quantification.

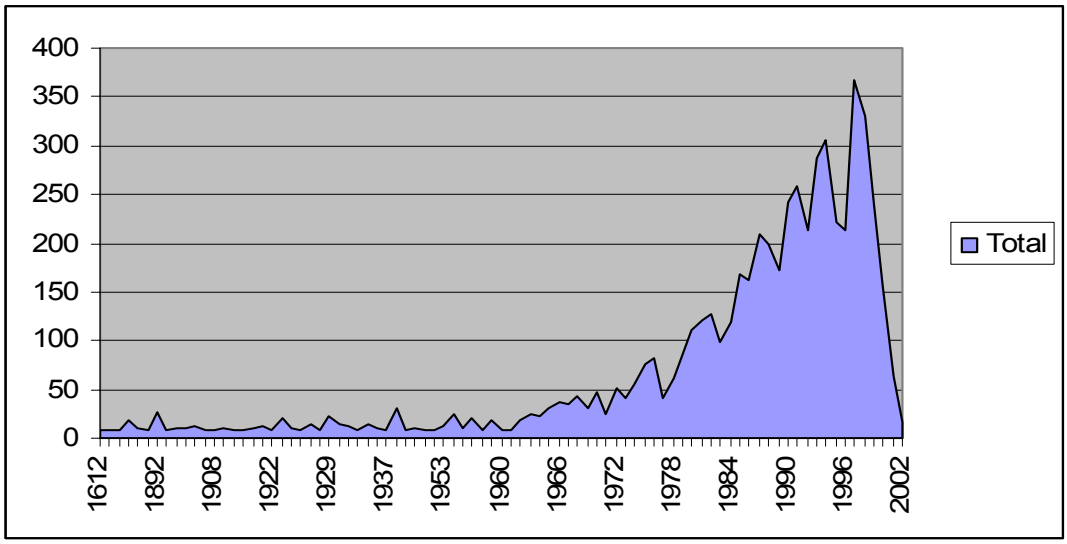
B. Dates

Il va de soi qu'une analyse des dates devrait pour être pertinente s'accompagner d'une étude des références bibliographiques, voire des événements les accompagnant, d'autant que

les dates sont très fortement corrélées aux éléments constitutifs des références citées, généralement soumises au format (Auteur, date) : les dates sont ainsi corrélées aux noms propres (+0.39 FT/+0.55 TE), aux parenthèses (+0.16 FT/+0.27) et aux abréviations (+0.28 FT/+0.26 TE), ces dernières étant essentiellement associées aux références (cf. *infra*).

Nous ne disposons malheureusement pas de ces données bibliographiques, bien trop coûteuses à rassembler relativement aux objectifs de notre étude.

Ces réserves exprimées, examinons le graphique ci-dessous, qui par souci de lisibilité ne présente que les dates mentionnées plus de sept fois dans l'ensemble du corpus – les autres références datées ne représentant que 10.5% du total obtenu :



Graphique : Répartition des dates apparaissant plus de sept fois dans le corpus (chiffres absolus)

Si la courbe observée est relativement stabilisée entre 1612 et 1960, on observe une croissance significative des dates de 1960 à 2000, année clé de voûte de constitution du corpus. La majorité des références datées se situe sur un empan de 20 ans, entre 1980 et 2000. Ce phénomène nous semble lié à la nécessité pour le chercheur de positionner sa recherche dans son champ scientifique d'investigation, d'abord actuel afin de s'intégrer ou de s'imposer dans sa communauté sociologique de rattachement, puis plus éloigné sur le plan temporel, afin de mentionner les travaux sinon fondateurs, du moins nodaux de son domaine scientifique.

La décroissance observée de 2000 à 1960 est irrégulière, certaines dates semblant plus fécondes, ou du moins plus notables, que d'autres : outre l'année 1997, qui représente le point culminant du graphique, mentionnons 1994, 1991 et 1987. Comme nous l'avons déjà souligné, il serait pertinent d'examiner les références bibliographiques auxquelles se rapportent ces dates, afin de déterminer les éventuelles autorités du corpus[24], ce qui participerait grandement à notre connaissance du champ scientifique linguistique français.

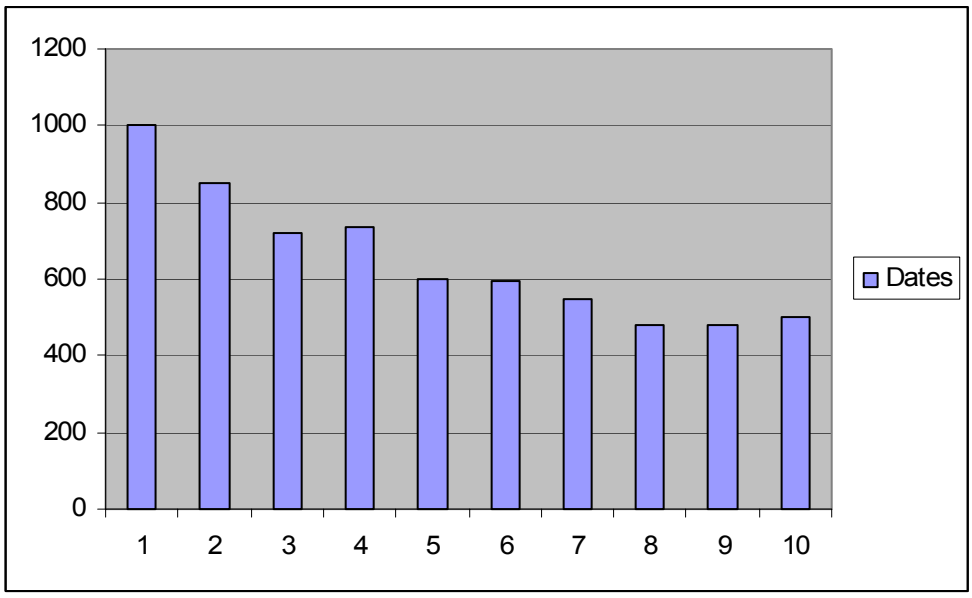
Les dates sont de manière générale associées à une dimension historico-narrative, qui émerge très nettement après extraction des exemples : elles sont ainsi fortement corrélées avec le passé simple (+0.22), les auxiliaires au passé simple (+0.21), le plus-que-parfait (+0.21) et

[24] Une date ne correspondant pas nécessairement à une référence bibliographique, une telle initiative serait appréciable.

l'imparfait (+0.20). On relève également plusieurs associations de la variable avec les marqueurs de troisième personne, indices de thématisation d'un objet ou d'une personne.

Les dates sont ainsi plus représentées dans les textes plus narratifs, qui, au regard des corrélations négatives de la variables, seraient moins structurés que l'ensemble des textes (-0.37 avec les indices de structuration ; -0.24 avec les indices de renvoi), contiendraient moins de ponctuations deux points (-0.26), de verbes au présent (-0.19), de connecteurs de conséquence (-0.19) et de verbes conjugués au futur (-0.16).

Enfin, soulignons qu'on observe une décroissance significative des dates au fil du texte : elles sont davantage corrélées au début d'article, dans lequel le chercheur annonce son (ses) hypothèse(s) scientifique(s) :



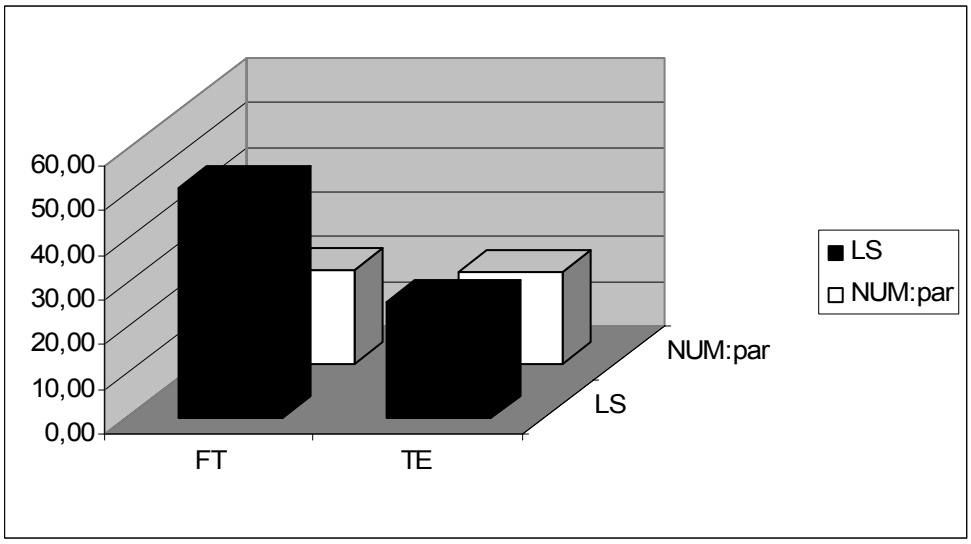
Graphique : Configuration tactique des dates

C. Ordinaux

Les ordinaux constituent enfin une catégorie finalement composite, qui rassemble les mentions aux siècles passés (e.g. *XVIIe/dix-huitième* siècle) et les adjectifs numéraux ordinaux (*première, deuxième*, etc.). L'importance de ces derniers associe l'ensemble des numéraux aux caractéristiques de l'énumération : on note ainsi une corrélation forte du descripteur aux connecteurs temporels (+0.26 FT/+0.25 TE) et aux virgules (+0.26 FT/ +0.22 TE).

Si l'analyse des adjectifs ordinaux les plus représentés présente globalement peu d'intérêt – *premier(ère)* représente ainsi plus d'un tiers des occurrences totales, il peut être intéressant de mentionner que le XIXe siècle est le siècle le plus représenté (33 occ.), bien avant les XVIIe et XVIIIe siècle.

3.4.8. Des indices de structuration textuelle et de renvoi dans le texte



Graphique : Répartition des marqueurs de structuration textuelle et de renvoi dans le texte (moyennes absolues par texte)

Bien que les indices de renvoi dans le texte ne baissent quasiment pas après extraction des exemples, on observe une diminution considérable des marqueurs de liste, liée au fait que les exemples de linguistique sont la plupart du temps numérotés.

Comme il l'a déjà été évoqué *supra*, les indices de structuration et de renvoi s'opposent au niveau textuel aux textes comprenant une dimension historico-narrative, ce qui nous semble plus précisément relever d'une opposition des textes historiques aux textes exemplifiés. La présence des deux points paraît en effet liée aux exemples, habituellement délimités dans le texte et généralement annoncés par deux points :

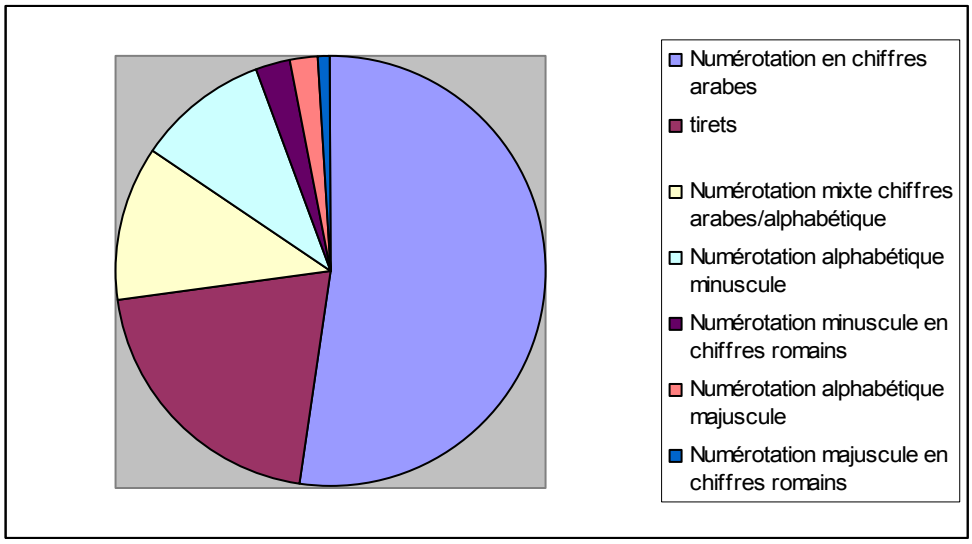
> Nous contrastons pour l'italien les restructurants optionnels (*volere*) et les non-restructurants (*desiderare*) :
>
> *(2)* *I* *a.* Giovanni vuole / desidera mangiarle.
>
> *a'.* Giovanni le vuole / * desidera mangiare. (108)

3.4.8.1. Indices de structuration textuelle

A. Listes

Outre les titres, que l'on traitera à part, on a pu relever six formats de liste principaux dans le corpus[25] :

[25] Contrairement aux articles de revue anglo-saxonne, le format des listes est diversement normé selon les revues : certaines revues préconisent un format particulier, tandis que le choix est libre dans d'autres, ce qui explique la diversité des formats relevés.



Graphique : Répartition des formats de liste observés dans le genre de l'article

Notons d'abord la nature ambiguë des listes numérotées en chiffres arabes qui renvoient aussi bien aux titres de niveau 1 de l'article qu'aux listes intégrées en son corps (liste de termes, de concepts, de modalités d'analyse, etc.).

S'il est impossible de dénombrer les énumérations introduites par un tiret, on remarque que les chiffres arabes sont utilisés pour numéroter les listes les plus longues – et essentiellement les listes d'exemples. Les formats *-1, -2, … n* et *1., 2., … n*, qui représentent 85% des numérotations en chiffres arabes, agencent des relevés qui peuvent comprendre jusqu'à 149 éléments, tandis que les numérotations alphabétiques minuscules organisent des listes de 9 éléments au plus, 6 pour les alphabétiques majuscules, 8 pour les minuscules en chiffres romains et 7 pour les majuscules en chiffres romains.

Le tableau ci-dessous rassemble les formats d'encodage des listes relevés au sein du corpus – en gris ont été surlignés les formats privilégiés de chaque type de numérotation :

Numérotation	Type		Numérotation	Total relevé
	Format	Exemple		
Chiffres arabes	**-N**	-1, -2, …, N	[1-108]	**2956**
	N.	1., 2., …, N	[0-149]	**1456**
	N	1, 2, …, N	[0-12]	**226**
	(N)	(1), (2)… N	[1-35]	**170**
	N)	1), 2), … N	[1-44]	**140**
	N/	1/, 2/, … N	[1-8]	**120**
	N-	1-, 2-, … N	[1-8]	**64**
	N°	1°, 2°, … N	[1-4]	**28**
Alpha minuscule	**[a-z].**	a., b., … [a-z]	[a-i]	**392**
	[a-z])	a), b), … [a-z])	[a-e]	**291**
	([a-z])	(a), (b), …([a-z])	[a-h]	**209**
	[a-z]	a, b, … [a-z]	[a-h]	**59**
	[a-z]-	a-, b-, … [a-z]-	[a-i]	**30**
	[a-z]/	a/, b/, … [a-z]/	[a-b]	**6**
Alpha majuscule	**[A-Z])**	A), B), … [A-Z])	[A-F]	**65**
	[A-Z]-	A-, B-, … [A-Z]-	[A-D]	**59**
	[A-Z].	A., B., … [A-Z].	[A-D]	**42**
	([A-Z])	(A), (B), …([A-Z])	[A-C]	**26**
	[A-Z]	A, B, … [A-Z]	[A-C]	**14**
	[A-Z]/	A/, B/, … [A-Z]/	[A-C]	**3**
Rom. minuscule	**(i+)**	(i), (ii), … (i+)	[1-8]	**215**
	i+)	i), ii), … i+)	[1-4]	**24**
	-i+-	-i, -ii, … -i+	[1-5]	**8**
	[i+]	i, ii, … [i+]	[1-3]	**3**
Rom. majucule	**I+.**	I., II., …, I+.	[1-7]	**52**
	I+)	I), II), … I+)	[1-6]	**16**
	I+	I, II, …, I+	[1-5]	**15**
	(I+)	(I), (II), … (I+)	[1-5]	**9**

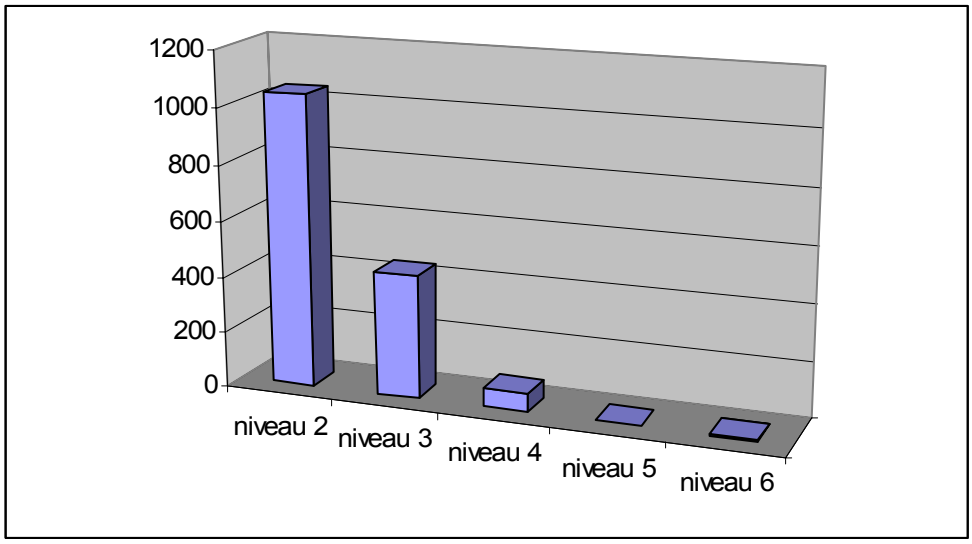
Tableau : Formats d'encodage des listes relevées au sein du corpus

On voit qu'à chaque type de numérotation est associé un indice typographique privilégié, ce qui stabilise la sémiotique du texte : le tiret semble ainsi dévolu aux chiffres, les parenthèses aux chiffres romains en minuscules et le point aux chiffres romains en majuscules. Seules les numérotations alphabétiques minuscules et majuscules sont associées à plusieurs typographies.

Quant aux numérotations mixtes, elles sont globalement associées aux exemples des textes, numérotées la plupart du temps (1a) ou (1'a).

B. Titres

Comme on l'a déjà mentionné, les titres de niveau 1 n'ont pas pu être séparés des simples listes. Il est en revanche possible d'analyser les titres de niveau supérieur de format 1.2. ou 1.2.3., format au demeurant le plus employé dans le genre de l'article :



Graphique : répartition des titres de niveau supérieur à 1 dans le genre de l'article (FT)

Nous n'avons relevé que deux occurrences de titres de niveaux 5 et 6, et aucune de niveau supérieur : le genre de l'article étant relativement court, une structuration excessivement complexe dessert en effet considérablement sa lecture, d'autant que le format de division pris en compte devient quasi indéchiffrable au-delà du niveau 3 (e.g. *2.1.2.1.1.*). Les divisions de niveau supérieur à trois sont ainsi plus volontiers introduites par un autre type de numérotation (alphabétique en général).

Si l'on ne prend en compte que le format observé, le genre de l'article contiendrait deux à trois grandes sections numérotées, et deux à trois niveaux de structuration en moyenne.

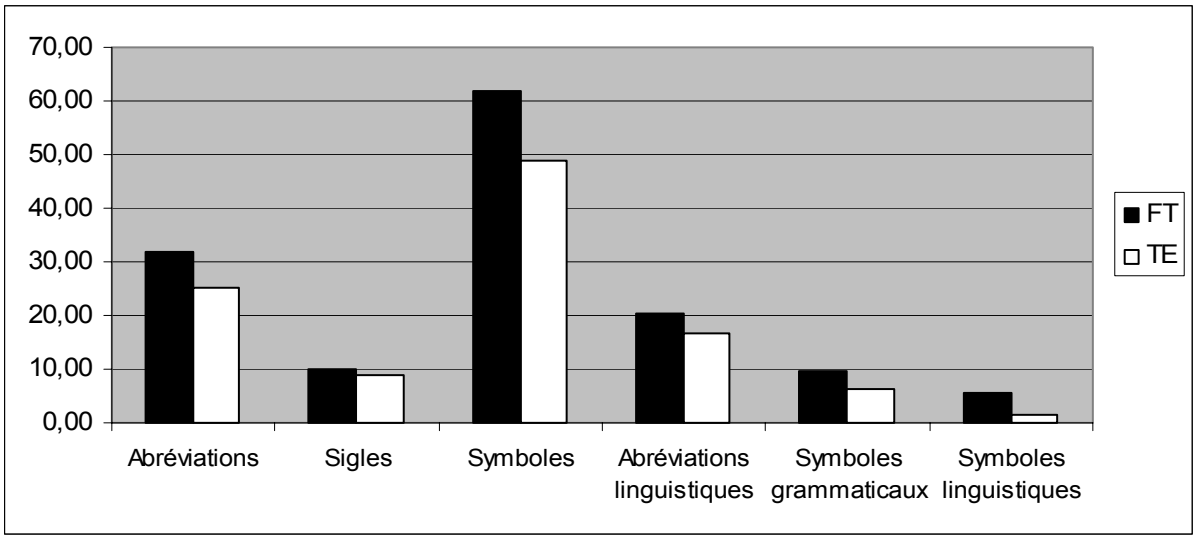
Fait notable, les parties les plus structurées des textes sont de loin les deuxièmes parties, qui contiennent au moins deux divisions dans plus de la moitié des textes, et qui sont quantitativement les premières à comporter des niveaux de structurations de profondeur trois à quatre. L'article de linguistique n'étant pas soumis à la structure IMRAD, le contenu de cette deuxième partie d'article varie très probablement selon les textes, ce qui limite l'interprétation. On tentera toutefois d'éclaircir ce phénomène à la lumière d'une analyse de la structure des textes effectuée sur un échantillon du corpus présentée *infra*.

C. Indices de renvoi

Les indices de renvois sont à quelques exceptions près exclusivement chiffrés ou mixtes (chiffres arabes/alphabétique), et concernent essentiellement les exemples des textes.

On ne relève curieusement aucune corrélation des indices de structuration avec les indices de renvoi, et les deux descripteurs ont au demeurant des corrélations bien distinctes. Les indices de renvoi sont ainsi fortement corrélés aux sigles linguistiques (+0.31), ce qui indiquerait un emploi privilégié des renvois dans certains sous-domaines linguistiques, plus intéressés par la syntaxe, et mobilisant probablement un nombre plus important d'exemples.

3.4.9. Des marques de formalisation



Graphique : Répartition des marques de formalisation (moyennes absolues par texte)

3.4.9.1. Symboles

Les symboles sont d'abord les éléments les plus représentés et les plus stables du genre de l'article (CV de 1.32 FT/1.44 TE). Le symbole le plus répandu est sans conteste 'X' qui représente près de 10% des symboles relevés dans le corpus, et près de 12% si l'on y ajoute sa version minuscule 'x'.Son homologue 'Y' est beaucoup moins présent (1.82% de l'ensemble des symboles observés et moins de 3% si l'on totalise l'ensemble des 'Y' et 'y' représentés) : l'association relationnelle X/Y ne semble donc correspondre qu'au quart des emplois de X.

Si X désigne en mathématiques l'inconnue d'une équation depuis Descartes, il semble avoir un statut plus générique en linguistique (X = tout élément) comme dans :

> elle correspond à l'explication de « Ce texte parle de X » par « Ce texte contient des expressions qui traduisent la notion de X ». (006)

Les opérateurs mathématiques '+' et '=' sont ensuite très représentés. On notera toutefois que '+' (il en va de même pour '-') n'additionne jamais des données mathématiques, mais des mots ou des concepts comme :

> langue = (rapports associatifs) + (syntaxe)
>
> discours = (rapports syntagmatiques) - (syntaxe) (001)

'+' et '-' sont également très présents dans les textes contenant des traits sémantiques comme :

> par exemple *guimbarde* implique [-mammifère], mais ce trait n'apparaît pas dans le signifié. En revanche, *guimbarde* possède le trait [+péjoratif], qui l'oppose à *voiture*, et qui bien évidemment n'appartient pas au référent. (008)

De manière générale, les symboles relevés sont d'ordre plus logique que mathématique, ce qui est n'est pas surprenant, la logique représentant encore aujourd'hui un outil majeur en grammaire et en linguistique.

On relève ainsi de nombreux symboles propres à la théorie des ensembles, ou à la logique des prédicats du premier ordre: Ø, →, *, p et q, etc.

3.4.9.2. Sigles

Les sigles et acronymes varient très peu après extraction des exemples : ils relèvent quasi tous des corps d'articles et sont généralement spécifiques à un texte, ou à un ensemble de textes se rapportant à un même objet (une revue thématique le plus souvent) : ainsi relève-t-on un emploi considérable des sigles LSF et LS dans les textes appartenant au numéro de Langue française sur la langue des signes. Les sigles les plus représentés dans l'ensemble du corpus renvoient naturellement au connu : ils renvoient à des institutions publiques notoires, comme la SNCF, l'UMR ou le CNRS, ou à des objets de nature institutionnelles comme le JT par exemple. On retrouve bien sûr de nombreux sigles se rapportant à l'enseignement, comme le DEUG ou le CM2. On notera enfin la nature quasi institutionnelle du TLF (*Trésor de la Langue Française*), qui fait partie des dix premiers sigles les plus représentés.

3.4.9.3. Abréviations

Les abréviations varient également peu après extraction : elles semblent donc relever davantage du corps de l'article que de ses exemples. Les abréviations forment une catégorie particulièrement hétérogène, et par conséquent complexe en termes d'interprétation. Elle permet néanmoins d'obtenir les éléments les plus abrégés du genre. Voici donc par ordre de fréquence les dix abréviations les plus utilisées, qui représentent 53.36% de l'ensemble du relevé : *p.* (page)*, cf., etc., ex.* (exemple)*, vs., al., M., J., G., P.* et *A.* Les cinq derniers éléments correspondent visiblement à des abréviations de prénoms, usuelles lors de la mention de références. On notera que la plupart des abréviations effectuées sont liées aux références : mention de pages, prénoms des auteurs, mention *et al.* – on remarque d'ailleurs que *et al.* est cinq fois plus utilisée que sa version longue *et alii*. – de même que les introductifs *cf.*, *i.e.* ou *v.* (pour voir), ainsi que *id., ibid., op.* ou *cit.*, également très représentés, ce qui explique les corrélations importantes des abréviations aux indices de références.

3.4.9.4. Formalisation linguistique

Force est d'abord de constater que les symboles grammaticaux éclairent peu notre connaissance du genre et du champ scientifique linguistique : leur présence dépend nettement de l'objet de l'article et varie donc largement selon les textes (CV de 3.42).

L'analyse des formalisations linguistiques (sigles linguistiques), plus représentées et plus stables (CV de 2.10) est comparativement plus significative : ainsi, les dix formalisations suivantes représentent 64.97% de l'ensemble des abréviations linguistiques relevées, ce qui est considérable : N (675 occ.), SN (571), V (286), N1 (225), RST (167), N2 (151), COD (98), MI (97), GN (96) et pp. (86). Ces éléments sont d'ordre syntaxique et sont essentiellement dévolus au groupe nominal : les abréviations N et SN représentent ainsi 33% de l'ensemble des éléments relevés. Hormis RST (pour *Rhetorical Structure Theory*), MI (pour *Monologue Intérieur*) et pp. (pour *participe présent*), l'ensemble des abréviations recensées est relativement bien répartie dans l'ensemble des textes.

Les symboles d'acceptabilité linguistique sont en principe caractéristiques des articles se rapportant au normatif et aux notions de bonne formation et de grammaticalité. On notera que près de la moitié des textes du corpus n'en contiennent aucun, soit un tiers des numéros de revues étudiés. Les textes et les revues qui en comportent le plus grand nombre relèvent visiblement du sous-domaine sémantique, et plus précisément d'une sous-branche de la sémantique intéressée par la description d'un marqueur (morphème, « mot du discours », etc.). Ces articles recourent aux symboles linguistiques pour valider l'acceptabilité sémantique d'un énoncé ou d'une séquence comme dans :

(5a) Henri a (une attitude + un comportement) de laxiste.

?Henri a pour (attitude + comportement) d'être laxiste.

Henri a comme (attitude + comportement) d'être laxiste.

(5'a) *Henri a son (attitude + comportement) dans le laxisme.

*Henri a pour (attitude + comportement) d'être dans le laxisme.

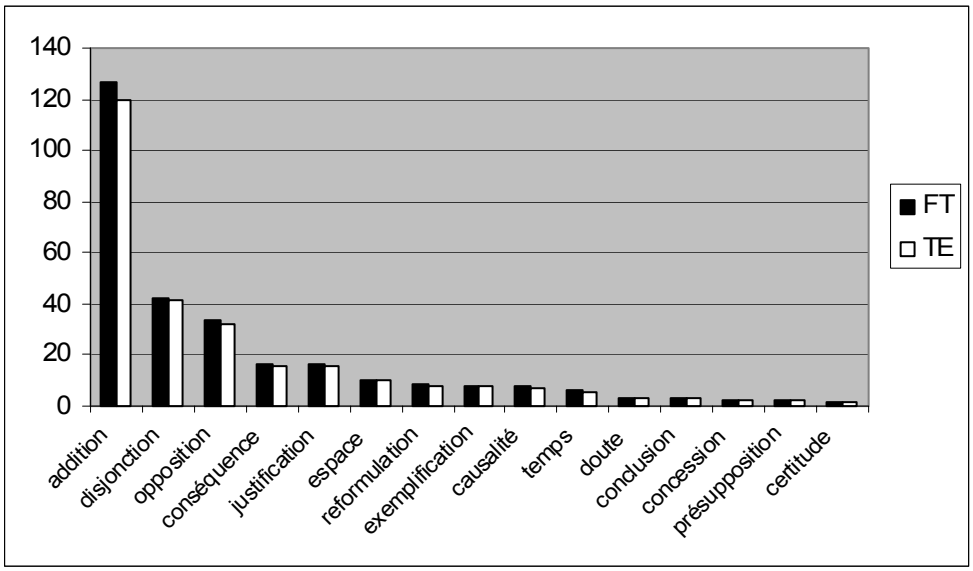
*Henri a comme (attitude + comportement) d'être dans le laxisme. (033)

Cette corrélation des symboles d'acceptabilité au sémantique entraîne naturellement un nombre beaucoup plus important desdits marqueurs dans les revues, numéros thématiques ou articles consacrés à l'étude d'un objet sémantique. Ainsi, seuls quatre des 23 articles de la *Revue de Sémantique et Pragmatique* dont nous disposons ne contiennent aucun symbole de bonne formation linguistique, tandis que les textes de la revue en comprennent en moyenne 21.8 par texte[26], ce qui est considérable. On relève également une moyenne de 26.25 marqueurs par article du numéro thématique *Du sens au sens* (*LINX*), 21.2 pour *Topicalisation et partition* (*Cahiers de Praxématique*) et 14.8 pour *Problèmes de classement des unités lexicales* (*Cahiers du CIEL*).

Ce descripteur nous permet donc déjà de mettre au jour une première opposition interne du corpus et de contraster les textes de sémantique grammaticale aux textes et aux revues historiques comme l'ensemble des articles *d'HEL* ou peut-être plus théoriques comme le numéro thématique *Contexte(s)* de *Scolia* qui n'en contiennent aucun.

3.4.10. Des connecteurs

Malgré ses insuffisances, la typologie de connecteurs que nous avons élaborée chapitre 2 peut nous fournir des indications éclairantes quant au genre de l'article :

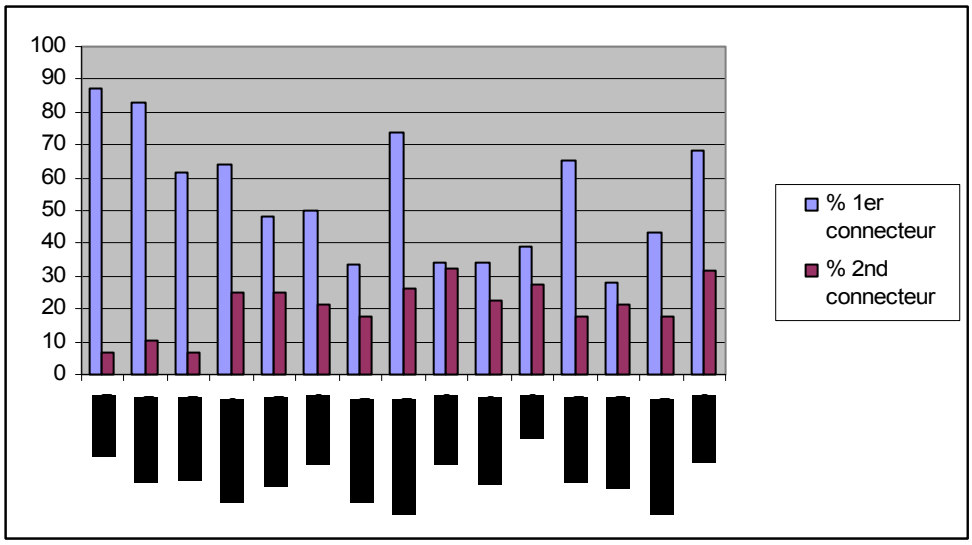


Graphique : Répartition des connecteurs avec/sans exemples (moyennes absolues par texte)

On observe d'abord peu de variation des connecteurs après extraction des exemples : l'ensemble des connecteurs semble ainsi faire partie du corps de l'article.

[26] Moyenne évidemment calculée sur les textes *contenant* ces symboles.

Toutes les classes présentent de manière constante deux connecteurs qui représentent en moyenne 74.85% de chaque classe :



Graphique : Proportions tenues par les deux premiers connecteurs de chaque classe

Voici d'ailleurs, à titre indicatif, les deux éléments les plus représentés de chaque classe :

Type de connecteur	Premier connecteur	Second connecteur
Addition	et	aussi
Disjonction	ou	soit
Opposition	mais	cependant
Conséquence	donc	alors
Justification	ainsi	en effet
Spatialité	ici	là
Reformulation	c'est-à-dire	en particulier
Exemplification	par exemple	notamment
Causalité	puisque	parce que
Temporalité	d'abord	ensuite
Doute	sans doute	peut-être
Conclusion	enfin	finalement
Concession	certes	bien sûr
Présupposition	a priori	apparemment
Certitude	tout à fait	certainement

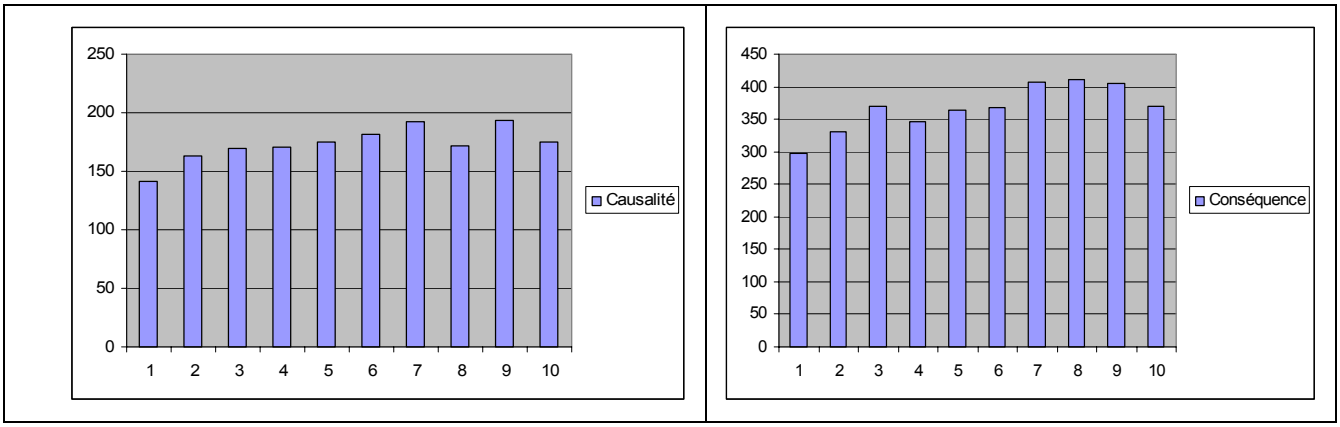
Tableau : Présentation des deux connecteurs les plus représentés de chaque classe

Ainsi, les connecteurs d'addition et de disjonction sont les plus représentés du genre de l'article, eu égard au nombre naturellement massif de 'et' et de 'ou' qui représentent respectivement 87.22 et 82.74% de leurs classes d'appartenance. Outre l'énumératif 'et', on observe un nombre significatif de connecteurs *d'une (d'autre) part*, qui présentent l'intérêt d'organiser la segmentation et la progression textuelle, contrairement aux simples énumératifs *aussi* ou *ainsi que*. Le plus stylé *de surcroît* (0.06% de l'ensemble du relevé) et la locution *non seulement ... mais encore* (0.25%) sont les connecteurs les plus marginaux du relevé.

La nature polémique ou argumentative de l'article et du discours scientifique en général transparaît particulièrement avec la proportion importante de connecteurs argumentatifs relevés : les connecteurs argumentatifs d'opposition détiennent en effet des proportions remarquables, qui sont à 61.64% liées à la présence du connecteur *mais*. On remarque un

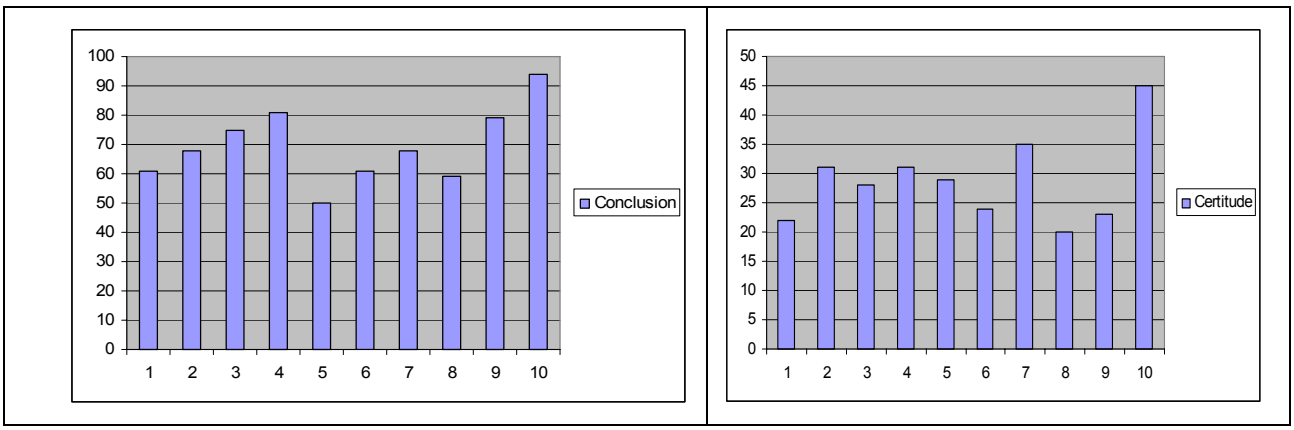
usage plus important (> 4% de l'ensemble du relevé) des oppositifs/restrictifs *cependant, or, en revanche* et *toutefois*, tandis que *bien que, malgré* et *outre* sont globalement peu représentés (>2%). Soulignons l'emploi très marginal de *a contrario*, dont on ne relève que 5 occurrences. De manière non surprenante, les connecteurs d'opposition sont corrélés aux négations (+0.25 TE) et aux points d'exclamation (+0.39), indices d'une dimension plus polémique. On relève également une corrélation significative avec les plus hypothétiques modaux au conditionnel (+0.24).

On notera que les relations de conséquence paraissent bien plus marquées que celles de causalité : *donc, alors* et *par conséquent* sont deux fois plus représentés que *puisque, parce que* ou *car*, phénomène qui sera à préciser lors de l'analyse ultérieure de la séquentialité du genre. Les connecteurs de conséquence et de causalité n'ont pas les mêmes corrélats, pas plus qu'ils ne sont corrélés entre eux : les connecteurs de conséquence semblent employés dans les textes plus exemplifiés (associations positives avec les impératifs, le pronom « *on* », les indices de renvoi… et négatives avec les dates, noms propres et abréviations), tandis que les connecteurs de causalité sont positivement corrélés aux noms propres et aux abréviations. Bien que les deux marqueurs soient tous deux associés aux négations, le connecteur de causalité relève peut-être davantage d'une dimension polémique, comme le démontre sa corrélation avec le connecteur d'opposition. On observe en revanche que les deux connecteurs ont des configurations tactiques ascendantes très similaires, avec un maximum atteint en fin d'article :



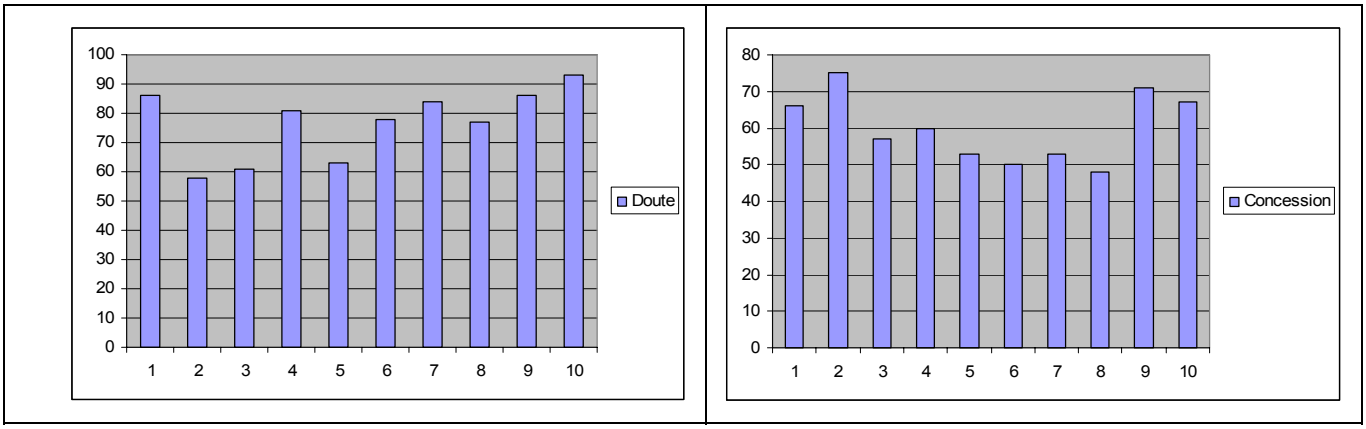
Graphique : Configurations tactiques des connecteurs de causalité et de conséquence

Les relations logiques s'intensifieraient donc en fin d'article, pour soutenir l'émergence de la conclusion finale et la résolution de l'hypothèse de départ, exaltées par les connecteurs de conclusion et de certitude :



Graphique : Configurations tactiques des connecteurs de conclusion et de certitude

et modérées par les connecteurs de doute et de concession, qui sont d'ailleurs statistiquement corrélés :



Graphique : Configurations tactiques des connecteurs de doute et de concession

3.5. A la recherche d'axes d'organisation du genre

Si les procédés classiques de la statistique descriptive adoptés précédemment nous ont permis de résumer l'information recueillie sur chaque descripteur, appréhendé isolément ou au sein de sa classe d'appartenance, elles demeurent insuffisantes : nous ne disposons en effet pour l'heure que d'une description relativement fragmentée du genre, bien que nous ayons rendu compte des corrélations textuelles des variables entre elles. De ce fait, les interrelations entre variables ne nous sont pas complètement inconnues, et on s'attend déjà à retrouver certaines régularités observées *supra*.

En revanche, nous ne sommes pas en mesure d'esquisser une véritable *structure* du genre. C'est là tout l'intérêt des méthodes factorielles, et plus précisément de l'Analyse en Composantes Principales (ACP), consacrée aux tableaux numériques comme le nôtre, dont le but est de « révéler ces interrelations entre caractères et de proposer une description de la population apte à suggérer une structure » (Auray, 1990).

En proposant un résumé descriptif de l'ensemble des observations effectuées sur les 135 descripteurs adoptés, l'ACP nous permettra ainsi :

- d'examiner les relations entre les textes et de repérer les groupes d'individus homogènes, ainsi que les individus au comportement atypique ;
- de construire un ensemble de variables artificielles « expliquant » l'ensemble des descripteurs pris en compte : ces variables permettent une réduction du tableau de données originel puisqu'au prix d'une perte d'information, il est possible de remplacer les 136 variables de départ par un ensemble beaucoup plus réduit de variables statistiques artificielles, *i.e.* les facteurs.

Le corpus étant très homogène, on présume que les structures obtenues seront évidemment moins prononcées, et probablement moins claires, que dans l'étude menée par Biber (1988) ; elles nous permettront néanmoins de faire émerger les champs de contraste principaux du genre de l'article.

L'ACP sera complétée d'une Classification Ascendante Hiérarchique, qui permettra de préciser les relations entre individus et de nuancer les résultats de l'ACP[27].

3.5.1. Analyse en Composantes Principales

Eu égard aux résultats obtenus dans la section précédente, où les corrélations étaient beaucoup plus significatives et plus stables dans les textes dont les exemples ont été enlevés, l'ACP a été menée sur ce dernier corpus.

3.5.1.1. Diagramme des valeurs propres

Nb	Valeur propre	% d' inertie	% cumulé	
1	10.90	7.79	7.79	**********
2	6.46	4.62	12.41	**********
3	5.40	3.86	16.27	**********
4	5.20	3.72	19.99	**********
5	4.09	2.93	22.92	**********
6	4.01	2.87	25.79	**********
7	3.67	2.63	28.41	**********
8	3.35	2.40	30.81	**********
9	3.33	2.38	33.19	**********
10	3.08	2.20	35.40	**********
11	2.88	2.06	37.46	**********
12	2.69	1.92	39.38	**********
13	2.66	1.91	41.29	**********
14	2.58	1.85	43.14	**********
15	2.44	1.75	44.88	**********
16	2.34	1.67	46.56	**********
17	2.20	1.58	48.13	**********
18	2.14	1.53	49.67	**********
19	2.09	1.50	51.17	**********
20	2.07	1.48	52.65	**********
21	1.94	1.39	54.05	**********
22	1.90	1.36	55.40	**********
23	1.86	1.33	56.73	**********
24	1.82	1.30	58.04	**********
25	1.73	1.24	59.27	**********
26	1.70	1.22	60.49	**********
27	1.68	1.20	61.69	**********
28	1.62	1.16	62.85	**********
29	1.61	1.15	64.00	**********
30	1.53	1.09	65.10	**********
31	1.44	1.04	66.13	**********
32	1.40	1.00	67.14	**********
33	1.38	0.99	68.13	**********
34	1.37	0.98	69.11	**********
35	1.33	0.96	70.06	**********
36	1.32	0.94	71.01	**********
37	1.23	0.89	71.89	**********
38	1.17	0.84	72.73	**********
39	1.17	0.84	73.57	**********
40	1.12	0.81	74.38	**********

Tableau : Diagramme des 40 premières valeurs propres (sortie DTM)

Le tableau ci-dessus décrit les 40 premiers facteurs de l'ACP, *i.e.* les 40 premières dimensions du genre. La première valeur propre, comprise entre 1 et 145 (soit le nombre total de descripteurs pris en compte), est ici égale à 10.90, chiffre finalement peu élevé qui reflète

[27] Les tests statistiques ont été effectués avec le logiciel DTM développé par Ludovic Lebart, que je remercie chaleureusement pour son aide précieuse, tant pour l'utilisation du programme que pour ses conseils d'interprétation.

la très grande homogénéité du corpus. De manière générale, plus la valeur propre est importante, plus elle résume de variables et plus le facteur est intéressant en termes de synthèse.

A chaque facteur (ou valeur propre) est associé un pourcentage d'inertie, présenté en troisième colonne et exprimé relativement à l'inertie totale du nuage observé : ce pourcentage permet d'apprécier l'importance relative du facteur dans le tableau. On voit ainsi que le premier facteur correspond à un pourcentage d'inertie de 7,79%, ce qui est significatif si l'on considère les 146 descripteurs pris en compte : il rendrait ainsi compte de 11,37 variables, tandis que les quatre premières valeurs propres en rendraient compte de 29.18.

Le diagramme présenté en colonne 5 présente l'allure de la décroissance de ces valeurs. Comme le rappellent Escofier et Pagès (1998) :

> Le principe de lecture de ce diagramme est le suivant : si deux facteurs sont associés à des valeurs propres presque égales, ils représentent la même part de variabilité et il n'y a pas lieu a priori de retenir l'un et non l'autre dans l'interprétation. Réciproquement, une forte décroissance entre deux valeurs propres successives incite à retenir dans l'interprétation les facteurs précédant cette décroissance.

On observe un palier dans la décroissance des quatre premiers facteurs ; au-delà, elle est lente et régulière, ce qui est d'ailleurs souvent observé dans la pratique. Les quatre premiers facteurs correspondent ainsi à des irrégularités dans la forme du nuage de points qui demandent à être interprétés, tandis que les facteurs qui suivent ne représenteraient que « l'inévitable bruit qui accompagne toute observation de nature statistique » (*ibid.*).

Afin d'apprécier la confiance que l'on peut accorder aux premières valeurs propres obtenues, observons leur stabilité à l'aide des intervalles d'Anderson. Comme le rappelle Lebart :

> L'ampleur de l'intervalle donne une indication sur la stabilité de la valeur propre vis-à-vis des fluctuations dues à l'échantillonnage supposé laplacien. L'empiètement des intervalles de deux valeurs propres consécutives suggérera donc l'égalité de ces valeurs propres. Les axes correspondants sont alors définis à une rotation près. Ainsi l'utilisateur pourra éviter d'interpréter un axe instable selon ce critère. (2003 : 205)

On observe ainsi le caractère significativement distinct de la première valeur propre, dans la mesure où les deux premières valeurs propres n'empiètent pas :

	Borne inférieure	**Valeur propre**	**Borne supérieure**
vp1	9.07	10.93	13.16
vp2	5.37	6.47	7.79
vp3	4.49	5.41	6.52
vp4	4.31	5.19	6.26

Tableau : Intervalles de confiance des quatre premières valeurs propres

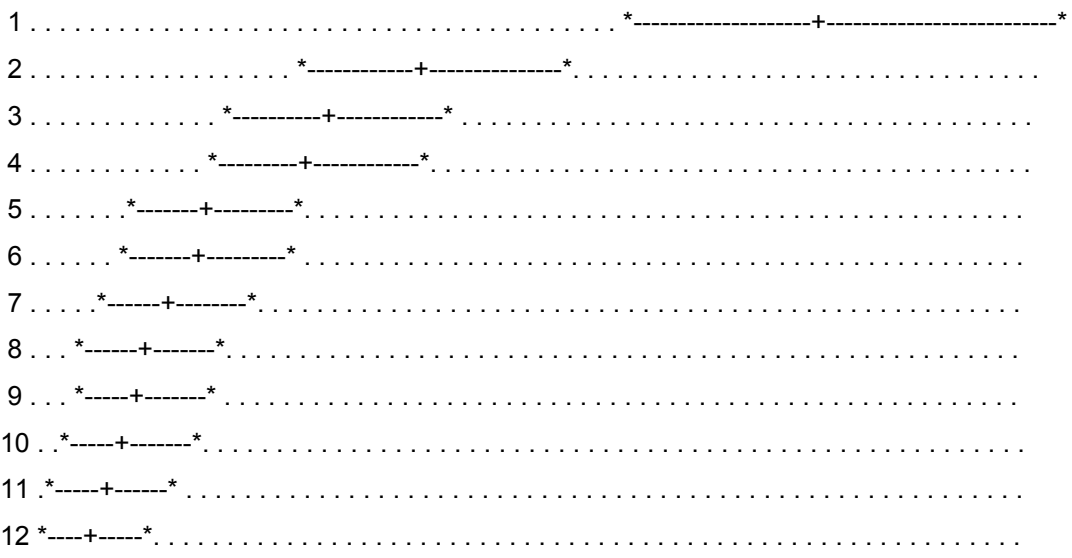


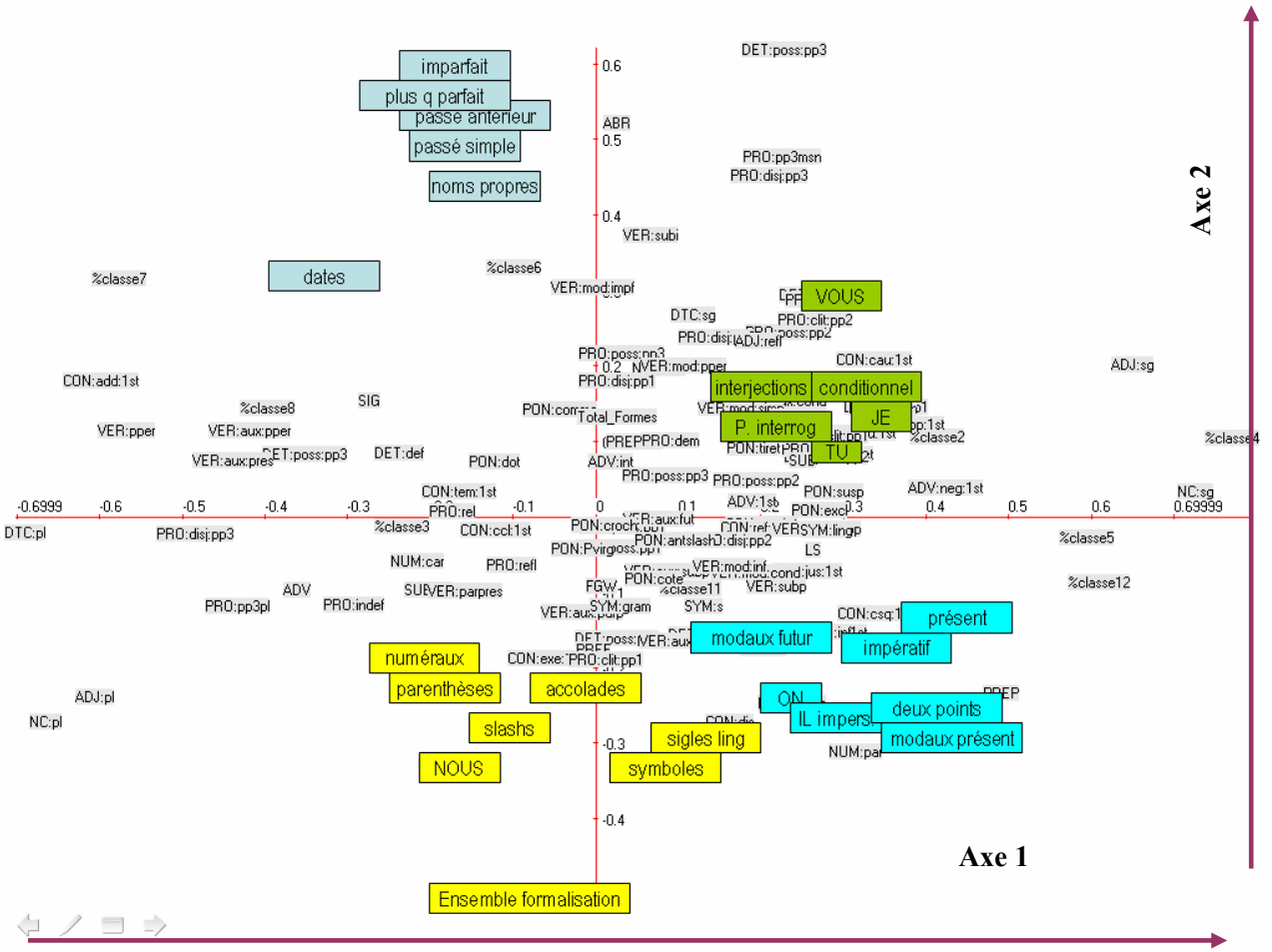
Figure : Position relative des intervalles

Les deux premiers axes principaux semblent ainsi les plus individualisés tandis que les intervalles de confiance des facteurs 3 et 4 empiètent largement.

3.5.1.2. Analyse des facteurs principaux

Si les valeurs propres et les pourcentages d'inertie évoqués précédemment encouragent la prise en compte de certains facteurs par rapport à d'autres, ils n'ont valeur que de pronostic, et il est globalement peu envisageable de déterminer mathématiquement les valeurs à prendre en compte – bien que de nombreux logiciels statistiques procèdent de la sorte. En effet, l'importance d'un facteur n'est pas un gage de son intérêt : un facteur de rang inférieur peut être beaucoup plus intéressant, précisément parce qu'il fait émerger des phénomènes moins visibles. En outre, un facteur de rang 1 n'est pas nécessairement interprétable, ce qui est particulièrement problématique en ce qui nous concerne.

Examinons d'abord le premier plan factoriel de l'ACP – on ne présentera ici que les représentations graphiques obtenues :



Graphique : Positionnement des variables sur les deux premiers axes factoriels

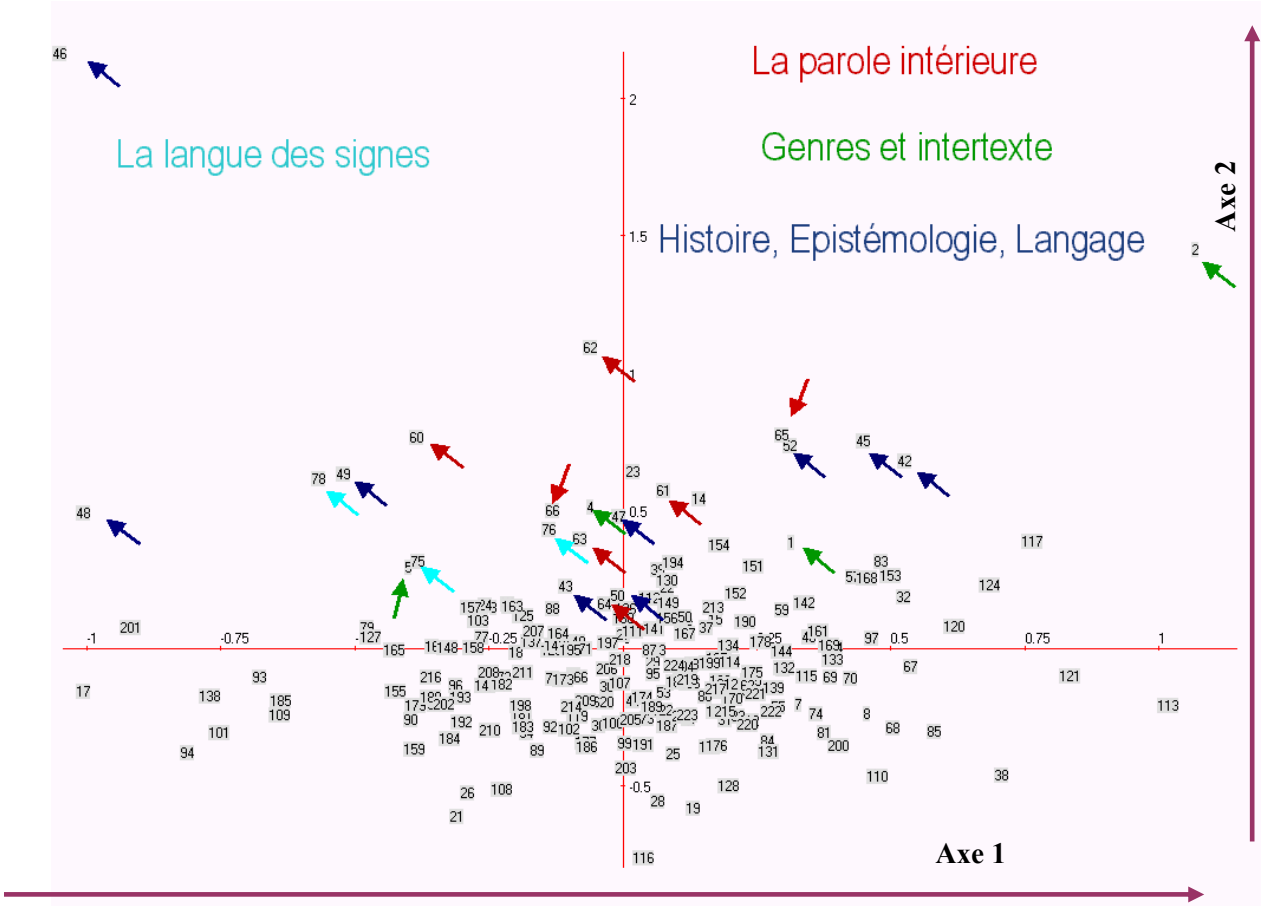
L'axe 1, qui représente 7.79% de la variance, est à ses extrémités marqué par une distinction grammaticale singulier/pluriel, qui présente un intérêt limité. Sur cette même dimension, les pronoms semblent s'opposer aux noms. On observe également que l'ensemble des descripteurs de la rhétorique scientifique (e.g. présent, pronoms *on* et *il* impersonnel, deux points, modaux présent, modaux futur, connecteurs d'opposition, etc.) s'oppose à un groupement de traits morphosyntaxiques défini par l'emploi intercorrélé du passé simple et de l'imparfait et de leurs homologues composés (en gris), qui dessine un *mode de narration plus historique* et plus proche du discours romanesque[28] (déjà observé *supra*).

Associé aux dates et aux noms propres, qui confirment bien la présence d'une composante historique, ce groupement s'oppose également sur le premier axe à un ensemble de descripteurs caractéristiques de l'oral (*en gris :* marques de première et de seconde personne, points d'interrogation et d'exclamation, interjections, etc.), spécifiques des articles plus exemplifiés travaillant sur corpus oraux, et sur le deuxième axe aux marqueurs de formalisation (*en jaune :* symboles, formalisations linguistiques, slashs, parenthèses et accolades).

[28] Il est d'ailleurs intéressant de souligner que cette opposition narrative (présent / passé) correspond au second facteur que Biber (1988) avait obtenu sur un corpus beaucoup plus hétérogène de textes anglais.

Cette opposition d'une dimension historico-narrative aux marqueurs d'une rhétorique scientifique est également observable sur le deuxième axe, qui extrait 4.62% du nuage de points : le facteur est *positivement corrélé* aux temps narratifs (imparfait, passé simple, auxiliaires imparfait et passé simple et modaux à l'imparfait), aux noms propres (et leurs abréviations) et aux marques de troisième personne, qui semblent indiquer la thématisation d'un objet (concept ou personne) ; il est au contraire *négativement corrélé* aux marqueurs de formalisation (symboles, formalisations linguistiques, slashs, parenthèses et accolades). En ce sens, il semble négativement associé aux textes plus exemplifiés (présence de renvois dans le texte, de ponctuations deux points, etc.). On note qu'il est également opposé aux pronoms personnels les plus caractéristiques de l'article (*on*, *il* impersonnel, *nous*) et aux numéraux. L'axe 2 oppose donc l'ensemble des caractéristiques principales du genre de l'article (*i.e.* l'ensemble des descripteurs que nous avons pris en compte et adapté aux spécificités du discours scientifique) aux particularités de la narration. En ce sens, il peut paraître paradoxal, et on pourrait penser qu'il s'agit d'un biais du corpus et écarter les textes qui contiennent une dimension historico-narrative ; elle nous semble pourtant particulièrement bien refléter le positionnement ambigu de la linguistique, encore marquée par la littérature.

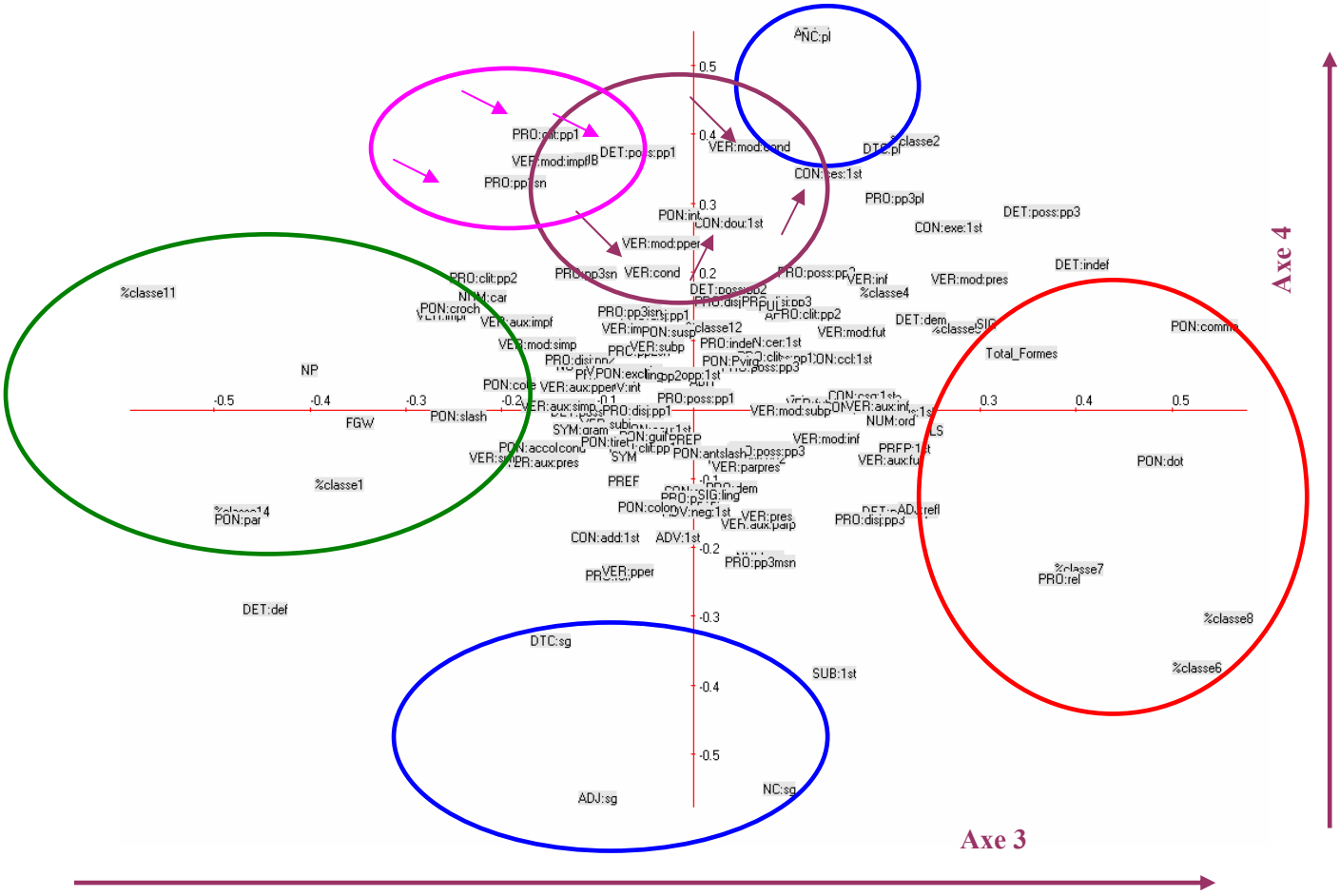
Cette tension des textes plus exemplifiés et historico-narratifs est particulièrement frappante lorsqu'on examine le positionnement des textes sur les deux premiers facteurs. On retrouve l'ensemble des textes appartenant aux numéros thématiques et aux revues plus historiques en positif sur l'axe 2 :



Graphique : Positionnement des individus sur les deux premiers axes factoriels

L'axe 3, qui représente 3,86% de la variance, semble se rapporter à la formalisation des textes et à la présence d'éléments de langue étrangère : l'ensemble des ponctuations, les numéraux, parenthèses, symboles, sigles, abréviations, slashs, accolades, éléments de langue

étrangère, etc. lui sont en effet négativement corrélés (en vert), tandis qu'il est positivement corrélé avec l'ensemble des autres classes de descripteurs, de même qu'avec les textes plus longs, avec virgules et points (en rouge). Il se pourrait qu'il soit lié aux différences entre textes plus théoriques et plus appliqués :



Graphique : Positionnement des variables sur les axes factoriels 3 et 4

Outre une opposition grammaticale singulier/pluriel (en bleu), déjà observée sur le premier axe factoriel, l'axe 4, qui résume 3,72% de la variance, est positivement corrélé à deux phénomènes : l'usage de la première personne du singulier (en rose) et celui d'un ensemble de descripteurs qui paraît rattaché au spéculatif : le conditionnel et ses modaux, de même que les connecteurs de doute et de concession (en violet). On remarque qu'à l'inverse, le présent ou les deux points lui sont négativement associés.

3.5.2. Approfondissement par cartes de Kohonen

Si les méthodes d'analyse factorielle permettent bien de visualiser les proximités et les oppositions des variables entre elles, le caractère linéaire de la projection entraîne des difficultés d'interprétation indéniables, liées notamment à la nécessité de visualiser un nombre important de cartes pour appréhender les interactions entre variables de manière pertinente.

Ainsi, l'examen des deux premiers plans factoriels que nous avons effectué nous a permis de mettre au jour certains contrastes et différents pôles d'opposition qui demeurent pourtant à valider en raison des biais qu'entraîne la visualisation des variables sur les axes linéaires de l'ACP : deux variables proches, voire presque recouvertes, sur deux axes peuvent en effet

s'avérer très éloignées l'une de l'autre. Observer les descripteurs dans un espace à trois dimensions[29] améliore la qualité de la description sans toutefois résoudre le problème.

C'est en raison de ces difficultés que nous avons choisi d'approfondir notre description en recourant aux *cartes de Kohonen*, fonction que propose le logiciel DTM de L. Lebart.

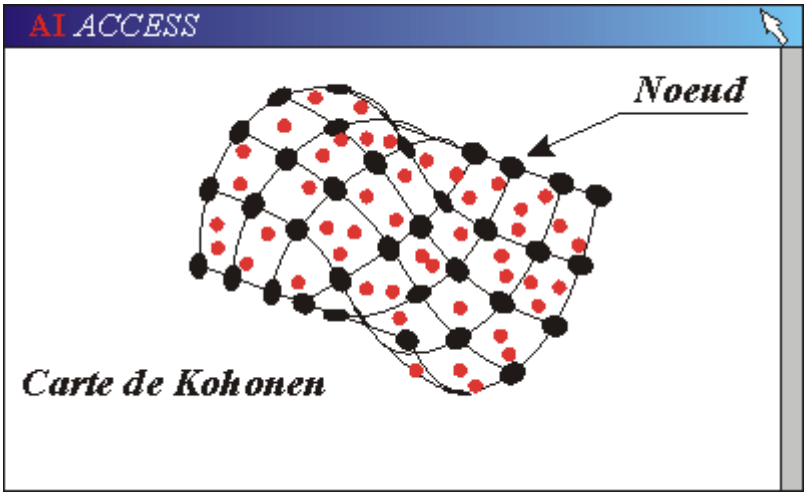
3.5.2.1. Des cartes de Kohonen (ou cartes auto-organisatrices)

C'est dans le début des années 80 que T. Kohonen a proposé la méthode des cartes auto-organisatrices (SOM, *Self-Organizing Maps*), qu'on présente généralement comme un cas particulier des réseaux de neurones. Dans la mesure où elles sont capables de s'étirer et d'épouser plus étroitement le nuage de points, les cartes de Kohonen peuvent être considérées comme un équivalent qualitatif et non linéaire de l'ACP.

L'algorithme de Kohonen est un algorithme de classification qui regroupe les observations en classes, en respectant la topologie de l'espace des observations. Une notion *a priori* de voisinage entre classes est ainsi définie :

> L'algorithme SOM est une variante de celui des K-Means qui, lors d'une itération, modifie non seulement un centre sélectionné comme étant le plus proche d'une donnée, mais aussi les centres voisins pour un graphe de voisinage fixé. Le graphe implique des interactions latérales entre centres qualifiés de voisins. (Priam : 2003 : 31)

On suppose généralement que les classes sont disposées sur une grille rectangulaire dans laquelle les voisins de chaque classe sont naturellement définis ; en d'autres termes, on construit une représentation bidimensionnelle d'une distribution multidimensionnelle et on dispose d'une représentation graphique unique des données dans l'espace de sortie, ce qui présente un intérêt indiscutable lorsqu'il s'agit d'observer un nombre important de données :



Graphique : exemple de carte de Kohonen

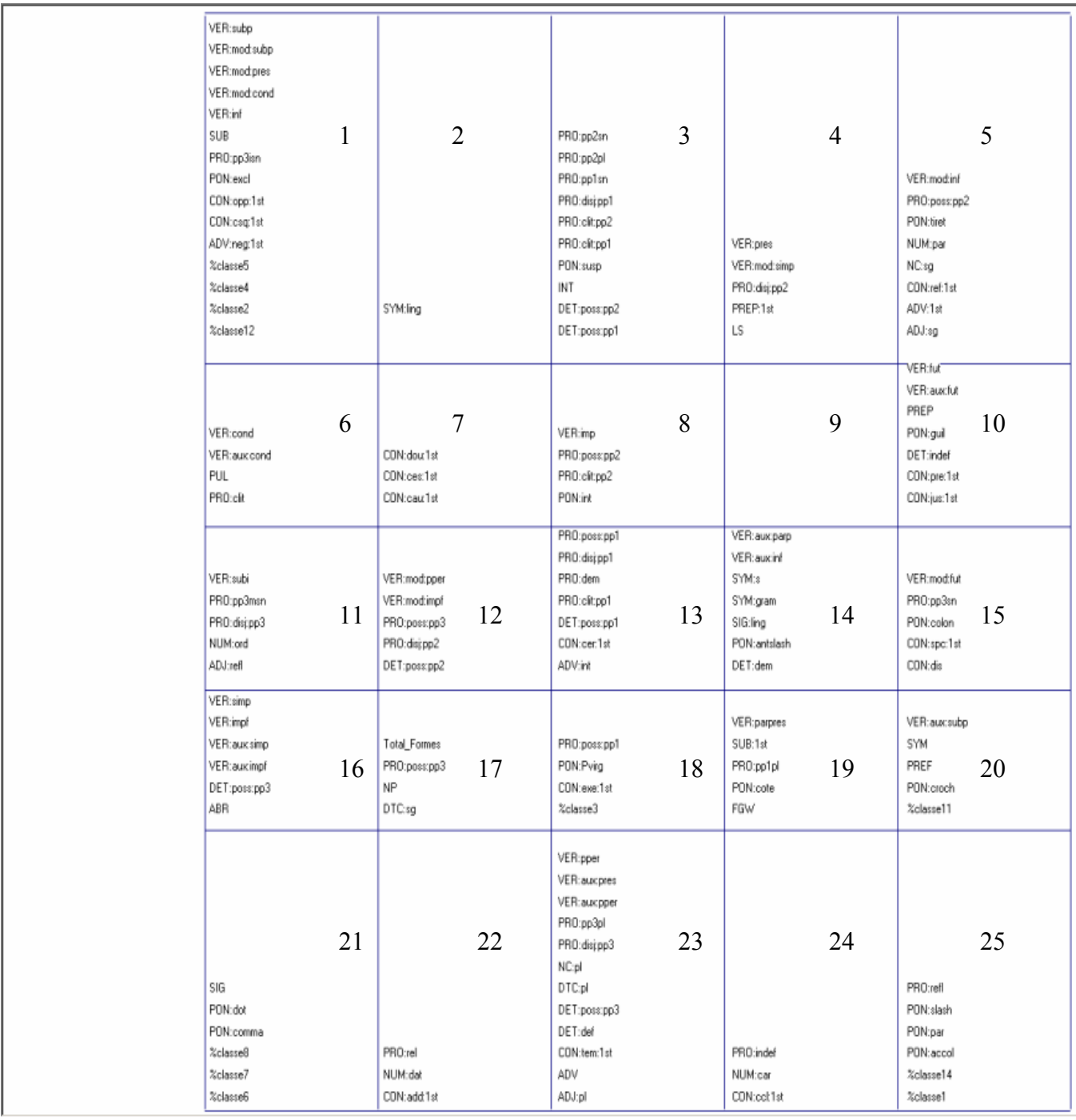
3.5.2.2. Examen de la carte de Kohonen obtenue

Nous avons opté pour une représentation des variables en 25 neurones (5*5), que nous avons numérotés. La carte obtenue permet de mettre en évidence des groupements de

[29] Ce que proposent divers logiciels de statistique, comme XLStat par exemple.

variables qui n'apparaissaient pas nécessairement sur les plans factoriels observés précédemment.

Si les pôles *historico-narratif* et *exemplification* mis au jour se démarquent nettement (neurone 16 et 3), le groupement « rhétorique scientifique » est ici scindé en plusieurs cases éloignées : les pronoms *il* impersonnel et *on*, qui semblaient proches, sont par exemple très opposés (classes 1 et 15). *on* est nettement associé au futur, aux deux points et aux connecteurs de spatialité, ce qui lui confère bien un rôle de guide du lecteur, tandis que l'impersonnel est entre autres descripteurs corrélé aux modaux présent et conditionnel, aux négations et aux connecteurs d'opposition et de conséquence, ce qui semble lui concéder la fonction de garant de l'objectivité / l'objectivation scientifique. La dimension *formalisation* de l'article se distribue enfin dans les classes adjacentes 20 et 25.



Graphique : Carte de Kohonen

3.5.3. Classification Ascendante Hiérarchique

L'ACP a été complétée d'une CAH, destinée à approfondir la description du genre à travers les regroupements textuels mis au jour. On présentera dans un premier temps les caractéristiques des regroupements obtenus, avant de les décrire relativement aux axes factoriels précédemment observés.

3.5.3.1. Analyse des partitions obtenues

Le nombre de classes devant être prédéterminé, nous avons choisi de partitionner le corpus en 12 classes. Les 12 partitions obtenues sont irrégulières : trois des douze classes ne contiennent qu'un individu, isolé pour ses spécificités par rapport à l'ensemble du corpus, tandis que les partitions restantes contiennent de 11 à 49 textes.

On analysera d'abord les singletons (textes limites ou « hors la loi ») de la CAH, avant d'examiner les neuf partitions plus importantes.

A. Description des textes limites

Les trois textes écartés, dont nous présentons les références bibliographiques ci-dessous, occupaient déjà une position marginale dans le graphique précédent (graphique 70) :

- **Classe 3 :** texte 46 : Romashko. « Vers l'analyse du dialogue en Russie » in Archaimbault. *Le dialogue, un objet d'étude?*, HEL 22, 2000.
- **Classe 11 :** texte 117 : Dostie et de Sève. « Du savoir à la collaboration. Étude pragma-sémantique et traitement lexicographique de *t'sais* » in Bergounioux, Gabriel. *Les connecteurs entre langue et discours*, RSP 5, 1999
- **Classe 12 :** texte 2 : Kanellos. « De la vie sociale du texte. L'intertexte comme facteur de la coopération interprétative » in Kanellos. *Sémantique de l'intertexte*, Cahiers de Praxématique 33, 1999.

Dans la mesure où ils ont été isolés, ces textes possèdent un profil qu'il est raisonnable de considérer comme « limite » sinon « hors genre », ce qui intéresse particulièrement la caractérisation du genre de l'article. Ils reflètent d'ailleurs les dimensions factorielles mises au jour :

♦ le texte 46 se situe ainsi à l'une des extrémités du facteur 2. C'est un article d'histoire de la linguistique de Sergej A. Romashko qui porte sur les recherches sur le discours dialogique en Russie dans les années 20. Le texte est rédigé au passé simple et à l'imparfait (et à leurs homologues plus-que-parfait et passé antérieur), temps dont il détient les proportions d'utilisation les plus élevées, ainsi qu'au passé composé (14ème rang si l'on ordonne les scores des 224 textes du corpus). On relève également de nombreuses formes passives au passé composé (avec l'auxiliaire *été* – 10ème rang) comme dans :

> S'il envisagea dans son livre *Les Problèmes de l'œuvre de Dostojevskij* (la deuxième édition **a été publiée** sous le titre plus connu de *Problèmes de la poétique de Dostojevskij*) les romans de Dostojevskij comme des romans dialogiques…

Le présent est en revanche très faiblement représenté – 218ème rang : il ne représente que 24,93% de l'ensemble des formes verbales du texte (moyenne de 38,49%), tandis que les modaux au présent n'en représentent que 1,07% (moy. de 5,05 et 223ème rang). Cette faible proportion de modaux est probablement corrélée avec le fait que l'article de Romashko est le texte qui contient le nombre le plus bas d'infinitifs (4,56% de l'ensemble des formes verbales *vs.* moy. de 18%).

L'isolement du texte n'est pas seulement lié à ses spécificités verbales : à l'instar de l'ensemble des textes historiques, il contient des dates, des noms propres et des abréviations. En revanche, les représentations de ces éléments sont particulièrement importantes (2ème rang pour les noms propres, 3ème pour les abréviations et 4ème pour les dates). L'extrait ci-dessous est particulièrement représentatif du contenu du texte : les six lignes suivantes comportent ainsi six noms propres, 5 abréviations et 4 dates :

> La situation linguistique s'orienta, elle aussi, vers le modèle européen : à partir de cette période, on essaya de prendre la langue parlée, l'usage — et non le texte d'autorité — comme base pour la langue « littéraire », pour le « russe standard » (Uspenskij 1985, p. 16). Mais les développements furent contradictoires (v. Zivov 1996), le slavon ne pouvant être refoulé totalement, ainsi, l'idée proclamée par A. A. Šaxmatov au tout début du XX^e^ siècle, selon laquelle le russe standard n'était en réalité que le slavon transformé, devint assez populaire ; c'est pourquoi Viktor Šklovskij déclarait : « La langue littéraire russe… est d'une origine étrangère à la Russie » (Šklovskij 1917 [1965], p. 95-96).

Comme on l'a déjà évoqué, les textes historiques contiennent en général un nombre plus important d'anaphoriques singulier et pluriel, ce qui suggère des textes dont la prise en charge fonctionne essentiellement sur un mode délocuté[30]. Le texte 46, qui contient une proportion très importante de pronoms, détient de fait le nombre le plus important de pronoms anaphoriques au singulier (*il* ou *elle*), naturellement associé à une représentation remarquable de déterminants possessifs et de disjoints de 3^ème^ personne (3^ème^ et 9^ème^ rangs). Par voie de conséquence, les pronoms personnels restants sont peu, voire non représentés : aucune occurrence de *nous* n'a ainsi été comptabilisée (seuls 14 textes n'en contiennent aucun) et le pronom impersonnel *il* est très peu employé (2 occ. comptabilisées).

En outre, on remarque un emploi notable des connecteurs d'addition (1^er^ rang), le texte n'étant nettement pas argumentatif. On relève d'ailleurs relativement peu de connecteurs de type autre : aucun connecteur de conclusion, très peu de connecteurs de disjonction (223^ème^ rang), etc.

Etant donné les spécificités du texte et sa dimension historico-descriptive, on relève une proportion importante d'adjectifs (12^ème^ rang) et de déterminants définis (10^ème^ rang).

On observe enfin une proportion importante de substantifs pluriels, qui représentent les 2/5^e^ de l'ensemble des noms communs (1/5^e^ en moyenne) : cette particularité est en partie liée aux modalités de référence du texte, qui désigne régulièrement les courants de pensée en termes de classes d'individus (les « philologues populistes », les dialectologues, etc.).

♦ le texte 117 est un article plus appliqué de sémantique de Gaétane Dostie et Suzanne de Sève, dédié à l'étude du marqueur discursif *t'sais*. C'est un article particulièrement exemplifié, qui comporte naturellement de nombreux *t'sais* dans son corps et maintes reprises des marqueurs observés, cf. :

> De plus, *t'sais* possède plusieurs emplois qui lui sont propres, comme le montre l'exemple (2) dans lequel la présence de *tu sais*/ *vous savez* est pour le moins étonnante.

Les exemples étant nombreux, l'article est particulièrement structuré (12^ème^ rang). Dans la mesure où nous n'avons procédé qu'à l'extraction des exemples décrochés des textes, il en demeure naturellement dans les articles qui en contenaient dans leurs corps[31]. Toutefois, de nombreux textes du corpus portent sur un marqueur discursif et n'ont pas pour autant été écartés ; l'objet du texte 117 présente donc des particularités qui entraînent son isolement. Comme nous l'avons déjà étudié *supra*, le genre de l'article contient une proportion particulièrement faible de marques de seconde personne ; or, l'objet même de l'article renferme le pronom *tu*, et on relève de très nombreuses marques de seconde personne. Le texte détient ainsi les proportions les plus élevées de *tu*, qui représente 24.14% de l'ensemble des pronoms, et du déterminant possessif de seconde personne du singulier. Mentionnons en

[30] Le délocuté renvoie à l'opposition *personne / non personne* de Benveniste et est associé à la *non personne*, au « il » (ou encore à « ce dont on parle »).

[31] Il s'agit essentiellement de reprises et de mentions aux exemples donnés ; on reprend généralement des fragments d'exemples pour les discuter.

outre les proportions considérables détenues par *vous* (2,53%/3ème rang) et le clitique *te* (4ème rang).

La première personne du singulier est également très présente, eu égard à la présence d'une relation *je/tu(vous)* induite par l'objet même de l'article : on relève une proportion importante de *je* (7,47% des pronoms relevés/9ème rang sur 224), du clitique *me* (11ème rang) et du déterminant possessif de première personne du singulier (9ème rang). La première personne est globalement employée pour gloser les acceptions du marqueur (je universel) :

II.1 *T'sais, T ≅*

> Je fais appel à ta collaboration et à ta compréhension afin de saisir la teneur de l'explication que je te communique au moyen du texte T.

Etant donné les proportions des pronoms de seconde personne et de première personne du singulier, on relève finalement peu d'impersonnels (avant-dernier rang, soit 13 occ./moyenne de 28 occ.) et de *on* (220ème rang, soit 16 *on*/moyenne de 32 occ.).

Les singularités de l'objet *t'sais* et de ses modalités d'analyse entraînent également des proportions marginales de verbes : le texte contient ainsi un nombre très important de formes verbales (3ème rang) et de présent (10ème rang). En revanche, on relève une proportion plus faible de modaux au présent (rang 202).

T'sais étant un marqueur oral, il est globalement appréhendé sur corpus oral :

> La plupart des exemples analysés sont tirés de la vie réelle, d'émissions de télévision ou de la Banque de données textuelles de l'Université de Sherbrooke. Nous avons parfois légèrement reformulé certains d'entre eux, afin de les rendre plus facilement interprétables.

On trouve ainsi de très nombreuses interjections, tant dans les exemples que dans le corps de l'article :

> Ajoutons qu'il peut être précédé, selon le contexte, d'expressions diverses comme *ah*, *bof, ben, non mais*, etc.

Le texte 117 est de fait celui qui en contient la proportion la plus importante. A fortiori, on relève de nombreux points d'interrogation (4ème rang), également spécifiques à l'oral :

> il appartient au paradigme d'expressions interactives qui sollicitent l'accord ou l'approbation comme *hein ?*, *n'est-ce pas ?*, *O.K. ?*, *tu comprends ?*, *tu me suis ?*, etc.

Davantage intéressé par la détermination de valeurs et d'acceptions sémantiques du marqueur, l'article est peu intéressé par la quantification : on relève très peu de chiffres (rang 202).

♦ le texte 2, rédigé par Kanellos, est enfin pour le moins particulier, dans la mesure où il est volontairement hors-genre : les première et dernière sections de l'article sont rédigées en dialogue (socratique),

> — De quoi est fait un texte ?
>
> — Pardon ?
>
> — Je vous demande : de quoi est fait un texte ?
>
> — La question paraît simple, et plutôt naturelle (…)

tandis que les sections 2 et 4 sont de format plus conventionnel, bien que *je* continue de s'adresser à *vous* :

> Je viens ainsi de justifier mon obsession de n'aborder mon sujet central, l'intertexte, qu'indirectement, à travers le texte. Il me fallait de toute façon le texte et à travers lui la sensibilité à une vision textuelle de la sémantique. Justement, pour faire mieux sentir tant la nécessité que les

limites d'un texte seul. Je peux vous prendre par la main, cher ami, et reprendre, avec vous, sans grande difficulté, tout ce que je viens de dire pour le texte-qualité, dans le cas de l'intertexte.

L'article est par conséquent très structuré (4ème rang), et il contient une proportion importante de marques de première personne du singulier et de seconde personne du pluriel – l'interlocuteur de *je* étant vouvoyé : on y relève un nombre considérable de disjoints *moi* (3ème rang), de *je* (6ème rang), de déterminants possessifs de première personne du singulier (6ème rang) et de clitiques de première personne du sg. (12ème rang). C'est le texte qui contient le nombre le plus important de *vous* clitiques (je *vous* demande), et une proportion à peine moins importante de pronoms *vous* (2nd rang).

L'article ayant trait à l'intertexte, il comporte de nombreux anaphoriques singulier (6ème rang), tant dans le corps des deux sections plus traditionnelles, que dans les parties dialogiques :

— (…) Il éclate d'emblée devant nos yeux en engageant nombre de domaines. Sans qu'on n'y puisse faire la moindre action. Et surtout il engage nombre de pratiques...

— Par conséquent, il convoque une multitude de schémas de compréhension et donc de stratégies de régulation de la dynamique signifiante. Nous sommes toujours d'accord ?

Outre les pronoms personnels, on relève différentes irrégularités statistiques, notamment en ce qui concerne les représentations des connecteurs. On dénombre en effet un nombre très important d'adverbes et de connecteurs (2ndes proportions les plus élevées) ; l'article de Kanellos est de surcroît le texte qui contient le nombre le plus élevé de connecteurs de concession (lié aux nombreuses marques de politesse). On y relève également un nombre très important de connecteurs de doute (rang 2) et d'opposition (rang 4).

Le style plus oralisé du texte entraîne les sur-représentations de ponctuations classiquement peu employées, comme les points d'interrogation (4ème rang) et de suspension (13ème rang). On y relève également de nombreux points, mais peu de ponctuations deux points (220ème rang).

Le ton étant spéculatif, on relève de nombreux conditionnels (8ème rang) :

Pour tout texte, il n'y aurait donc pas *un* intertexte mais *plusieurs* intertextes, chacun correspondant à une intention de compréhension placée au sein d'une pratique (…)

Notons enfin que les modalités de citation du texte sont plus littéraires que scientifiques : les références sont présentées en notes, et on ne relève donc aucune date dans le corps de l'article. Si cette modalité de citation est partagée par l'ensemble des textes du numéro, on relève toutefois maintes références de type (*auteur, date*) dans les autres articles.

Synthèse

Les textes limites reflètent ainsi les dimensions obtenues précédemment et nous apportent des indications intéressantes quant aux éléments frontaliers du genre, qui pourraient éventuellement compromettre son identification. S'il est évident qu'un article n'est normalement pas rédigé en format dialogue, l'isolement du texte historique 46 illustre les modalités rédactionnelles bien distinctes des textes d'histoire de la linguistique. Le texte 117 démontre quant à lui les difficultés posées par l'étude morphosyntaxique des textes de linguistique, qui bouleversent les statistiques d'un marqueur, ou d'un ensemble de marqueurs, dès lors qu'ils s'y attachent ; on voit toutefois que les descripteurs de seconde personne et de première personne du singulier sont particulièrement sensibles, car finalement peu caractéristiques du genre.

B. Description des classes restantes

Le tableau qui suit présente les caractéristiques générales (positives et négatives) des 9 classes restantes :

Classe	Nb de textes	Textes	Caractéristiques	
			Positives	Négatives
1	34	7 8 10 19 33 34 35 38 67 68 69 70 74 81 85 92 110 113 115 116 120 124 131 136 139 153 161 162 176 178 190 191 200 220	Pronoms Déterminants indéfinis Prépositions Verbes Adverbes/connecteurs Conn. conséquence Subordonnants Négations Conn. opposition	Conn. addition Noms Déterminants définis Adverbes Participes passés Dates
2	12	24 29 32 37 40 44 45 51 86 133 142 215	je me clitique mon, ma, mes Guillemets simples Modaux imparfait Numéraux Eléments étrangers	Prépositions Noms Adverbes Indices structuration Adjectifs
4	11	53 65 80 83 87 95 97 107 189 196 222	Conditionnel (temps simple, auxiliaires et modaux) Numéraux ordinaux Connecteurs de doute et de justification	Points Numéraux son, sa, ses Auxiliaires présent
5	49	5 11 13 28 31 47 50 57 61 62 64 82 88 96 98 103 104 121 122 123 130 137 144 145 148 149 150 151 152 154 155 156 158 163 165 177 184 186 188 195 199 204 206 207 210 214 217 219 224	Connecteurs d'exemplification Dates Abréviations Symboles linguistiques son, sa, ses/(le)sien Points virgules	Numéraux Symb., sigles et abr. Dét. Notre, nos / pron. nous Impératifs Points Chiffres
6	20	1 9 27 36 59 73 84 106 118 128 129 132 134 135 170 171 197 209 221 223	Singulier Renvois dans le texte Formalisation linguistique Conn. justification et reformulation	Pluriel Conn. temporalité Dates Noms propres Impératifs
7	21	12 15 16 20 21 25 26 54 55 99 100 102 105 108 109 140 143 159 166 203 205	Ponctuations On Eléments étrangers Participes présent Conn. disjonction Adverbes Parenthèses et accolades Guillemets Déterminants définis Futur Deux points	Déterminants Points Verbes Subordonnants Adverbes et connecteurs Pronoms Conn. conséquence Virgules Ils Conditionnel (temps simple et auxiliaires) Son, sa, ses
8	24	23 30 39 41 42 43 52 56 58 63 72 89 111 112 114 146 167 168 169 175 187 194 212 213	Nous clitique Dét. notre, nos Nous Déterminants défini	Déterminants indéfinis Pluriel ils Conn. disjonction

			Singulier Impératifs	Structuration on Pronoms indéfinis
9	28	48 60 66 90 91 93 94 101 119 147 160 172 173 174 179 180 181 183 185 192 193 198 201 202 208 211 216 218	Auxiliaires présent et passé Ils Participes passés Pluriel Parenthèses Nous (nos, notre, (le)nôtre) Numéraux	Pronoms Singulier Deux points Présent Futur Dét. Son, sa, ses Je Modaux présent on
10	22	3 4 6 14 17 18 22 49 71 75 76 77 78 79 125 126 127 138 141 157 164 182	Sigles Points Dét. Leur(s)	Subordonnants Parenthèses Ponctuations

Tableau : Caractéristiques des 9 classes

La classification ayant été effectuée sur les textes hors exemples, les phénomènes linguistiques participant à la formation des partitions relèvent globalement des caractéristiques du corps de l'article : les textes agrégés du fait des spécificités de leur objet sont plus marginaux.

♦ Deux des neuf classes sont essentiellement liées à un usage spécifique des marques personnelles de première personne, du singulier pour la classe 2 et du pluriel pour la classe 8. On observe dans les douze textes de la classe 2 la présence importante du pronom personnel *je*, qui renvoie dans 11 textes sur 12 à l'auteur des textes[32]. Ce dernier se met ainsi particulièrement en avant dans l'ensemble des textes de la classe :

> Enfin, après l'établissement d'une grille de comparaison des typologies et des commentaires que cette comparaison conduit à faire, je proposerai la mienne. (024)
>
> La manière la plus sûre d'aborder un mot est sans aucun doute de considérer l'ensemble de ses usages. Si j'ai violé ce principe dans l'étude des prépositions françaises (Vandeloise 1986), c'est parce que, méthodologiquement, il me paraissait difficile de tenter de prime abord une étude exhaustive. (037)
>
> De fait, je ne pense pas que les grammairiens soient par intuition « arrivés au seuil de la découverte de la complétive ». (045)

Je étant un pronom somme toute peu fréquent, la représentation du pronom et des marques de première personne dans les textes est finalement très raisonnable en termes quantitatifs, mais suffit à isoler les textes. Il en va différemment des marques de première personne du pluriel, deux fois plus représentées en moyenne : les textes de la classe 8 contiennent ainsi un nombre particulièrement massif de *nous*, employé comme *je* pour référer à l'auteur :

> Les conditions de traduction de ce rapport et notamment l'urgence **nous** imposaient pratiquement de trouver une solution à ce problème en l'absence de texte original complet. Dès lors, **nous nous** sommes placés dans la situation d'un enquêteur cherchant à retrouver le chaînon manquant permettant de tout expliquer. (023)
>
> **Nous** admettons que chaque acte textuel active une information que **nous** appelons, à la suite de Bally (1965 [1932]), un propos. **Nous** n'utilisons pas le terme de " rhème " (…) (039)

[32] Le texte 40 se distingue en effet des autres par un usage universalisant de *je*, comme dans :

"si j'énonce P, c'est que je suppose que tu n'es pas au courant de P, et je souhaite que tu en prennes note"

> Le fait que **nous nous** intéressions aux relations sémantiques entre le verbe et ses compléments ne signifie pas que **nous** cessons de **nous** intéresser aux relations sémantiques entre le préfixe et la base. **Nous** considérons les points de vue interne et externe comme étant dans une relation de complémentarité nécessaire. En d'autres termes, pour l'étude sémantique de RE, **nous** adopterons un point de vue à la fois interne et externe. Ce n'est qu'en articulant les deux points de vue que l'on peut, selon **nous**, aboutir à un traitement sémantique adéquat du préfixe RE. En fait, ainsi que **nous** le verrons dans la section 3. (…) (072)

On a d'abord envisagé qu'un emploi aussi intensif de *nous* était possiblement corrélé aux styles d'auteur[33], dans la mesure où les deux articles de Laurendeau dont nous disposons, qui diffèrent pourtant grandement en termes d'objets (« L'alternance futur simple / futur périphrastique : une hypothèse modale » / « Condillac contre Spinoza : une critique nominaliste des glottognoses »), sont regroupés au sein de la classe 8. On relève en outre au sein de la même classe un article co-écrit par Cortès et Kriegel (030), alors que le texte 18 de Cortès n'y est pas associé : il nous semblait en effet plus que probable qu'un texte co-écrit différait stylistiquement d'un texte rédigé individuellement, notamment au niveau d'un recours peut-être plus fréquent à *nous*. Pourtant, on ne trouve dans la classe 8 que 3 des 31 textes co-écrits, et a fortiori, les deux textes de G. Kleiber que nous détenons ne sont pas regroupés, bien que Kleiber emploie de nombreux *nous* dans les deux cas.

En négative on observe dans la classe 8 un usage plus modéré des formes plurielles (et des pronoms anaphoriques pluriel), de même qu'une représentation moins importante de l'indéfini *on*. Après examen des corrélations textuelles de *nous*, on relève une opposition probablement concurrentielle du pronom avec *je* (-0,34) et *on* (-0,2), visiblement liée aux valeurs exclusive/inclusive de *nous*. Il semble que les textes de la classe 8 aient tendance à recourir à *nous* de manière plus indifférenciée, tant pour référer à l'auteur que pour inclure le lecteur. La présence des formes plurielles nécessiterait bien entendu un examen attentif de l'ensemble des substantifs et des objets étudiés dans chacun des articles, ce qui n'est malheureusement pas envisageable pour l'heure.

♦ La classe 4 se distingue essentiellement par un usage important du conditionnel (temps simples, auxiliaires et modaux) et des connecteurs de concession et de doute, qui lui sont d'ailleurs corrélés au niveau textuel. Il fédère des textes globalement plus spéculatifs et contenant peu de quantification (cf. corrélation négative avec les numéraux), à l'exception toutefois des individus 080 et 083 qui s'intéressent précisément à l'hypothétique, et comprennent de ce fait un nombre important de verbes conjugués au conditionnel :

> Dans ces deux énoncés, le conditionnel (*serait* / *aurait vendu*) ne peut s'interpréter qu'en fonction du statut de la qualification apposée à gauche (083)

Cette présence importante du conditionnel – bien entendu associée à d'autres proportions de descripteurs – semble suffire à regrouper les quatre textes de Combettes[34] dont nous disposons. Ces trois articles sont en effet plus spéculatifs et le conditionnel a très souvent valeur de questionnement :

> Comment, surtout, éviter une sorte de cercle, qui consisterait à prendre la cause pour l'effet, en considérant, par exemple, que tel fait de langue est la trace, le résultat, de la constitution progressive de la "phrase complexe", de l'hypotaxe, alors qu'au contraire la grammaticalisation des faits de subordination résulterait de l'existence même de ces phénomènes ? (196)
>
> on ne peut affirmer, en particulier en raison de l'absence de documentation suffisante sur la période romane primitive, que les formes qui seraient les plus proches du système latin et que le français

[33] Facteur de variation que nous aborderons *infra,*dans la seconde partie analytique de la thèse.

[34] Soulignons que l'un de ces trois textes est co-écrit avec S. Prévost.

> ne ferait en quelque sorte que continuer, seraient les premières à se trouver intégrées dans la fonction de détermination nominale. (107)

De surcroît, on y relève de nombreuses réserves, et un certain recul quant à l'étude effectuée :

> On remarquera au passage l'intérêt qu'il y aurait à procéder à une étude d'ordre typologique, qui permettrait de caractériser le français par rapport à d'autres systèmes linguistiques. (222)

> Rappelons que nous n'abordons pas ici la question de l'ordre des syntagmes, limitant nos observations aux marqueurs dont le rôle est assumé par une expression particulière ; une étude plus complète devrait évidemment prendre en compte les deux aspects de la problématique, dans la mesure où certaines relations peuvent être établies entre les deux domaines (222)

On retrouve cet emploi spéculatif du conditionnel dans l'ensemble des autres textes, tant pour nuancer la valeur et la portée des hypothèses proposées que pour modérer, voire contester, celles d'autres instances :

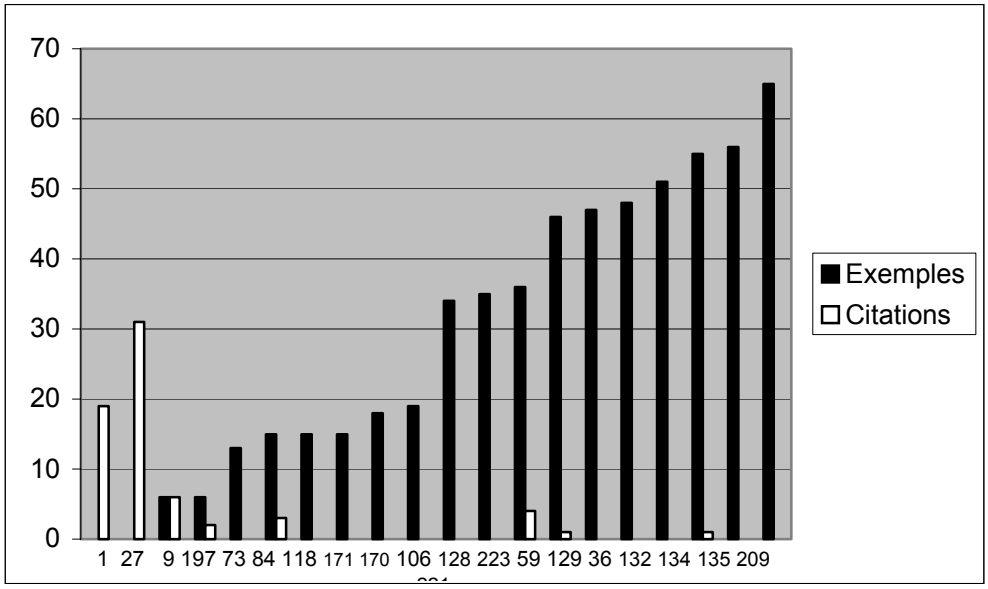
> On peut alors formuler l'hypothèse d'une validation plus générale de l'analyse que nous venons d'effectuer : le schéma mis à jour ne serait pas seulement le principe d'organisation des phrases (primitives) mais aussi celui des syntagmes (effectivement) nominaux. On aurait ainsi identifié un "schéma fondamental" applicable récursivement aux structures linguistiques. (189)

> Il n'est donc plus besoin de faire apparaître une instance extérieure au locuteur lui-même pour que la production endophasique se trouve justifiée : l'acte de nommer le réel mentalement ou d'opérer des prédications verbales accompagnerait automatiquement les processus mentaux de représentation et d'analyse. (065)

♦ Les textes de la classe 6 sont particulièrement appliqués et globalement dédiés aux études syntaxico-sémantiques de marqueurs. Ils sont en effet caractérisés par une présence importante de renvois dans le texte, qui sont comme on l'a vu *supra* nettement liés aux textes exemplifiés, et de marqueurs de formalisation linguistique, eux-mêmes globalement syntaxiques. On y trouve enfin peu de dates et de noms propres, en effet peu caractéristiques des textes plus exemplifiés.

La classe 6 comprend ainsi 5 articles (sur un total de 15) du numéro 5 de la revue *Scolia*, consacrés aux « Problèmes de sémantique et de relations entre micro- et macro-syntaxe », de même que quatre études explicitement dévolues à l'analyse d'un ou deux marqueurs (textes 36, 84, 221 et 223). On note en outre que les trois textes de Paillard (dont un co-écrit) sont réunis.

Le graphique ci-dessous illustre bien cette forte présence d'exemples dans la très grande majorité des textes de la classe ; seuls 2 textes n'en contiennent aucun et ne contiennent que des citations, ce qui indiquerait une dimension plus théorique :



Graphique : Eléments constitutifs des textes de la classe 6

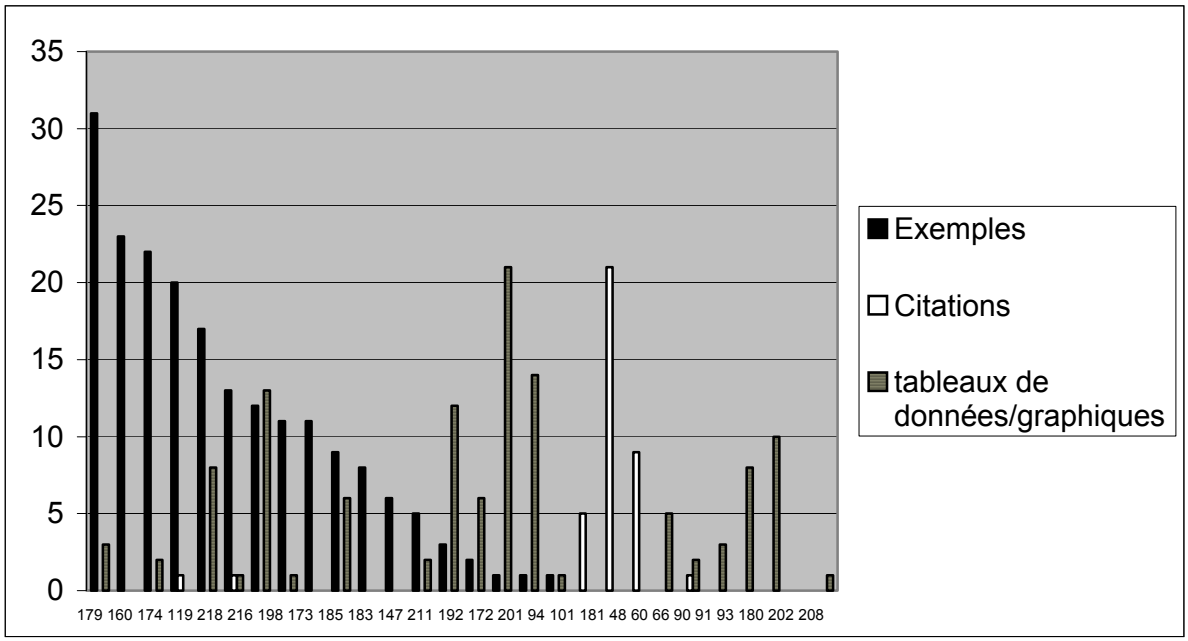
♦ La classe 7, caractérisée par un nombre important de ponctuations, d'éléments étrangers, de participes présent et de *on*, semble rassembler les travaux sur le lexique et les études classificatoires ou typologiques. On y trouve ainsi 4 articles (sur un total de 9) du numéro des cahiers du *CIEL* intitulé « Théories et Pratiques du Lexique », de même que deux textes (sur 7) de celui dénommé « Problèmes de classement des unités lexicales ». Six des douze articles du numéro 45 de *LINX*, intitulé « Invariants et variables dans les langues. Études typologiques », s'y trouvent également rassemblés, et les textes restants sont globalement dans une démarche typologique, ou tout au moins systématique.

♦ La classe 9 semble se caractériser par un usage important du passé composé et des formes passives au présent (auxiliaires présent), assorti à un usage plus restreint du présent et du futur. Le pronom *nous* est privilégié à *je* et *on*, et on y relève de nombreux anaphoriques pluriel. Le pluriel (noms, adjectifs et amalgames) y est d'ailleurs sensiblement plus représenté que le singulier. Fait intéressant, la classe comprend un tiers des onze articles co-écrits (11 textes sur 31) : l'usage du *nous* combiné au passé composé serait ainsi, entre autres caractéristiques, plus spécifique à ces textes.

A l'instar des articles de Laurendeau et de Combettes, les textes de Bergounioux et de Chevrot sont regroupés au sein de la même classe 9, ce qui semble confirmer la présence de styles d'auteur dans le genre de l'article. En outre, on y trouve presque un tiers des textes de la revue *Verbum* dont nous disposons (11 textes sur 36), 5 articles sur 11 du numéro 2 de la revue *Syntaxe et Sémantique* (« La sémantique du lexique verbal ») et 4 textes sur 11 du numéro 42 de *LINX* (« Approches sociolinguistiques du plan phonique »). Certains numéros thématiques, et certaines revues semblent ainsi posséder des caractéristiques morphosyntaxiques communes qui permettent de regrouper leurs contributeurs par classification.

On remarque qu'une proportion significative des articles de la classe est dédiée à la phonétique/phonologie : quatre textes, soit un tiers du numéro thématique « Approches sociolinguistiques du plan phonique » (LINX 42) s'y trouvent rassemblés, de même que deux textes du numéro 24-01 de *Verbum* « Y a-t-il une syntaxe au-delà de la phrase ? » se rapportant à la prosodie et un texte de Didelot-Zerr qui s'attache également à la prosodie et la phonétique.

Les textes de la classe 9 semblent globalement plus appliqués : plus de la moitié des textes contient des exemples, et la plupart des textes qui n'en comportent pas incluent des tableaux de données ou des graphiques commentés. On ne relève que trois textes qui comprennent un nombre important de citations :



Graphique : Eléments constitutifs des textes de la classe 9

Le passé composé semble ainsi typique des textes plus appliqués et des comptes rendus d'expérimentation :

> Nous avons cherché des exemples d'anaphoriques proximaux et démonstratifs indirects dans de l'écrit planifié, mais n'en avons pas trouvé. (174)
>
> Dans l'étape d'implantation des données d'ULSID à l'ICP, nous avons harmonisé la transcription entre les lexiques, en adoptant les symboles conventionnels de l'API (1996) et en conservant l'ensemble des informations sur le découpage syllabique. (101)

♦ La classe 10 fédère les textes contenant un usage important de sigles combiné à un emploi restreint de subordonnants et de ponctuations – essentiellement les parenthèses. Etant donné le caractère peu significatif des sigles, la classe 10 contient un ensemble de textes pour le moins disparates : ainsi y trouve-t-on trois articles appliqués du numéro non thématique 7 de la *Revue de Sémantique et Pragmatique* et trois textes du volume 33 des *Cahiers de Praxématique*, regroupant des textes plus théoriques articulés autour de la « Sémantique de l'intertexte ». On relève toutefois quelques régularités : l'ensemble des articles du numéro thématique « La langue des signes : enjeux institutionnels et linguistiques » s'y trouve en effet rassemblé, du fait de son emploi important des sigles « LSF » et « LS ». A fortiori, la classe 10 contient les deux textes de Siblot dont nous disposons, ce qui souligne encore une fois la présence de styles personnels au sein du genre de l'article.

♦ Les classes 1 et 5 sont les deux partitions les plus importantes mises au jour : elles rassemblent respectivement 34 et 49 individus, soit plus d'un tiers du corpus. En ce sens, elles sont plus délicates à saisir – la classe 1 nous semble particulièrement peu interprétable. Au regard de la classe 5, un quart du corpus total comprendrait des connecteurs d'exemplification, des dates et des abréviations, de même que des symboles linguistiques, ce qui nous semble caractéristique des textes se rapportant à la norme linguistique. De manière

générale, la démarche serait peu quantitative (peu de numéraux en général et de cardinaux en particulier) et les proportions tenues par la première personne du pluriel seraient peu significatives.

Synthèse

Il convient d'abord de souligner que certaines variables semblent plus sensibles que d'autres en termes de variation. Les proportions de pronoms personnels semblent ainsi particulièrement stabilisées au sein du genre. On voit en effet à quel point les différences de répartition de ces marqueurs influencent le partitionnement du corpus : si les textes 2 et 117 ont été en grande partie isolés du fait de la proportion importante de marques de seconde personne qu'ils contiennent, les pronoms personnels de première personne jouent un rôle non moins majeur dans la formation des classes 2, 8 et 9.

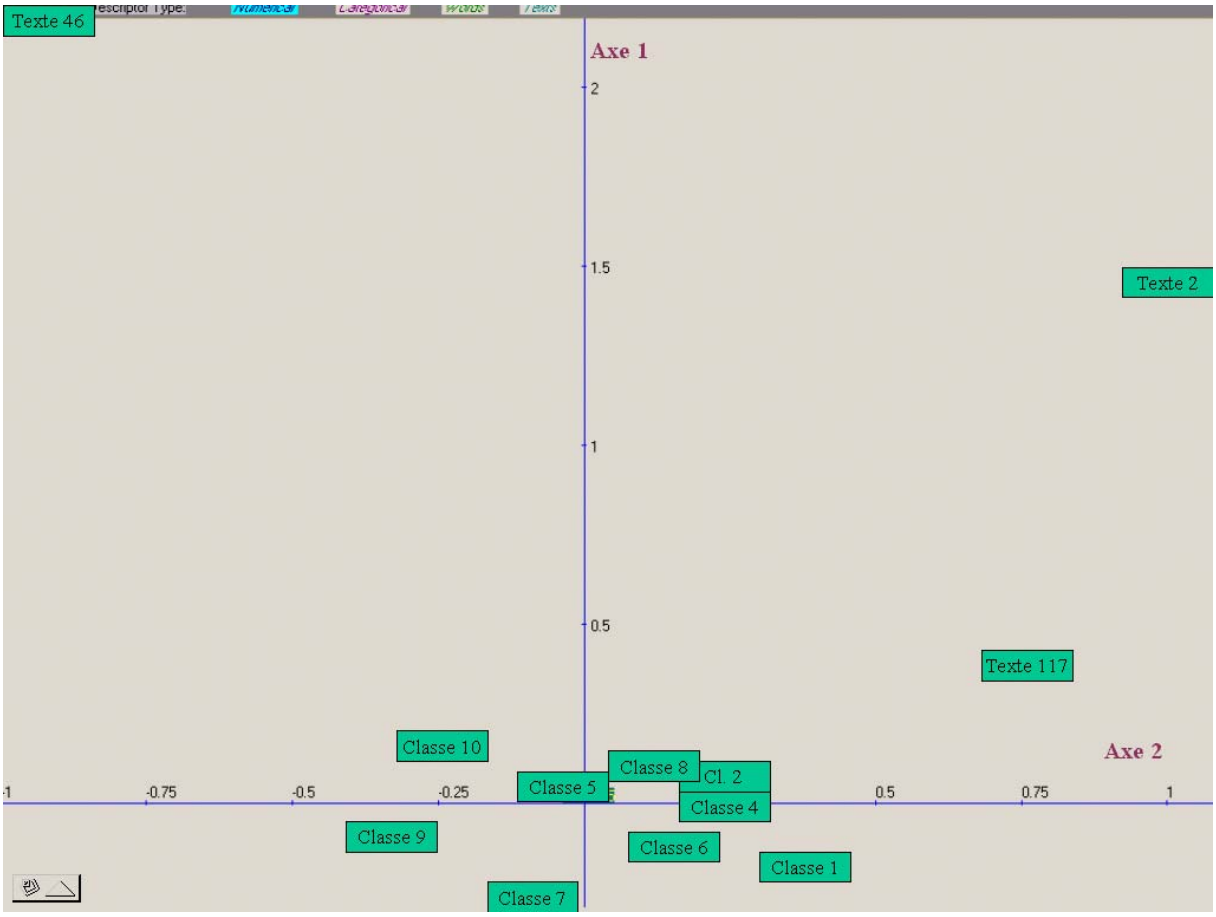
Il semble en aller de même de certains temps verbaux : les textes plus spéculatifs, qui se singulariseraient par un usage important du conditionnel, se trouvent ainsi isolés.

Enfin, on a pu constater que les descripteurs morphosyntaxiques regroupaient souvent les textes de même auteur, de même numéro thématique et de même revue, stabilité qui demeure à confirmer à partir de corpus plus adaptés : les variations stylistiques et domaniales seront ainsi éprouvées par la suite (partie analytique II), ce qui nous permettra de mettre au jour les descripteurs qui pourraient spécifiquement représenter des axes de variation générique.

3.5.3.2. Position des partitions sur les axes factoriels

Les douze classes mises au jour reflètent bien les groupements intercorrélés de variables précédemment observés.

Le graphique suivant illustre d'abord les positions marginales des trois textes limites : le texte historique 46 se situe sur le versant positif de l'axe 1, et négatif de l'axe 2, c'est-à-dire au niveau des caractéristiques narratives des textes historiques, tandis que les textes 2 et 117 sont plus proches, dans la mesure où leurs spécificités sont essentiellement liées à un usage singulier des marques personnelles :

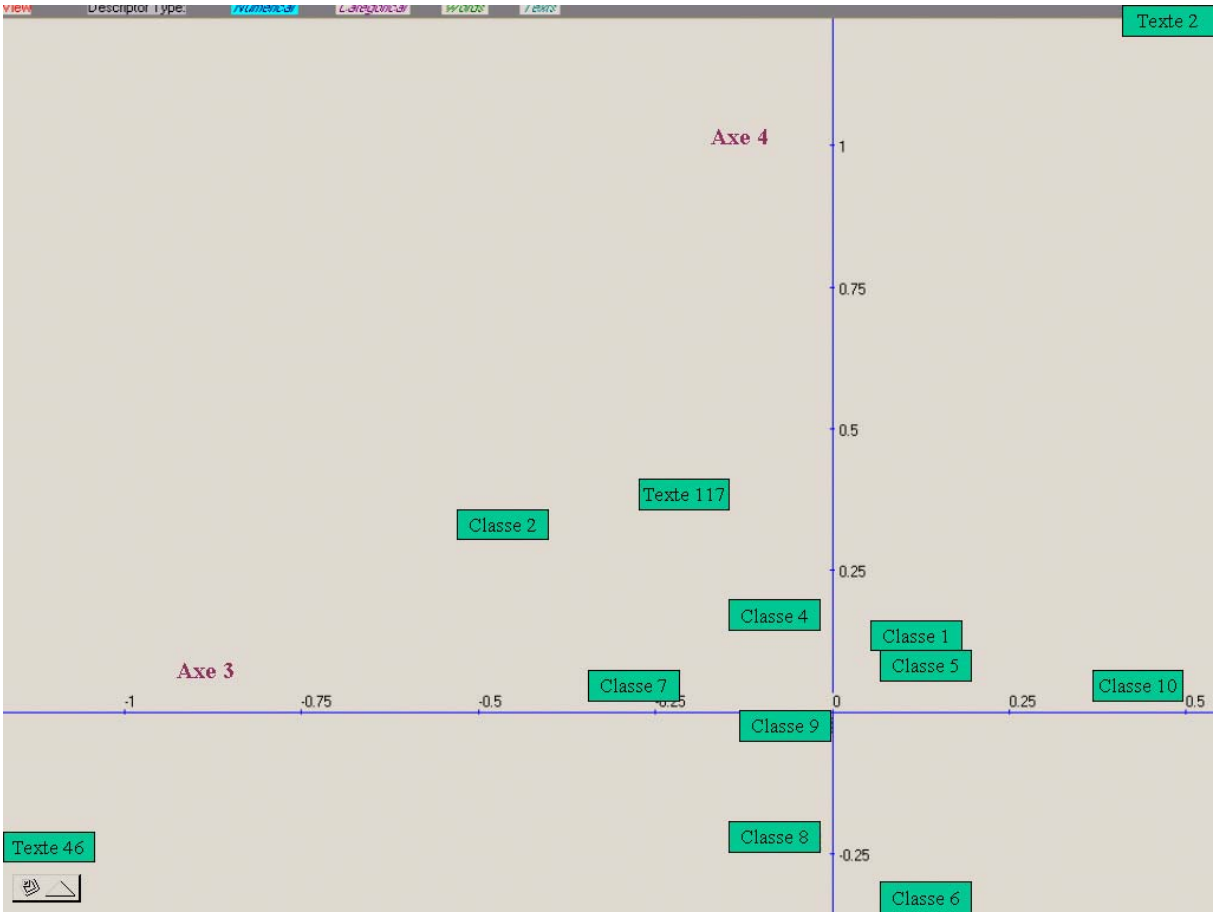


Graphique : positionnement des douze classes sur les axes factoriels 1 et 2

La classe 8, qui nous avait semblé relever essentiellement d'un emploi important des marques de première personne du pluriel ne se positionne pas au niveau de ces descripteurs : la présence de nombreux éléments au singulier dans cette classe la positionne davantage du côté du singulier, c'est-à-dire sur le versant positif de l'axe 1. Les classes 2 et 9 se positionnent par contre au niveau des marqueurs de première personne du singulier et du pluriel.

La classe 4 se situe bien au niveau du conditionnel, ce qui est plus visible encore au regard des axes factoriels 3 et 4.

La classe 5, qui contient presque un quart du corpus, est particulièrement centrée sur le premier graphique : on peut penser que ce regroupement correspond, avec la classe 1, au *noyau* dur du genre.



Graphique : positionnement des douze classes obtenues sur les axes factoriels 3 et 4

3.5.4. Approfondissement de la structure générique par « peeling »

L'analyse factorielle et la CAH menées dans le sections précédentes ayant fait apparaître des éléments frontaliers (textes 2 et 46) entraînant un centrage de l'ensemble des individus et entravant finalement l'observation de la structure générique, nous avons dans un deuxième temps de l'analyse procédé à un *peeling*, c'est-à-dire à un « épluchage » progressif des points aberrants.

Si la technique est couramment employée en analyse de données statistiques, elle nous a d'abord semblé problématique, dans la mesure où nous avons cherché à préserver l'intégrité du corpus ; cette question nous semble d'ailleurs représenter l'objet d'un éventuel débat en linguistique de corpus : faut-il procéder d'emblée à une extraction des textes frontaliers avant d'engager l'analyse ? Quel statut doit-on accorder aux individus 'aberrants' d'un corpus ? Cette procédure n'ouvrirait-elle finalement pas la porte à d'autres types d'extractions, fondées sur d'autres critères ?

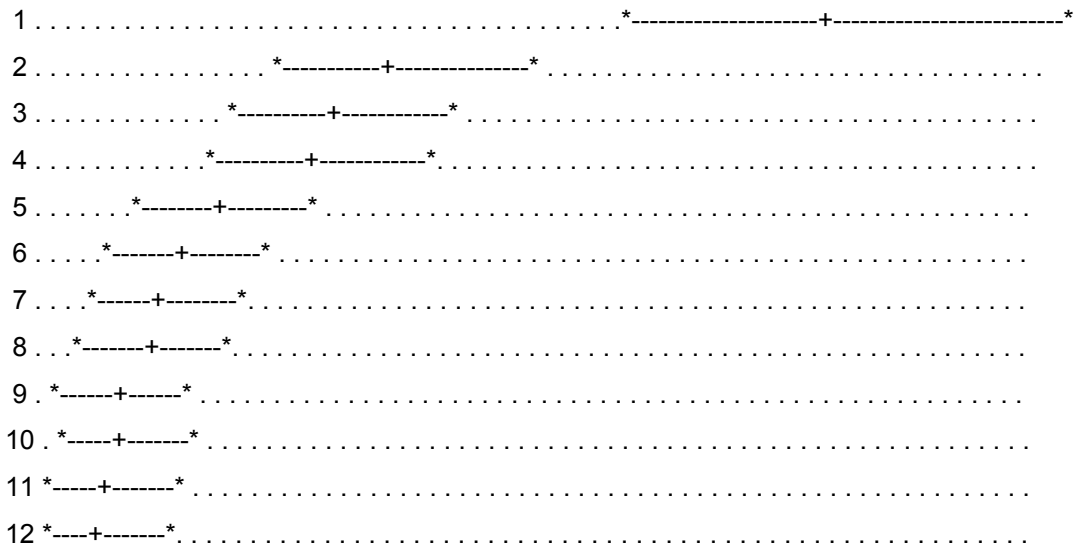
Dans la mesure où notre objectif est à la fois descriptif d'un corpus et d'un genre, puisque le premier est supposé *représentatif* du second, nous avons finalement choisi de restreindre le premier afin d'appréhender le second de manière plus précise : une seconde ACP a ainsi été

menée après suppression des individus 2 et 46, dont les caractéristiques ont été restituées *supra*.

Le nouveau diagramme des valeurs propres obtenu diffère globalement peu du précédent tandis que la première valeur propre demeure toujours la plus significativement distincte :

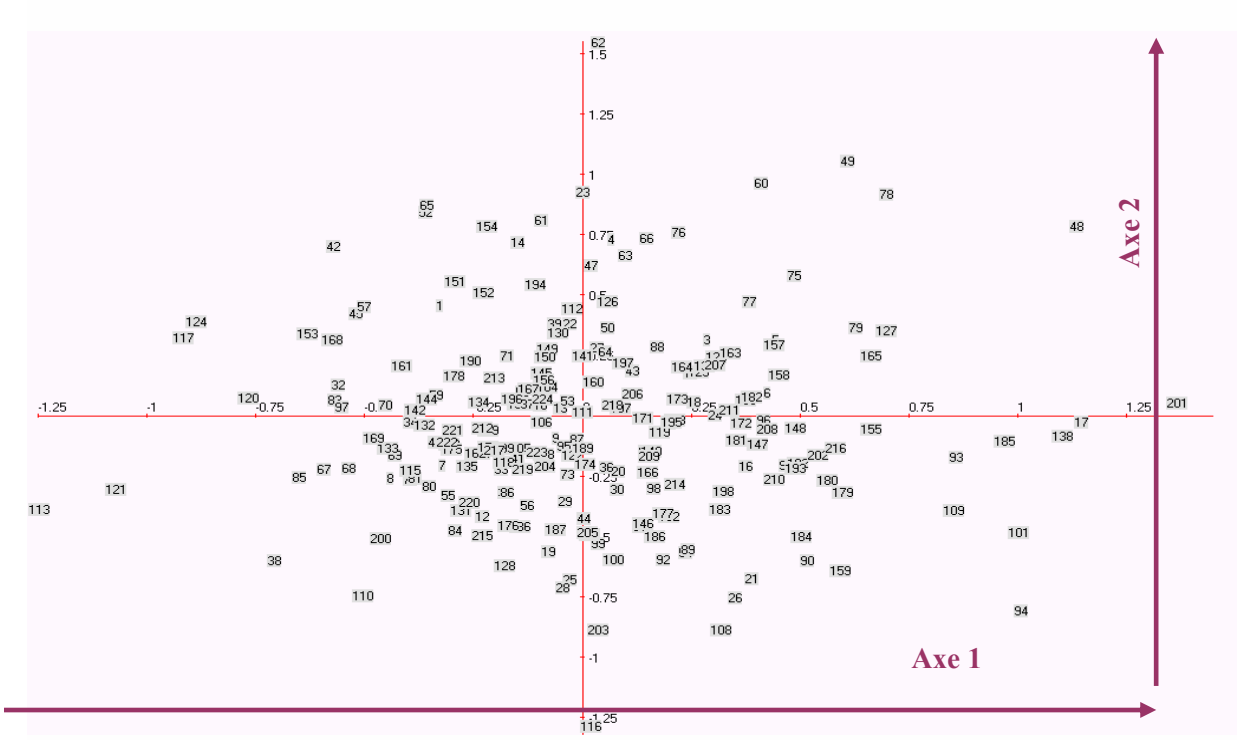
Nb	Valeur propre	% d' inertie	% cumulé	
1	10.59	7.62	7.62	**********
2	5.99	4.31	11.93	**********
3	5.38	3.87	15.81	**********
4	5.17	3.72	19.53	**********
5	4.13	2.98	22.51	**********

Tableau : Diagramme des 5 premières valeurs propres (sortie DTM)



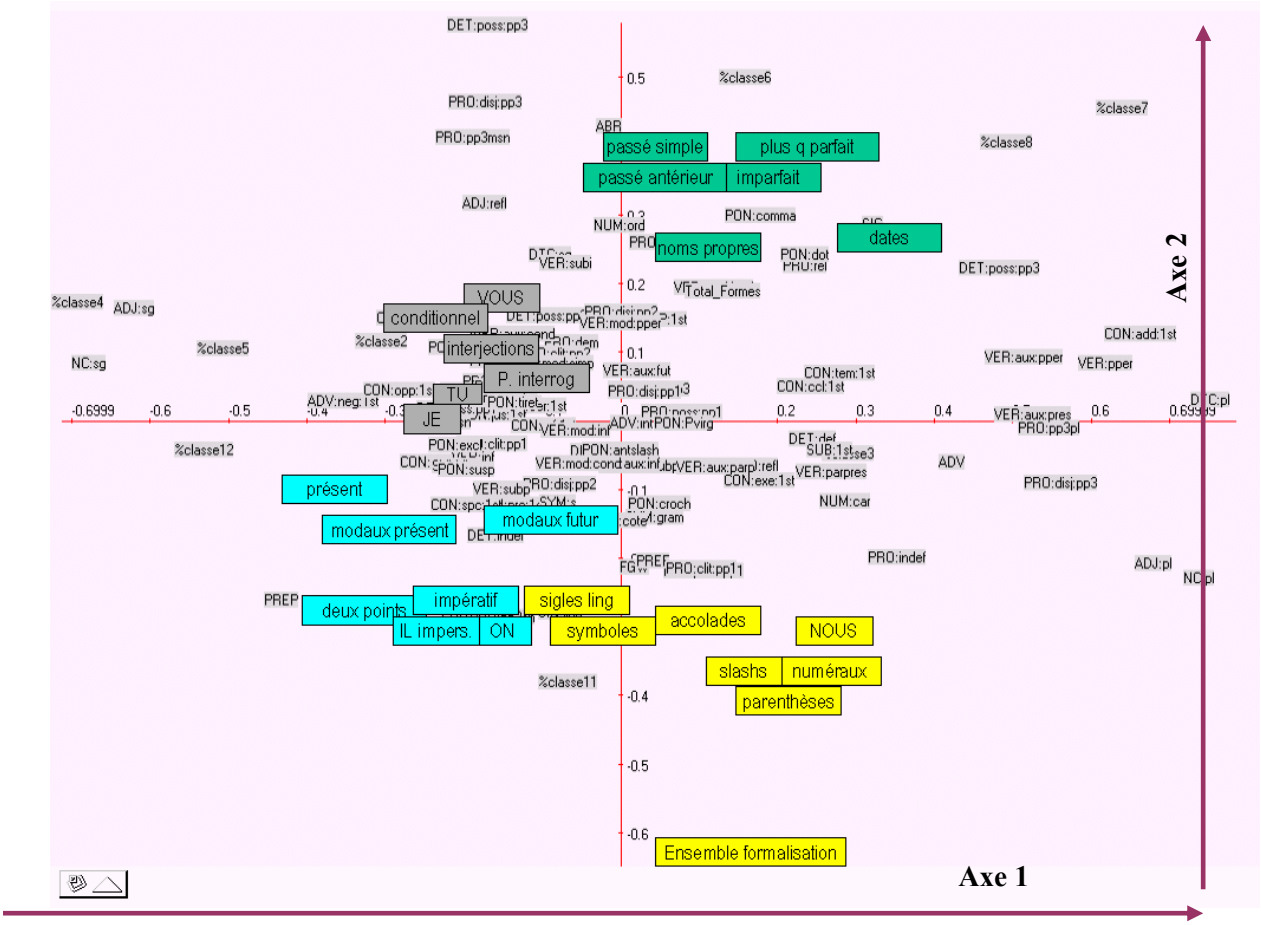
Graphique : Position relative des intervalles

Si l'on observe le premier plan factoriel obtenu, le nuage des individus obtenu est significativement plus observable :



Graphique : Positionnement des individus sur les deux premiers axes factoriels

Malgré quelques changements, les quatre pôles mis au jour demeurent (fort heureusement) stabilisés, ce qui valide – ou du moins n'infirme pas – les axes d'organisation générique mis au jour dans les sections précédentes :



Graphique : Positionnement des variables sur les deux premiers axes factoriels

Etant donné qu'il est statistiquement plus pertinent, c'est d'abord sur ce plan factoriel que seront fondées les comparaisons dans le cadre des parties contrastives à venir (styles, genres et domaines, langues, *v. supra*).

3.5.5. Variables illustratives supplémentaires

Nous observerons enfin les variables typologiques supplémentaires que nous avions à titre d'hypothèse associées aux textes du corpus, à savoir :

- ✓ le sexe de l'auteur ;
- ✓ le numéro de revue et la revue de provenance ;
- ✓ l'année de publication.

Rappelons que l'auteur est neutralisé dans le corpus ASLF, et fera l'objet d'une analyse spécifique dans le chapitre 7.

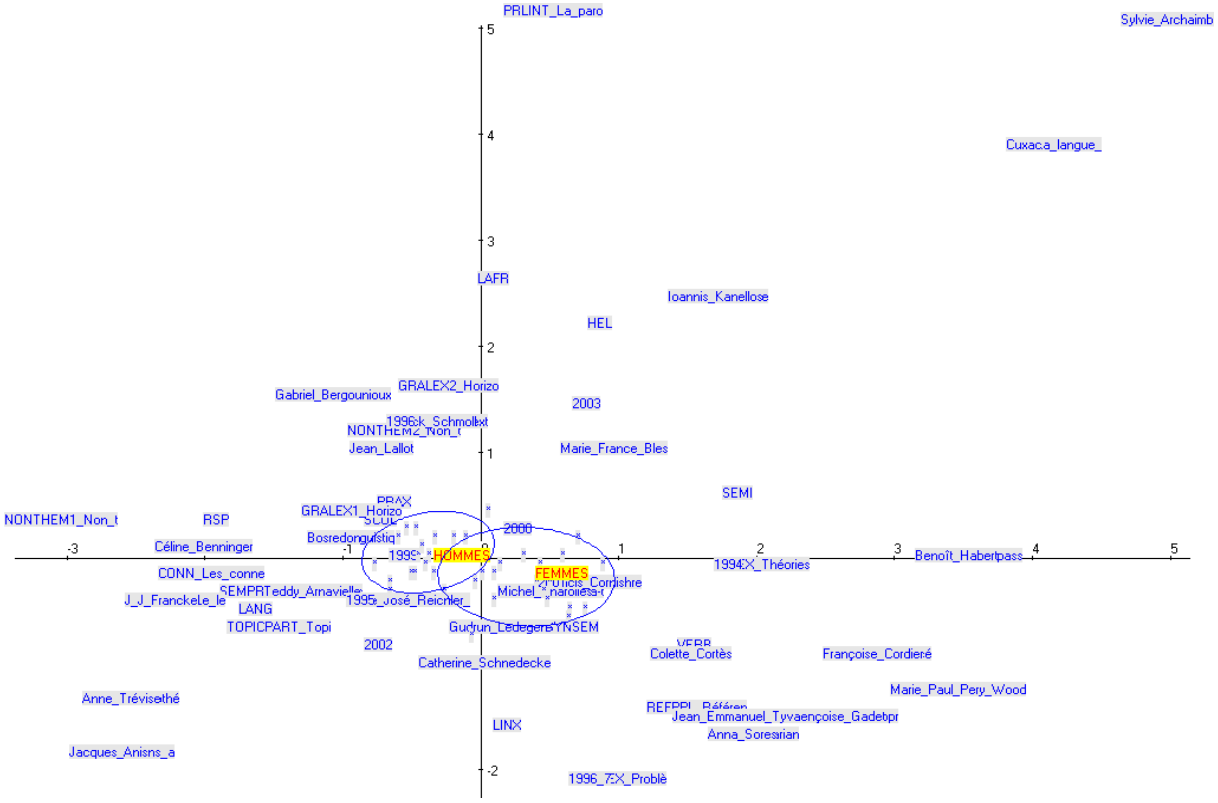
Les analyses présentées ci-après mobilisent la technique du *bootstrap* (Lebart *et al.*, 2003 : 205-210), que nous seront amenée à utiliser par la suite.

3.5.5.1. Sexe

La répartition des sexes dans le corpus est globalement équilibrée : 121 hommes pour 103 femmes. Soulignons que nous n'avons pas construit le corpus en prenant cette variable en

considération.

Observons la position des deux modalités homme et femme de la variable sexe sur le premier plan factoriel (corpus après peeling) :

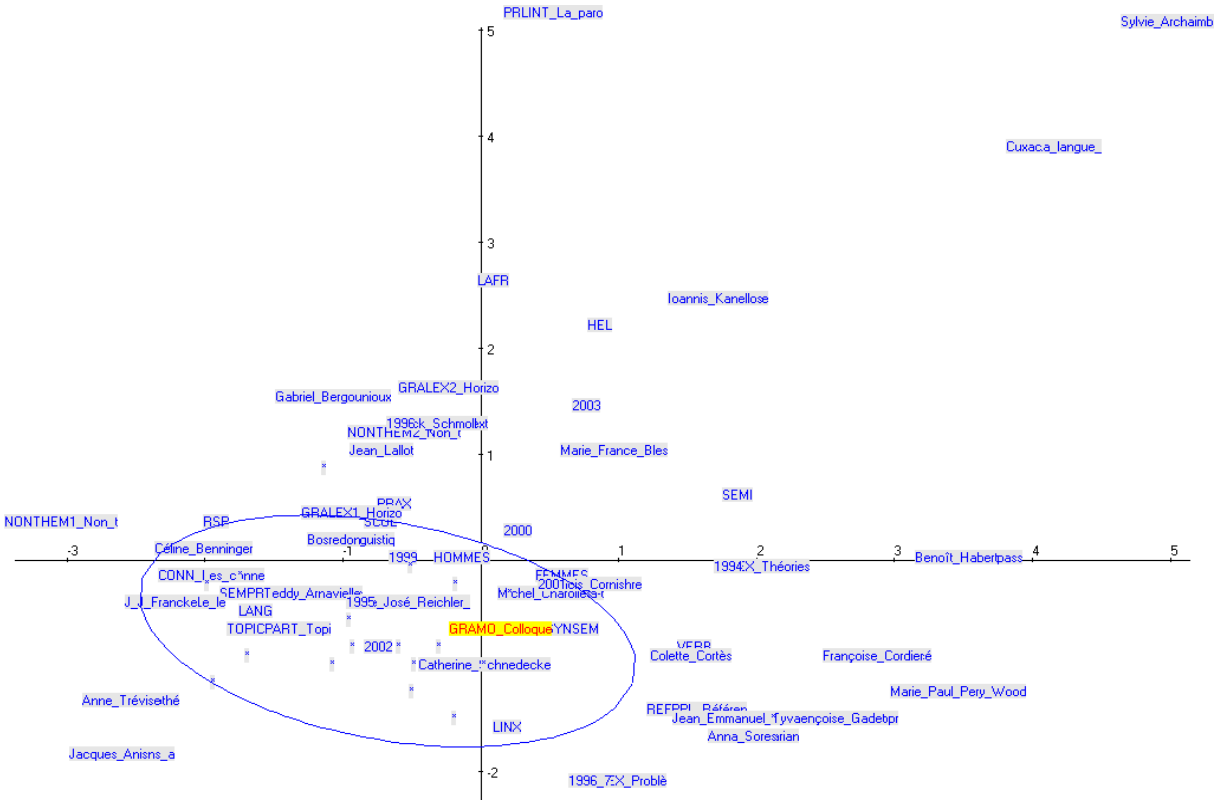


Graphique : Ellipses de confiance autour de la variable sexe – premier plan factoriel

Hommes et femmes sont globalement centrés, ce qui indique une incidence très faible, sinon absente de la variable sexe sur les caractéristiques des textes. On observe à la limite que les femmes sont plus orientées vers le pôle formel.

3.5.5.2. Revue, et numéro thématique

Observons d'abord le numéro thématique 31 constitué d'actes, et non d'articles :

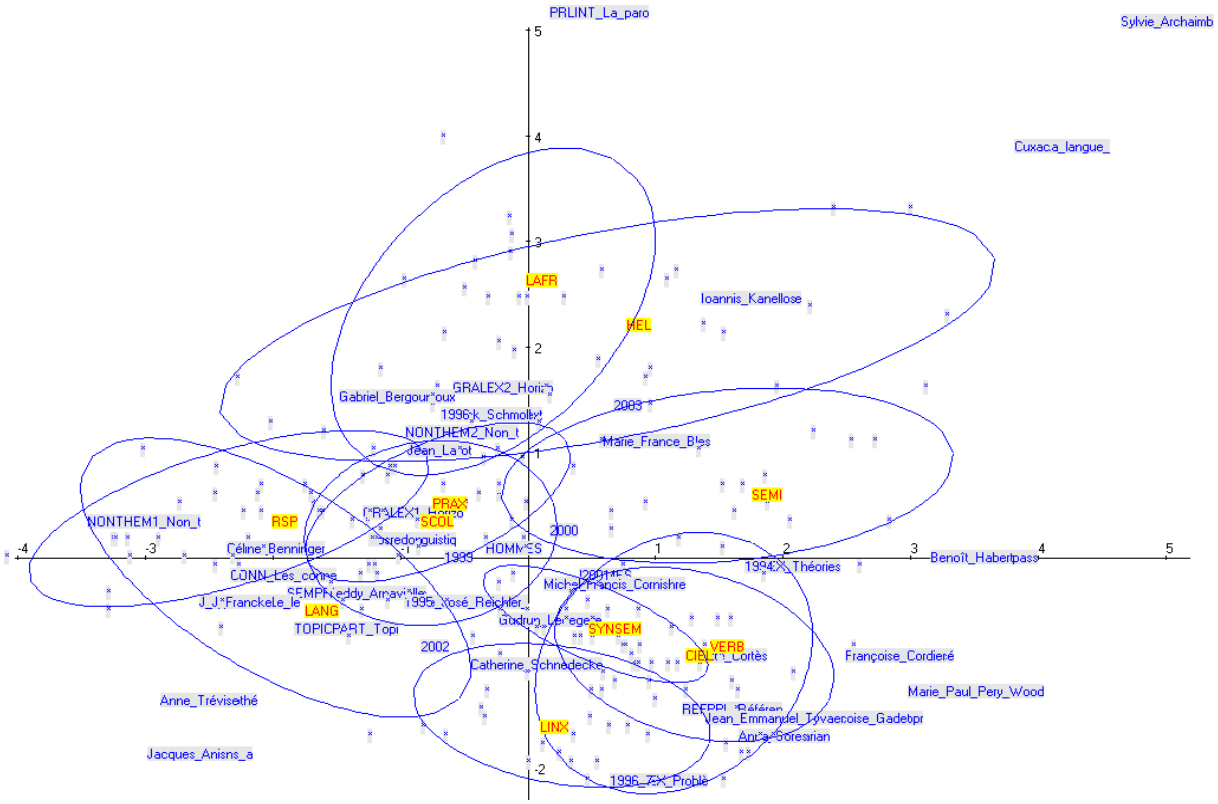


Graphique : Ellipse de confiance autour du numéro thématique 31

La revue est globalement centrée, et n'est en aucun cas isolée, ce qui illustre bien la grande proximité des deux genres.

Bien que leur situation sur les axes les associent aux orientations mises au jour, l'ensemble des numéros thématiques et des revues n'apparaissent pas comme significativement distincts sur le premier plan factoriel : les ellipses sont généralement grandes et se recouvrent largement.

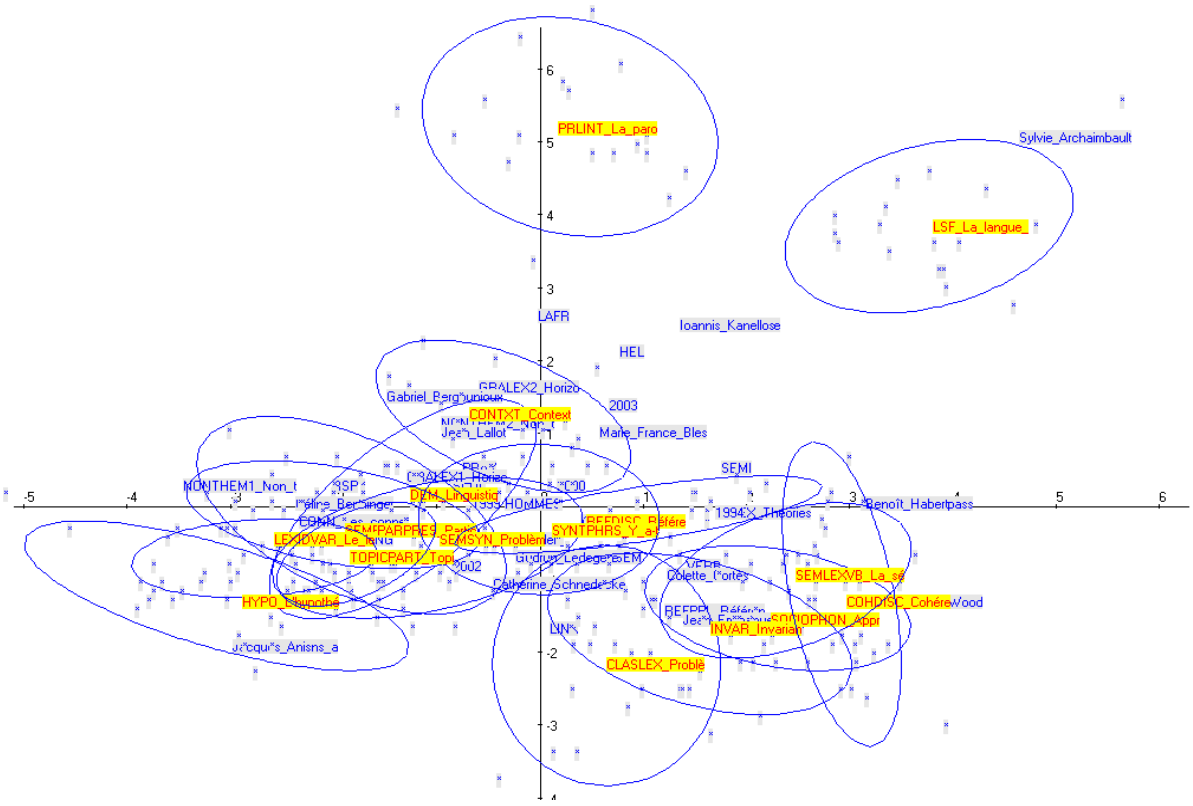
Aucune revue n'est ainsi significativement distincte (ellipse non recouverte) :



Graphique : Ellipses de confiance autour des 11 revues de linguistique

En revanche, on observe d'importantes différences entre les revues : la revue *Langue française* est ainsi très significativement distincte des revues *CIEL, LINX* et *RSP*. *HEL* se distingue également nettement, tandis que *LINX* et *CIEL* sont indiscernables.

Les numéros thématiques se distinguent plus nettement encore : deux numéros thématiques de la revue Langue Française sont ainsi très significativement distincts : le numéro sur la Parole intérieure édité par Bergounioux et celui de Cuxac « La langue des signes : enjeux institutionnels et linguistiques » :

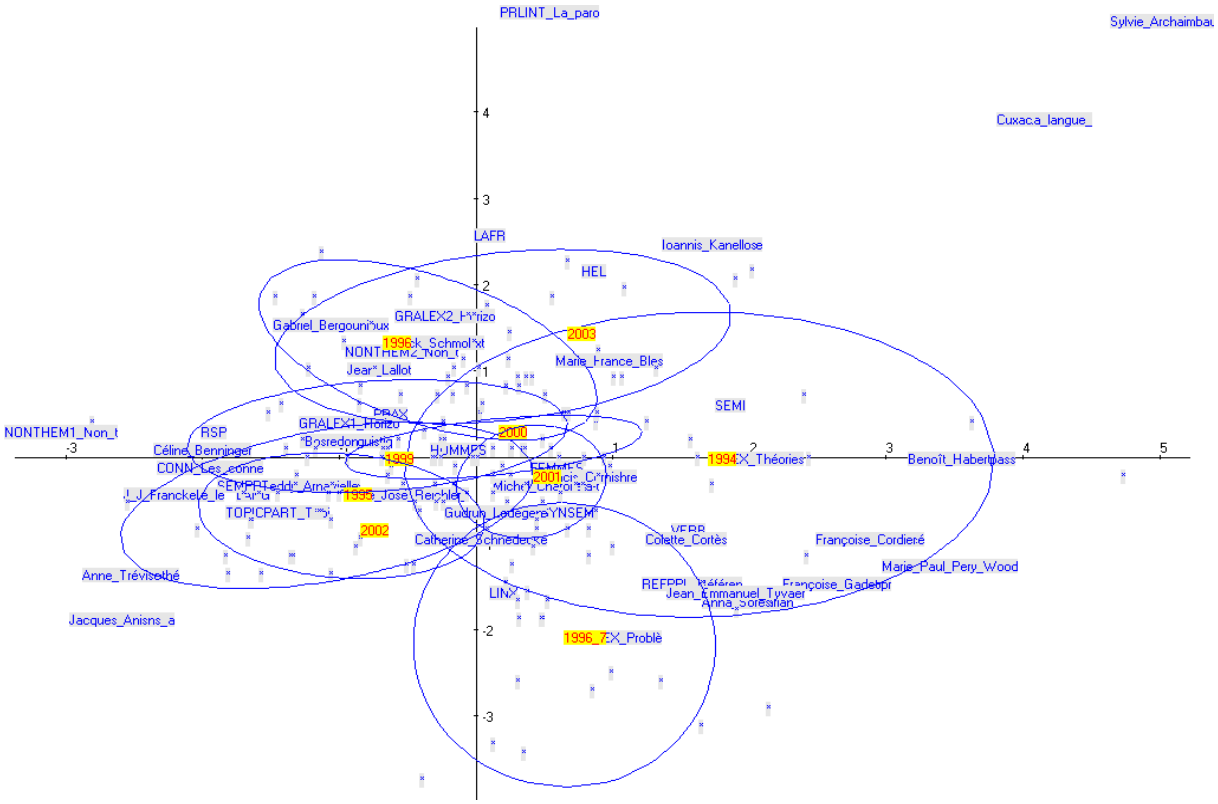


Graphique : Ellipses de confiance autour des numéros de revue

Les deux revues se situent ainsi sur le pôle le plus historique et le moins appliqué, ce qui correspond bien à leur orientation scientifique générale.

3.5.5.3. Année de publication

La variable année est moins pertinente encore, ce qui n'est pas surprenant étant donné notre effort synchronique – le contraire aurait été problématique :



Graphique : Ellipses de confiance autour de la variable année

3.6. Conclusion

Le présent chapitre a ainsi permis de dresser un profil morphosyntaxique général du genre de l'article : après avoir décrit les variables au sein de leur catégorie d'appartenance, en observant leur distribution et leurs corrélations, une ACP assortie d'une CAH nous a permis de mettre au jour certains axes d'organisation du genre, que nous n'aurions pu saisir par d'autres biais. Au final, le genre de l'article s'avère articulé autour de quatre pôles opposant le passé au présent et deux types d'orientations en linguistique : formalisation et exemplification (dimension interlocutive et plus orale).

La méthodologie adoptée est progressive : après avoir déterminé les caractéristiques générales du genre de l'article en le contrastant à trois autres genres et discours[35], nous avons observé les éléments les plus marginaux du genre de l'article. Dans un troisième temps de l'analyse, ce sont les catégories morphosyntaxiques générales qui ont été précisément observées : nous avons particulièrement mis l'accent sur les pronoms personnels en raison de leur plus grande interprétabilité – ils demeurent en effet et de manière non anodine les marqueurs les plus observés du discours scientifiques, bien que l'ensemble des descripteurs ait été précisément observé. On soulignera que certaines catégories morphosyntaxiques n'ont été observées qu'indirectement, à travers leurs corrélations avec les descripteurs approfondis : c'est le cas par exemple des prépositions, des adverbes et des subordonnants.

Comme nous l'avons déjà évoqué, nous avons été confrontée à une avalanche de données

[35] Rappelons que les données Cordial doivent être maniées avec précaution.

que nous avons tenté de gérer et de saisir au mieux : nous avons ainsi privilégié ce qui nous paraissait le plus interprétable, en demeurant dans un cadre objectivé. Par exemple, les représentations en déciles présentées ne résultent pas d'un choix arbitraire, dans la mesure où la totalité des graphiques a été observée et objectivée en amont : seules les configurations dessinant une forme ascendante ou descendante ont été présentées, et intégrées à nos analyses.

Les données et les résultats présentés peuvent ainsi paraître d'intérêt inégal : par exemple, le relevé systématique des marqueurs de liste de l'article est naturellement moins stimulant que celui des dates, qui donnent une approximation de la durée de vie des textes scientifiques – de manière incomplète car nous n'avons pas procédé à un relevé systématique des références citées, afin de distinguer les références qui font date des autres.

Le niveau morphosyntaxique, qui demeure dans les faits l'un des seuls niveaux opérationnels dont on dispose, nous a globalement paru constituer un point d'entrée descriptif pertinent qui, loin d'être restrictif, ouvre au contraire de nombreuses voies interprétatives, dans des cadres disciplinaires distincts. Le caractère systématique de nos analyses offre ainsi aux scoliastes futurs des lieux de comparaison qui nous ont fait défaut.

Bien que nous ayons abordé la séquentialité de l'article à travers l'analyse des déciles textuels de ses catégories, le présent chapitre est globalement demeuré au niveau du corps de texte : il convient dès lors d'examiner le genre de l'article en termes de structuration textuelle, et de sections.