Chapitre 9 : Genre et langues

Directions de recherche

L'analyse contrastive interlangue des genres est rarement menée sur le plan quantitatif, et *a fortiori* sur un plan morphosyntaxique global ; on recense ainsi au sein du courant ESP une multitude d'analyses génériques interlangues consacrées à l'observation d'un marqueur ou d'un phénomène linguistique local, mais aucune entreprise n'a à notre connaissance cherché à contraster les régulations globales d'un genre dans deux langues différentes, à partir d'un jeu de descripteurs équivalents.

La présente section est ainsi particulièrement exploratoire – et donc imparfaite : sa vocation première demeure de mesurer l'ampleur des problèmes posés par une telle démarche, qui pourra être poursuivie et affinée dans des travaux futurs.

Le corpus français a ainsi été confronté à la collection d'articles anglo-saxons (présentée chapitre 2), et en amont de l'analyse et de la détermination de lieux d'équivalence et d'opposition se pose le problème des descripteurs employés.

Les entreprises d'annotation étant développées dans des cadres monolingues, elles demeurent de manière générale peu adaptées aux traitements multilingues malgré la similitude, voire l'identité des systèmes d'annotation dans chaque langue. En effet, la mise en contraste des langues requiert un point de comparaison, un *tertium comparationis*, toutefois difficile à déterminer dans notre cas, le genre étant fondamentalement multidimensionnel.

On rencontre le plus souvent deux cas de figure en matière d'annotation de corpus multilingue (Neumann et Hansen-Schirra, 2003) : le corpus peut d'abord être divisé en deux sous-ensembles annotés séparément. Les problèmes de comparabilité sont ainsi repoussés et n'apparaissent qu'au moment de l'interprétation des résultats. La seconde méthode privilégie un sens de contraste en fondant l'annotation sur une langue, l'autre (les autres) devant être adaptée(s), ce qui pose souvent problème.

Etant donné qu'il est exclu, pour des raisons évidentes de temps et de coût[1], de reconduire la même expérience d'entraînement sur les textes anglais, c'est vers le second cas de figure que notre démarche s'orientera – soulignons que la première alternative, comparativement plus coûteuse, pose également des problèmes de comparabilité.

A fortiori, on recense un nombre plus important d'étiqueteurs de l'anglais que du français, ce qui élargit la gamme de choix et facilite la sélection d'un outil et d'un système d'étiquetage approchant sinon similaire : l'étiquetage trop robuste du *Penn TreeBank* devant être écarté, c'est finalement sur l'étiqueteur CLAWS[2] (*Constituent Likelihood Automatic Word-tagging System*, UCREL, Lancaster) que s'est arrêté notre choix.

Développé à partir de l'étiquetage manuel du corpus Brown par G. Leech et son équipe, le système d'étiquetage, fondé sur des règles (*rule-based*), est en constant développement depuis

---

[1] Rappelons que l'ensemble de la procédure a demandé six mois de développement et de corrections.

[2] http://www.comp.lancs.ac.uk/ucrel/claws/

1981 ; la version actuelle (CLAWS7) propose 137 étiquettes, si l'on écarte les tags positionnels[3] :

TAG	Descriptif	TAG	Descriptif	TAG	Descriptif
APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)	**NN2**	plural common noun (e.g. books, girls)	**RPK**	prep. adv., catenative (about in be about to)
AT	article (e.g. the, no)	**NNA**	following noun of title (e.g. M.A.)	**RR**	general adverb
AT1	singular article (e.g. a, an, every)	**NNB**	preceding noun of title (e.g. Mr., Prof.)	**RRQ**	wh- general adverb (where, when, why, how)
BCL	before-clause marker (e.g. in order (that),in order (to))	**NNL1**	singular locative noun (e.g. Island, Street)	**RRQV**	wh-ever general adverb (wherever, whenever)
CC	coordinating conjunction (e.g. and, or)	**NNL2**	plural locative noun (e.g. Islands, Streets)	**RRR**	comparative general adverb (e.g. better, longer)
CCB	adversative coordinating conjunction (but)	**NNO**	numeral noun, neutral for number (e.g. dozen, hundred)	**RRT**	superlative general adverb (e.g. best, longest)
CS	subordinating conjunction (e.g. if, because, unless, so, for)	**NNO2**	numeral noun, plural (e.g. hundreds, thousands)	**RT**	quasi-nominal adverb of time (e.g. now, tomorrow)
CSA	as (as conjunction)	**NNT1**	temporal noun, singular (e.g. day, week, year)	**TO**	infinitive marker (to)
CSN	than (as conjunction)	**NNT2**	temporal noun, plural (e.g. days, weeks, years)	**UH**	interjection (e.g. oh, yes, um)
CST	that (as conjunction)	**NNU**	unit of measurement, neutral for number (e.g. in, cc)	**VB0**	be, base form (finite i.e. imperative, subjunctive)
CSW	whether (as conjunction)	**NNU1**	singular unit of measurement (e.g. inch, centimetre)	**VBDR**	were
DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)	**NNU2**	plural unit of measurement (e.g. ins., feet)	**VBDZ**	was
DA1	singular after-determiner (e.g. little, much)	**NP**	proper noun, neutral for number (e.g. IBM, Andes)	**VBG**	being
DA2	plural after-determiner (e.g. few, several, many)	**NP1**	singular proper noun (e.g. London, Jane, Frederick)	**VBI**	be, infinitive (To be or not... It will be ..)
DAR	comparative after-determiner (e.g. more, less, fewer)	**NP2**	plural proper noun (e.g. Browns, Reagans, Koreas)	**VBM**	am
DAT	superlative after-determiner (e.g. most, least, fewest)	**NPD1**	singular weekday noun (e.g. Sunday)	**VBN**	been
DB	before determiner or pre-determiner capable of pronominal function (all, half)	**NPD2**	plural weekday noun (e.g. Sundays)	**VBR**	are
DB2	plural before-determiner (both)	**NPM1**	singular month noun (e.g. October)	**VBZ**	is
DD	determiner (capable of pronominal function) (e.g any, some)	**NPM2**	plural month noun (e.g. Octobers)	**VD0**	do, base form (finite)
DD1	singular determiner (e.g.	**PN**	indefinite pronoun,	**VDD**	did

[3] De la même manière que nous avons pris en compte certaines locutions et expressions (e.g. les connecteurs *par exemple* ou *d'autre part*), CLAWS annote de nombreuses séquences de type *for example* ou *no matter how*, respectivement étiquetées *for*_REX21 *example*_REX22 et *no*_RGQV31 *matter*_RGQV32 *how*_RGQV33.

	this, that, another)		neutral for number (none)		
DD2	plural determiner (these,those)	**PN1**	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)	**VDG**	doing
DDQ	wh-determiner (which, what)	**PNQO**	objective wh-pronoun (whom)	**VDI**	do, infinitive (I may do... To do...)
DDQGE	wh-determiner, genitive (whose)	**PNQS**	subjective wh-pronoun (who)	**VDN**	done
DDQV	wh-ever determiner, (whichever, whatever)	**PNQV**	wh-ever pronoun (whoever)	**VDZ**	does
EX	existential there	**PNX1**	reflexive indefinite pronoun (oneself)	**VH0**	have, base form (finite)
FO	formula	**PPGE**	nominal possessive personal pronoun (e.g. mine, yours)	**VHD**	had (past tense)
FU	unclassified word	**PPH1**	3rd person sing. neuter personal pronoun (it)	**VHG**	having
FW	foreign word	**PPHO1**	3rd person sing. objective personal pronoun (him, her)	**VHI**	have, infinitive
GE	germanic genitive marker - (' or's)	**PPHO2**	3rd person plural objective personal pronoun (them)	**VHN**	had (past participle)
IF	for (as preposition)	**PPHS1**	3rd person sing. subjective personal pronoun (he, she)	**VHZ**	has
II	general preposition	**PPHS2**	3rd person plural subjective personal pronoun (they)	**VM**	modal auxiliary (can, will, would, etc.)
IO	of (as preposition)	**PPIO1**	1st person sing. objective personal pronoun (me)	**VMK**	modal catenative (ought, used)
IW	with, without (as prepositions)	**PPIO2**	1st person plural objective personal pronoun (us)	**VV0**	base form of lexical verb (e.g. give, work)
JJ	general adjective	**PPIS1**	1st person sing. subjective personal pronoun (I)	**VVD**	past tense of lexical verb (e.g. gave, worked)
JJR	general comparative adjective (e.g. older, better, stronger)	**PPIS2**	1st person plural subjective personal pronoun (we)	**VVG**	-ing participle of lexical verb (e.g. giving, working)
JJT	general superlative adjective (e.g. oldest, best, strongest)	**PPX1**	singular reflexive personal pronoun (e.g. yourself, itself)	**VVGK**	-ing participle catenative (going in be going to)
JK	catenative adjective (able in be able to, willing in be willing to)	**PPX2**	plural reflexive personal pronoun (e.g. yourselves, themselves)	**VVI**	infinitive (e.g. to give... It will work...)
MC	cardinal number, neutral for number (two, three..)	**PPY**	2nd person personal pronoun (you)	**VVN**	past participle of lexical verb (e.g. given, worked)
MC1	singular cardinal number (one)	**RA**	adverb, after nominal head (e.g. else, galore)	**VVNK**	past participle catenative (e.g. bound in be bound to)
MC2	plural cardinal number (e.g. sixes, sevens)	**REX**	adverb introducing appositional constructions (namely, e.g.)	**VVZ**	-s form of lexical verb (e.g. gives, works)
MCGE	genitive cardinal number, neutral for number (two's, 100's)	**RG**	degree adverb (very, so, too)	**XX**	not, n't
MCMC	hyphenated number (40-50, 1770-1827)	**RGQ**	wh- degree adverb (how)	**ZZ1**	singular letter of the alphabet (e.g. A,b)
MD	ordinal number (e.g. first, second, next, last)	**RGQV**	wh-ever degree adverb (however)	**ZZ2**	plural letter of the alphabet (e.g. A's, b's)
MF	fraction, neutral for number (e.g. quarters, two-thirds)	**RGR**	comparative degree adverb (more, less)		
ND1	singular noun of direction (e.g. north, southeast)	**RGT**	superlative degree adverb (most, least)		

NN	common noun, neutral for number (e.g. sheep, cod, headquarters)	**RL**	locative adverb (e.g. alongside, forward)	
NN1	singular common noun (e.g. book, girl)	**RP**	prep. adverb, particle (e.g about, in)	

Tableau: UCREL CLAWS7 Tagset

(http://www.comp.lancs.ac.uk/ucrel/claws7tags.html)

Bien que le jeu d'étiquettes de CLAWS ait une granularité élevée, il diffère substantiellement de celui que nous avons développé, ce qui entrave l'entreprise comparative en limitant les points de comparaison possibles, eux-mêmes limités par les spécificités du jeu d'étiquettes élaboré pour le français, construit indépendamment d'une comparaison interlangue ultérieure[4]. Idéalement, il conviendrait de développer un système d'étiquetage en vue d'une comparaison interlangue, mais cette entreprise constituerait elle-même l'objet d'une étude considérable, voire d'un autre doctorat.

Si l'on s'en tient à notre objectif plus modeste d'adaptation, on observe de nombreux tags qui présentent peu, voire aucun intérêt pour la description du discours scientifique (*e.g.* catégorisation sémantique des jours de la semaine, des mois ou des noms locatifs de type *street*). *A fortiori*, les temps verbaux ne sont pas nettement distingués : les temps du futur et du conditionnel ne sont ainsi pas isolés en l'état, *will ('ll, wo* – forme négative) et *would ('d, should, could*) étant regroupés avec l'ensemble des auxiliaires modaux au sein de la catégorie **VM**.

Bien qu'il soit de toute manière peu envisageable de constituer des jeux d'étiquettes parfaitement équivalents[5], certaines adaptations doivent être considérées afin d'améliorer sinon de forcer les points d'équivalence. Cette notion d'*équivalence* est loin d'être évidente, dans la mesure où elle n'est pas traductologique : notre corpus d'étude n'est pas un bitexte aligné de segments de type *original en langue A / traduction en langue B* – qui serait d'ailleurs difficile à rassembler, surtout si l'on souhaite comme nous conserver une direction linguistique français → anglais : les articles français ne sont en effet jamais, sauf très rares exceptions, traduits en anglais. Dans cette perspective, il ne s'agit pas de contraster l'expression morphosyntaxique d'un rendu identique dans deux langues, mais plutôt d'observer les points d'équilibrage morphosyntaxique du genre dans chaque langue, et de déterminer si les descripteurs utilisés dessinent des structures génériques stabilisées – qu'il serait alors pensable de contraster.

L'entreprise est délicate et ne peut méthodologiquement et en essence s'apparenter aux études contrastives mises en œuvre précédemment (*v.* chapitres 7 et 8), dans la mesure où les genres et les discours ne sauraient être confondus avec les langues, qui s'organisent en systèmes linguistiques distincts – bien que non hermétiques.

Si l'application du système de descripteurs au domaine – pour le moins éloigné – de la mécanique a montré l'inadéquation d'1/5[e] des descripteurs (*v.* chap. 8), il s'avère évidemment impensable d'étiqueter un texte anglais avec un système d'étiquetage du français et le passage

[4] La comparaison interlangue était présente au début de l'entreprise, mais cette dernière étant d'abord fondée sur le français, nous n'avons pas réduit le système de descripteurs élaboré dans le chapitre 2 à cet axe contrastif afin de garantir une pertinence descriptive maximale de la langue et du genre de l'article français.

[5] Les langues anglaise et française constituent bien entendu des systèmes linguistiques fondamentalement différents qu'il serait naïf de mettre totalement en parallèle au niveau morphosyntaxique – au mieux pourrait-on envisager des catégories d'équivalences sémantiques, mais ce n'est pas notre propos.

d'une langue à l'autre nécessite un développement bien plus considérable que celui d'un discours ou d'un genre à l'autre. Cette observation peut paraître triviale mais elle nous semble illustrer certains aspects du problème posé : bien que ce soit les structures organisationnelles du genre qui nous intéresse et que nous cherchons à contraster, comme nous l'avons fait d'un genre, d'un domaine ou d'un style à l'autre, l'équivalence des catégories entre elles doit également être pensée afin d'atteindre le premier objectif.

Dans cette perspective et étant donné que le point d'entrée linguistique adopté est morphosyntaxique et centré sur l'unité-mot ou *token*, les équivalences doivent être mesurées au niveau catégoriel, ce qui est hautement problématique, le français et l'anglais ayant une morphosyntaxe bien évidemment distincte. Dans certains cas, et lorsque les catégories sont spécifiques à la langue observée, nous serons donc tenue de nous référer à la notion d'équivalence traductologique : par exemple, le génitif anglais 's nous semble se rattacher à la classe des prépositions, dans la mesure où le français le rendra le plus souvent par la préposition *de*.

Après avoir discuté et présenté les classes d'équivalence finalement constituées et leur contenu (1.1.), nous contrasterons la morphosyntaxe des deux corpus au niveau plus local des sous-systèmes linguistiques élaborés (1.2.), et au niveau plus global de leur structure générique (1.3.).

9.1. Adaptation des descripteurs et construction d'équivalences

Une fois le corpus étiqueté[6], différents remaniements ont été effectués afin d'obtenir un ensemble de traits comparable au jeu d'étiquettes adopté pour le français.

Les noms des descripteurs ont d'abord été harmonisés, dans la mesure du possible : **JJ** a ainsi été renommé **ADJ**, **NN1 NC :sg** , etc.

Dans la mesure où il est probable que l'organisation générique de l'article scientifique anglo-saxon soit fondamentalement distincte de celle mise au jour pour le français, et qu'elle mobilise des descripteurs différents, aucune des étiquettes proposées par CLAWS – même les plus improbables – n'a dans un premier temps été supprimée ; ce n'est pas *a priori* mais dans le cadre d'une seconde étape statistique (*v. infra* 3.1.) qu'elles seront éventuellement isolées, afin d'objectiver l'analyse menée. En revanche, plusieurs catégories ont du être créées afin de maximiser les comparaisons possibles.

Les variables étant calculées relativement à leurs classes morphosyntaxiques d'appartenance, nous avons élaboré un ensemble de catégories globales relativement équivalentes aux 15 classes françaises, dans lesquelles ont été intégrés les descripteurs.

9.1.1. Classes morphosyntaxiques

Rappelons d'abord que le système de descripteurs du français se répartissait en 15 classes morphosyntaxiques :

[6] Les textes ont été traités un à un *via* WMatrix, qui présente l'intérêt de proposer une version XML de l'étiquetage. Notons cependant que cette version n'était pas bien formée XML, ce qui s'est avéré fort problématique, et d'autre part que différentes étiquettes non annoncées s'étaient greffées à l'étiquetage (*e.g.* différentes étiquettes sémantiques de type <unclear>, des tags de type <T=[chiffre]>, etc.).

1. Formalisation (symboles, sigles et abréviations)	9. Particules (*e.g.* semble[**-t-**]il)
2. Adverbes et connecteurs	10. Eléments de langue étrangère
3. Adjectifs	11. Ponctuations
4. Pronoms (personnels, disjoints, relatifs, etc.)	12. Subordonnants
5. Verbes	13. Interjections
6. Déterminants	14. Numéraux (cardinaux, ordinaux, indices de liste et de renvoi, etc.)
7. Noms (communs et propres)	15. Préfixes
8. Prépositions et amalgames	

Tableau : Classes morphosyntaxiques du système d'étiquetage du français

Après observation des catégories proposées par CLAWS, c'est un ensemble de 14 classes équivalentes que nous avons constituées, ensemble fondé sur le précédent modèle :

1. Formalisation	9. Catégorie composite (**FU**)
2. Adverbes et connecteurs	10. Eléments de langue étrangère
3. Adjectifs	11. Ponctuations
4. Pronoms	12. Subordonnants
5. Verbes	13. Interjections
6. Déterminants	14. Numéraux
7. Noms	
8. Prépositions	

Tableau : Classes morphosyntaxiques du système d'étiquetage de l'anglais

Les descripteurs, qui ont parfois dû être modifiés, adaptés, voire créés, ont dans un second temps été répartis dans les quatorze classes, dont on présentera brièvement le contenu.

9.1.2. Contenu des quatorze classes

Soulignons d'abord que les classes 9, 10 et 13 ne contiennent qu'un descripteur unique, de contenu non équivoque. La catégorie composite (**FU**) est constituée d'éléments ou de segments qui correspondent le plus souvent à des erreurs de segmentation initiale liées à la segmentation particulière que nécessite le traitement des textes scientifiques (*v.* chapitre 2). Ainsi, CLAWS segmente difficilement les slashs, qu'il ne reconnaît pas lorsqu'ils sont accolés aux mots qu'ils délimitent ; les séquences de type « NNS/NS communication » sont ainsi étiquetées :

<w ana="FU">NNS/NS</w>

<w ana="NC:sg">communication</w>[7]

Les éléments de langue étrangère (**FGW**) ont été manuellement vérifiés[8], CLAWS ayant effectué de très nombreuses erreurs d'étiquetage de cette catégorie.

[7] Le format présenté n'est pas celui de CLAWS, mais le format XML que nous avons adopté.

Les partitions finalement obtenues sont loin d'être satisfaisantes et demanderaient davantage d'élaboration, mais elles représentent un premier pas exploratoire dans un domaine d'observation à notre connaissance encore vierge.

9.1.2.1. Classe 1 : formalisation

Ont été considérés comme éléments potentiellement rattachés à la formalisation des textes scientifiques les formules (**FO**) et les marqueurs de type *[A-Za-z]* ou *[A-Za-z]'s* (**ZZ1, ZZ2, ZZ221, ZZ222**).

Nous avons esquissé une classe **ABR** (abréviations) qui n'est malheureusement pas équivalente à celle du français, dans la mesure où sa constitution aurait demandé un développement démesuré, eu égard à la visée exploratoire du présent chapitre. Cette classe ne contient donc pour l'heure que les & ; la catégorie a d'ailleurs été créée suite à l'étiquetage problématique de ces éléments (rappelons que & est un caractère illicite en XML, car mobilisé par la syntaxe concrète du langage formel).

Les symboles et marqueurs linguistiques n'ont pas été pris en compte étant donné l'important travail manuel que leur prise en compte nécessiterait.

9.1.2.2. Classe 2 : adverbes et connecteurs

A. Adverbes

La classe d'*adverbes* contient l'ensemble des catégories préfixées par **R** dans le système d'étiquetage initial. De manière générale et malgré les problèmes certains qu'elles posent parfois, nous avons globalement suivi les catégorisations proposées par les concepteurs de CLAWS ; ainsi, les locutions *that is* ou *for example* sont incluses dans la classe des adverbes, alors qu'elles s'inscriraient de manière plus pertinence dans celle des connecteurs. A ces étiquettes [R*][9], nous avons adjoint la catégorie des négations **XX**, de même que les post-déterminants comparatif et superlatif **DAR** et **DAT**.

B. Connecteurs

Les connecteurs français et anglais ne pourront globalement pas être comparés, dans la mesure où l'idée d'une typologie équivalente des objets a dû être écartée : les connecteurs se distribuent en effet dans des catégories et sous des étiquettes diverses et on admettra aisément que l'élaboration d'une table d'équivalences représente un travail substantiel que nous ne pouvons mener dans le présent cadre.

Les conjonctions de coordination (**CC**), de même que le connecteur *but* (**CCB**) et la catégorie hétérogène **CS** (e.g. if, because, unless, so, for), ont été incluses dans la classe des connecteurs.

---

[8] A noter qu'il demeure encore de nombreuses erreurs d'étiquetage de ces éléments, dans la mesure où nous n'avons pas pu vérifier *tous* les segments ; la catégorie a bien été vérifiée, mais les éléments de langue étrangère n'ont pas été exhaustivement considérés.

[9] Représentation sous forme d'expression régulière (chaîne de caractère préfixée par R et comportant 0 ou + de caractères suivant R).

9.1.2.3. Classe 3 : adjectifs

L'ensemble des tags préfixés par **J** dans le système initial a été rassemblé dans la catégorie des *adjectifs*. Les catégories pré- et post-déterminants définies par CLAWS étant problématiques, dans la mesure où nous n'en disposons pas d'équivalentes, elles ont globalement été réparties dans les classes existantes. Les catégories **DA** (*same, such, latter*) et **DA1** (*such, former, same*) ont ainsi été incluses dans la classe des *adjectifs* dans la mesure où *même* ou *tel* ont été étiquetés comme tels en français. On notera d'ailleurs que la catégorie **ADJ:refl** française que nous avions créée n'a plus aucune raison d'être en termes de comparaison, dans la mesure où elle équivaut aux réflexifs *himself/herself* qui ne correspondent qu'à une seule forme.

On voit d'ailleurs là le caractère hautement problématique de l'analyse contrastive multidimensionnelle : si l'examen d'un marqueur formel ou d'un objet sémantique est comparativement aisé à mettre en œuvre d'une langue à l'autre, la prise en compte d'un nombre important de descripteurs, *a fortiori* d'un ensemble de descripteurs supposés exhaustif d'un niveau d'analyse linguistique spécifique, est éminemment complexe et pose de sérieux problèmes d'équivalence.

9.1.2.4. Classe 4 : pronoms

Les catégories préfixées par **P** ont été regroupées dans la classe des *pronoms*. Si les pronoms personnels et clitiques pourront être comparés, les disjoints ne seront pas observables, les phénomènes d'emphase pronominale différant d'une langue à l'autre, tandis que les réflexifs et les pronoms possessifs sont confondus dans des classes globales ne différenciant pas le trait /personne/ en leur sein.

A ces éléments, nous avons adjoint les *pré-déterminants* **DB** (*all, half*) et **DB2** (*both*), les *déterminants* **DDQ** (*which, what*)**, DDQGE** (*whose*) et **DDQV** (*whichever, whatever, whatsoever*), de même que l'« existentiel » *there* (**EX**), qui nous semble avoir un statut pronominal.

Notons que l'observation de l'impersonnel *il/IT* et de l'indéfini *on* ne sera pas possible, les *it* anaphorique et impersonnel n'étant pas distingués, et l'indéfini *ONE* étant confondu au sein de la classe PN1, qui englobe l'ensemble des pronoms indéfinis (*anyone, everything, etc.*). La vérification au cas par cas de ces éléments n'étant raisonnablement pas envisageable, nous avons choisi d'abandonner ces marqueurs.

Enfin, il nous faut souligner que les pronoms démonstratifs ne sont pas distingués des déterminants possessifs, ce qui est problématique.

9.1.2.5. Classe 5 : verbes

Les éléments préfixés par **V** ont été inclus dans la classe des verbes. On notera l'ajout de la catégorie **TO** (*to* infinitif) à la classe, bien que la catégorie soit relativement redondante de celle des *infinitifs*[10] (**VVI**).

Les auxiliaires modaux *will ('ll, wo* – forme négative) et *would ('d, should, could*) ont été respectivement étiquetés **VER:fut** et **VER:cond**, afin d'être comparés aux temps du futur et du conditionnel français. On soulignera que ces auxiliaires ne sont pas les seules modalités d'expression du futur et du conditionnel en anglais, mais que ces réserves s'appliquent

[10] Ceci dit, les deux catégories ont des décomptes sensiblement différents : 132.47 **VVI** *vs.* 92.83 **TO** en moyenne par texte.

également au français (le conditionnel peut ainsi s'exprimer par la biais d'un imparfait dans le cadre des subordonnées en SI par exemple) – on atteint ici les limites du mot et de la forme. De la même manière, il n'est possible d'observer que les impératifs de type *let us* (**VM21/22**). Les séquences de type « compare the examples[11] » ne peuvent être incluses, car *compare* sera étiqueté **VD0** (simple present).

Soulignons également que les catégories verbales proposées par CLAWS sont souvent de type *un tag/une forme* (**VBM**/*am,* **VBZ**/*is,* **VBR**/*are,* **VDZ**/*does,* **VHZ**/*has*, etc.), ce qui limite certes les erreurs d'étiquetage et confère au tagger un taux de précision élevé (96-97% pour CLAWS), mais affaiblit l'intérêt et la cohérence linguistique de l'annotation.

9.1.2.6. Classe 6 : déterminants

La constitution de la classe des *déterminants* s'avère globalement peu problématique : les catégories articles (**AT** et **AT1**) correspondent respectivement aux déterminants définis et indéfinis (**DET :def** et **DET :indef**) que nous avons pris en compte pour observer le français, tandis que les classes fusionnées **DD1** et **DD2** équivalent à la catégorie des déterminants démonstratifs **DET :dem**. Enfin, les déterminants possessifs **DET :poss** correspondent à l'étiquette **APPGE** – dans laquelle le trait /personne/ n'est pas distingué.

Seule modification effectuée, le rajout de la catégorie **DA2** (*few, several, many*), dont les éléments constitutifs s'inscrivent dans la catégorie des déterminants indéfinis que nous avions élaborés.

Encore une fois, l'entreprise peut paraître naïve et discutable étant donné les fonctionnements distincts des deux langues en matière de détermination : le déterminant défini anglais est par exemple loin d'être systématiquement rendu par un défini français, mais peut être également traduit par un démonstratif, un possessif, voire par l'article ø qu'il est bien évidemment plus que délicat de traiter par le biais d'un étiquetage fondé sur le mot. Ces réserves doivent cependant être nuancées par la spécificité de notre objectif, qui vise à comparer les profils quantitatifs du genre de l'article dans les deux langues dans une approche globale qui s'écarte fondamentalement de la perspective traductologique locale traditionnelle.

9.1.2.7. Classe 7 : noms

Si la classe des *noms* française n'incluait que trois variables (noms communs singulier et pluriel, et noms propres), CLAWS propose un panel d'étiquettes bien plus large, dans lequel on recense une vingtaine de catégories sémantiques – d'ailleurs globalement peu ou non adaptées à notre objet : noms locatifs (*islands, streets*), de temporalité (*day(s), week(s))*, jours de la semaine, mois, etc. La catégorie **NPM2** (mois au pluriel) est d'ailleurs absente des textes (*v. infra*) mais comme nous l'avons déjà souligné, aucune des étiquettes proposées n'a été écartée à ce stade de l'expérience.

9.1.2.8. Classe 8 : prépositions

Les quatre tags préfixés par **I** ont été pris en compte – de même que les versants locutionnels de la catégorie **II** (*e.g.* **II21/22** *as far, apart from…*). Après réflexion, la catégorie *génitif* (**GE**) a été adjointe à cette catégorie, le génitif ainsi marqué correspondant régulièrement à une préposition *de* en français.

[11] Que l'on rendra dans la plupart des cas par « Comparons les exemples ».

9.1.2.9. Classe 11 : ponctuations

Les ponctuations sont bien prises en compte par CLAWS, mais on soulignera que leur statut et leur étiquetage diffère d'une ponctuation à l'autre. En règle générale, l'étiquette reprend le signe de ponctuation identifié (*e.g.* <w ana=":">:</w> ou <w ana=",">,</w>) : il en va ainsi des virgules, des points, des points-virgules, des deux points, des points d'interrogation, de suspension et d'exclamation, des parenthèses et des tirets.

Le traitement des guillemets simples est problématique, dans la mesure où le tagger les identifie quasi-systématiquement comme génitif lorsqu'elles sont accolées en fin de mot. Ce choix est justifié par des séquences comme :

> This is also true of speakers' and hearers' understanding (001)
>
> <w ana="NC:pl">speakers</w>
>
> ➔ <w ana="GE">'</w>
>
> <w ana="CC">and</w>
>
> <w ana="NC:pl">hearers</w>
>
> ➔ <w ana="GE">'</w>
>
> <w ana="NC:sg">understanding</w>

mais pose d'évidents problèmes lorsque le guillemet simple a une valeur de délimitation des mots, proche du guillemet, comme dans :

> The more conventionalized ones are relatively fixed constructions, close to 'fixed patterns' (…) (001)
>
> <w ana="ADV">close</w>
>
> <w ana="II">to</w>
>
> ➔ <w ana="ADJ">'fixed</w>
>
> <w ana="NC:pl">patterns</w>
>
> ➔ <w ana="GE">'</w>

On voit ici que le guillemet simple fermant est reconnu comme génitif, tandis que l'ouvrant n'est pas distingué car non pris en compte dans l'étape antérieure de segmentation.

CLAWS inclut pourtant les guillemets et les guillemets simples dans une même catégorie 'YQUO' qui fonctionne très bien pour les premiers, le guillemet étant globalement peu ambigu ; peu d'occurrences d'étiquetage du guillemet simple comme 'YQUO' ont d'ailleurs pu être relevées dans le corpus, les concepteurs ayant visiblement minimisé les usages du guillemet simple – ou plus exactement, le corpus sur lequel a été fondé CLAWS en contenait trop peu pour que le problème soit correctement traité.

Les slashs sont tout aussi problématiques que les guillemets simples, dans la mesure où ils sont systématiquement catégorisées 'FO' (formules), et ne sont pas correctement distingués ; la plupart du temps, ils demeurent accolés au mot, ce qui entrave évidemment le processus d'identification et de catégorisation et les séquences /[mot]/ sont donc étiquetées 'FU' (résidus). On soulignera cependant que lorsque l'élément précédant le slash est identifié par CLAWS, ce dernier confère à la séquence l'étiquette de l'élément reconnu : par exemple, la séquence « native speaker/non-native speaker » sera tokenisée [native][speaker/non-native][speaker] et le second token sera catégorisé **NC:sg** eu égard à l'identification du premier *speaker*.

Les problèmes posés par les guillemets simples et les slashs n'ont pas pu être réglés dans le cadre de notre étude ; il conviendrait idéalement d'adapter le programme de tokenisation

que nous avions développé pour le français à l'anglais, et d'entraîner TnT sur le corpus anglais, tâche que nous avons choisi d'écarter ici.

Notons enfin qu'aucune accolade, aucun crochet et aucun antislash n'a pu être observé dans le corpus anglais.

9.1.2.10. Classe 12 : subordonnants

Aux éléments de la catégorie BCL (*Before Clause Marker*), nous avons adjoint les étiquettes CSN (*than*), CST (*that*), CSW (*whether* et *whether or not*) et les locutions préfixées par DDQV (*no matter what*).

La catégorie se recouvre quelque peu avec celle des connecteurs, dans la mesure où cette dernière contient certains éléments que nous avions étiquetés *subordonnants* et non *connecteurs* (*e.g.* la conjonction *si* est par exemple considéré comme un subordonnant dans le système français, tandis que *if* est inclus dans la catégorie des connecteurs en anglais).

9.1.2.11. Classe 14 : numéraux

Si CLAWS catégorise les numéraux, force est de constater que sa typologie diffère sensiblement de la nôtre ; outre les cardinaux et les ordinaux, qui sont de toutes façons pris en compte par l'ensemble des étiqueteurs, CLAWS distingue les chiffres séparés par des tirets, ce que nous n'avions pas considéré – à l'exception des configurations où ces séquences correspondaient à des marqueurs de structuration ou à des listes : après observation des textes, cette catégorie s'avère intéressante, puisqu'elle renvoie quasi-systématiquement aux mentions de pages dans les références citationnelles. Outre ces éléments, les fractions (*two thirds*) sont également étiquetées, ce qui peut éventuellement préciser les modalités de quantification linguistique. On notera enfin que le numéral *one* et son versant arabe 1 est regroupé au sein de la catégorie MC1, distinction qui paraît *a priori* singulière.

Les catégories française et anglaise sont ainsi distinctes, dans la mesure où les titres et les indices de liste ne seront pas comparables ; *a fortiori*, leur non prise en compte entraînera certains biais dans l'analyse (augmentation des points et des numéraux) que nous avions évité pour le français.

Les dates étant facilement identifiables, nous les avons récupérées à l'aide d'expressions régulières.

9.1.3. Table des équivalences

La table qui suit synthétise les remarques précédemment émises, et propose une vision globale de la comparabilité des deux systèmes de descripteurs :

<table>
<tr><th>Classe</th><th>Catégories spécifiques au tagset français</th><th colspan="2">Catégories communes</th><th>Catégories spécifiques au tagset anglais</th></tr>
<tr><td rowspan="2">1. Formalisation</td><td rowspan="2">ABR
SIG
SIG :ling
SYM :ling
SYM :gram</td><td colspan="2">ABR (&)</td><td rowspan="2">ZZ2
ZZ221(2)</td></tr>
<tr><td>SYM</td><td>ZZ1
FO</td></tr>
<tr><td>2. Adverbes et connecteurs</td><td>Distinction de 15 types de connecteurs</td><td colspan="2">ADV, ADV:int, ADV:neg, etc.</td><td>Distinctions spécifiques (comparatifs,</td></tr>
</table>

<table>
<tr><td></td><td></td><td colspan="2"></td><td>connecteur BUT, etc.)</td></tr>
<tr><td>3. Adjectifs</td><td>Trait [nombre]</td><td colspan="2">ADJ</td><td>Distinctions spécifiques (comparatifs, superlatifs, caténatifs, etc.)</td></tr>
<tr><td rowspan="6">4. Pronoms</td><td rowspan="6">+ tags spécifiques PRO :indef, PRO :dem</td><td colspan="2">Pronoms personnels PRO :pp1sn, PRO :pp2, PRO :pp3msn, PRO:pp1pl, PRO:pp3pl</td><td rowspan="6">+ tags spécifiques pour whichever, whatever, whoever, each other, one another, both, etc.</td></tr>
<tr><td>il impersonnel</td><td>PPH1/EX</td></tr>
<tr><td>Indéfini ON</td><td>PN1</td></tr>
<tr><td>Clitiques</td><td>PPIO1, PPHOI1, PPIO2, PPHO2</td></tr>
<tr><td>ADJ:même + PRO:refl se</td><td>Réflexifs PNX1, PPX1, PPX2</td></tr>
<tr><td>Pronoms possessifs + [personne]</td><td>PPGE</td></tr>
<tr><td rowspan="13">5. Verbes</td><td rowspan="13">Subjonctifs présent et imparfait, participe présent, auxiliaires conditionnel, subjonctif présent et imparfait, infinitif, futur, participes présent et passé, imparfait, passé simple</td><td colspan="2">VER:cond, VER:mod:cond, VER:fut</td><td rowspan="13">+ Forme progressive future VVGK
Forme progressive VVG, VBG, VDG, VHG
Ought et used VMK</td></tr>
<tr><td>VER:inf</td><td>Distinction TO, VVI, VBI, VHI, VDI,</td></tr>
<tr><td>VER:pres</td><td>Distinction VBM, VBZ, VBR, VBR, VD0, VVZ, VDZ, VV0</td></tr>
<tr><td>VER:aux:pres</td><td>VHZ, VH0</td></tr>
<tr><td>VER:simp, VER:impf</td><td>Simple past VVD, VBDR, VBDZ, VDD, VHD</td></tr>
<tr><td>VER:pper</td><td>VVN, VBN, VDN, VVNK (bound to)</td></tr>
<tr><td>VER:aux:pper</td><td>VHN</td></tr>
<tr><td>VER:mod:[temps]</td><td>VM</td></tr>
<tr><td>VER:imp</td><td>VM21(22)</td></tr>
<tr><td rowspan="2">6. Déterminants</td><td rowspan="2"></td><td colspan="2">DET:def, DET:indef, DA2, DET:dem</td><td rowspan="2">DET:dem inclut également les pronoms démonstratifs
+ DA2 (few, several, many)</td></tr>
<tr><td>DET:poss:[personne]</td><td>APPGE</td></tr>
<tr><td>7. Noms</td><td></td><td colspan="2">NC:sg, NC:pl, NP</td><td>+ nombreuses distinctions (localités, titres, numéraux, mois, etc.)</td></tr>
</table>

8. Prépositions		PREP	IF, II, IO, IW, GE	
		Locutions PREP:[attribut positionnel]	II[attribut positionnel]	
9. Composite				FU
10. Eléments de langue étrangère		FGW		
11. Ponctuations	Antislashs, accolades, crochets	Ensemble des ponctuations		Caractère peu faible des slashs et des guillemets simples
12. Subordonnants		SUB	CSN, CST, CSA, CSW, locutions BCL et DDQV	Recouvrements avec la catégorie des connecteurs
13. Interjections		INT		
14. Numéraux	Structuration textuelle LS, NUM :par	NUM:car, NUM:ord, NUM:dat		Distinction fractions, cardinaux avec tirets, cardinaux singulier (*one* et 1)
15. Préfixes	PREF			

Tableau : Synthèse des lieux d'équivalences des deux systèmes d'étiquetage

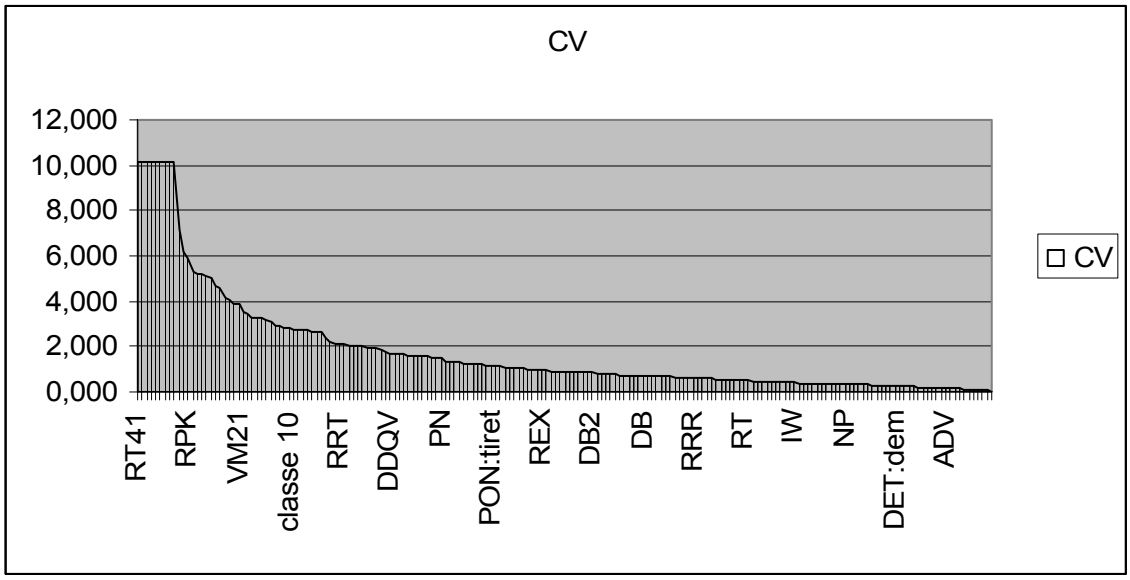
9.2. Statistiques descriptives comparées

Les exemples des textes anglais n'ayant pas été balisés, l'entreprise descriptive et comparative sera menée sur textes entiers.

9.2.1. Nouveaux seuils statistiques

9.2.1.1. Coefficient de variation

La variabilité de l'article anglo-saxon est d'abord très similaire à celle du genre français : plus des deux tiers des variables observées ont également un CV inférieur à 2 (soit 200%) : les éléments ayant un CV supérieur à 2 peuvent donc être considérés comme plus aléatoires, et de fait moins caractéristiques du genre :



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

9.2.1.2. Seuil de corrélation

Si l'on reprend la formule précédemment adoptée :

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

n = nombre d'individus

on obtient un seuil de 0.2, qu'on arrondira à 0.25 en raison des comparaisons multiples.

9.2.2. Eléments marginaux du genre

Le jeu d'étiquettes de CLAWS ayant une granularité particulièrement élevée et n'étant pas spécifiquement adapté à l'observation des textes scientifiques, de nombreux descripteurs sont de manière peu surprenante absents de la majorité des textes du corpus :

Catégorie générale	Catégorie spécifique		% valeur 0
	Etiquette	Descriptif bref	
Formalisation	**ZZ2**	Formes a's	99,029
	ZZ221	Formes A' S	86,408
	ABR	Abréviations	66,990
Adjectifs	**JJ21**	Adjectifs par excellence, tout court	80,583
Adverbes et connecteurs	**RGQV31**	no matter how	99,029
	RT31	from now on	99,029
	RT41	for the time being	99,029
	RPK	about (to be)	97,087
	RL21	on board	96,117
	RT21	by now	96,117
	RG21	far from	90,291
	REX41	that is to say	87,379
	RGQV	however	87,379
	RR41	for the most part, from time to time, all of a sudden	86,408

	RRQV	whenever, wherever	79,612
	RR31	and so on, as it were	55,340
	RRT	Superlatif (best, longest)	52,427
	RGQ	wh degré (how)	50,485
Connecteurs	**CS41**	in as much as, in so far as	98,058
	CS31	as far as, as soon as	67,961
Subordonnants	**DDQV31**	no matter what	97,087
	CSW31	whether or not	76,699
Prépositions	**II41**	from the view to, on the part of, in the light of	72,816
Numéraux	**MF**	Fractions (two thirds, quarters)	83,495
Noms	**NN131**	sine qua non	99,029
	NN221	dramatis personae	99,029
	NNU21	per cent	99,029
	NPD2	Jours de la semaine au pluriel	99,029
	NN21	lingua franca	94,175
	NNA	titres post (e.g. Jr.)	93,204
	NNO2	NC numéraux pluriel (hundreds, thousands…)	89,320
	NNU1	Unités de mesure sg (inch)	87,379
	NPD1	Jours de la semaine singulier	87,379
	NNL1	NC locatifs (islands, streets…)	82,524
	ND1	Noms de direction (south, southeast…)	77,670
	NNO	NC numéraux neutres (dozen, hundred…)	75,728
	NNB	Titres (e.g. Mr., Prof.)	73,786
	NNU2	Unités de mesure pl (feet)	63,107
	NPM1	Mois au singulier	63,107
Ponctuations	**PON:slash**	Slashs	65,049
	PON:cote	Guillemets simples	64,078
	PON:excl	Points d'exclamation	61,165
Pronoms	**PNQV**	WHOEVER	99,029
	PNX1	Réflexif ONESELF	94,175
	PN121	NO ONE	93,204
	PPX121	one another	84,466
	PNQO	Relatif WHOM	72,816
	PRO:pp1sn	Pronom I	71,845
	PPGE	Pronoms possessifs (mine, yours…)	71,845
	DDQV	whichever, whatever, whatsoever	65,049
	PN	pronoms NONE et PLENTY	56,311
	DDQGE	relatif whose	52,427
Verbes	**VVNK**	bound to	95,146
	VM21	let's	90,291
	VVGK	GOING TO	72,816
	VHN	HAD participe passé	63,107
	VMK	OUGHT et USED	62,136
	VBM	AM	54,369

Tableau : Eléments marginaux du genre de l'article

Si les pronoms possessifs sont apparus comme les grands marginaux du genre de l'article français (de plus de 90% des textes), on notera qu'il en va différemment des textes anglais, où les descripteurs ne sont absents que des deux tiers des textes. Outre les différences de construction linguistique séparant les deux langues, ce phénomène nous semble d'une part lié à la présence importante d'exemples oraux dans les textes mobilisant ces variables, et d'autre part au fait que les anglo-saxons répugnent moins à marquer la relation de possession personnelle que les français :

> In my view, in the analysis of nonce-word tests – including *mine* presented in the following section – several questions arise (046)

> (…) the reader of the narrative must be cautioned, however, that the choice of detail is *mine*, and the kind of information I choose to pass on may be different from the experience of a reader/viewer who is able to witness the actual sequence (…) (056)

Notons également que le *you* anglais est loin d'être marginal, et n'apparaît d'ailleurs pas dans ce tableau au contraire du *I*, dont le versant *je* était loin d'être contingent en français. Le phénomène est pour le moins contre-intuitif, eu égard au nombre important d'études clamant la mise en avant de l'auteur anglo-saxon *via* la première personne du singulier et l'effacement tout occidental de l'auteur français.

Si le subjonctif imparfait et le passé simple étaient peu représentés dans le genre français, ce sont les temps de l'impératif, du futur de type *going to* et du pluperfect qui semblent contingents en anglais. Absent de 90% des textes, l'impératif anglais de type *let's* ne nous semble pas désavouer la présence d'un style dialogique ou *reader-friendly* : si les Français recourent volontiers aux impératifs de première personne, de type « examinons » plutôt que « examinez », les anglo-saxons sont connus pour préférer les structures impératives de seconde personne de type « examine » plutôt que « let's examine » ; ces différences ne pourront malheureusement pas être observées, CLAWS n'identifiant pas ce type d'impératif.

Les résultats obtenus sur les ponctuations sont très similaires aux observations que nous avions inférées du corpus français : bien que biaisées, les guillemets simples sont relativement marginaux, tandis que les points d'exclamation sont globalement peu usités.

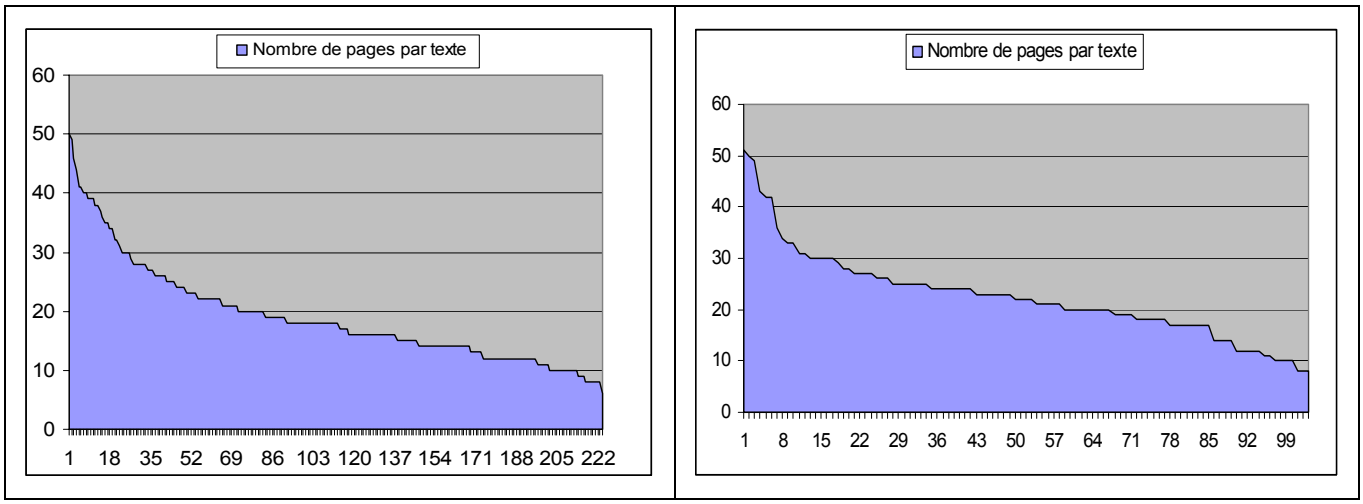
Mentionnons également la peu surprenante faible représentation des fractions et des formes de type a's, de même que la moins attendue sous-représentation des connecteurs *however* et *that is to say* – liées probablement à une préférence accordée à d'autres items pour exprimer la même relation (*e.g. id est* ou *i.e.* à la place de *that is to say*).

On soulignera enfin le caractère marginal – et attendu – des jours de la semaine, des noms numéraux, des noms locatifs et de direction ; une représentation élevée de ces items aurait été pour le moins singulière.

9.2.3. Eléments de description et de contraste du genre dans les deux langues

9.2.3.1. De la longueur des textes

L'article de linguistique anglo-saxon semble d'abord avoir une longueur moyenne sensiblement plus importante que le genre français (22.5 pages *vs.* 19.1) :



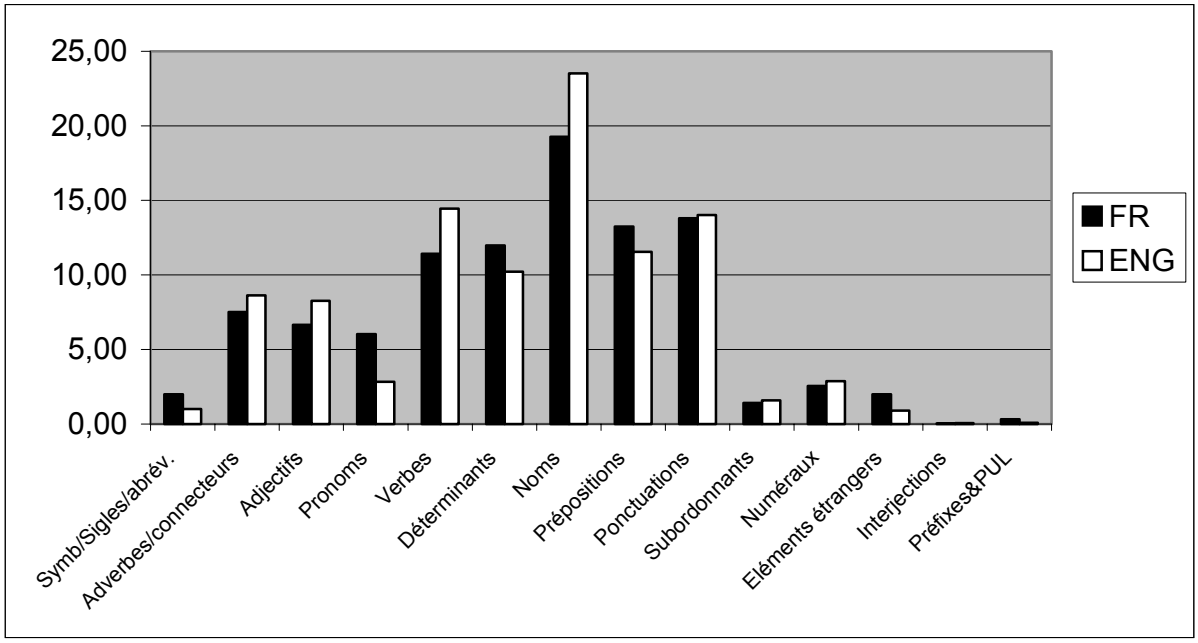
Graphique : Nombre de pages par texte en français (gauche) et en anglais (droite)

Plus un texte est long, plus il contient de pronoms de 3e personne du singulier (+0.36), de génitifs (+0.28), de parenthèses (+0.32), de verbes conjugués au *simple present* (+0.3), de connecteurs *but* et de subordonnants (+0.27) ; à l'inverse, les articles plus courts sont significativement corrélés aux formes *been* (-0.34), aux points (-0.33), aux déterminants (-0.28), aux conjonctions de coordination (-0.27), aux superlatifs (-0.25) et aux pronoms personnels *we* (-0.24).

Les textes plus courts ne sont pas comme en français corrélés aux chiffres ; s'il semble que les articles plus longs contiennent dans les deux langues une syntaxe plus complexe (les textes longs sont globalement corrélés aux subordonnants), il y aurait peut-être une opposition entre textes plus descriptifs et textes-rapports (usage de phrases courtes, de *been*, de coordination et de *we*), qui sera à confirmer, voire à infirmer par la suite.

9.2.3.2. De la répartition des classes linguistiques

Le graphique qui suit présente les répartitions des classes dans les deux langues ; les éléments français préfixes et PUL sont ici contrastés à la classe composite anglaise :



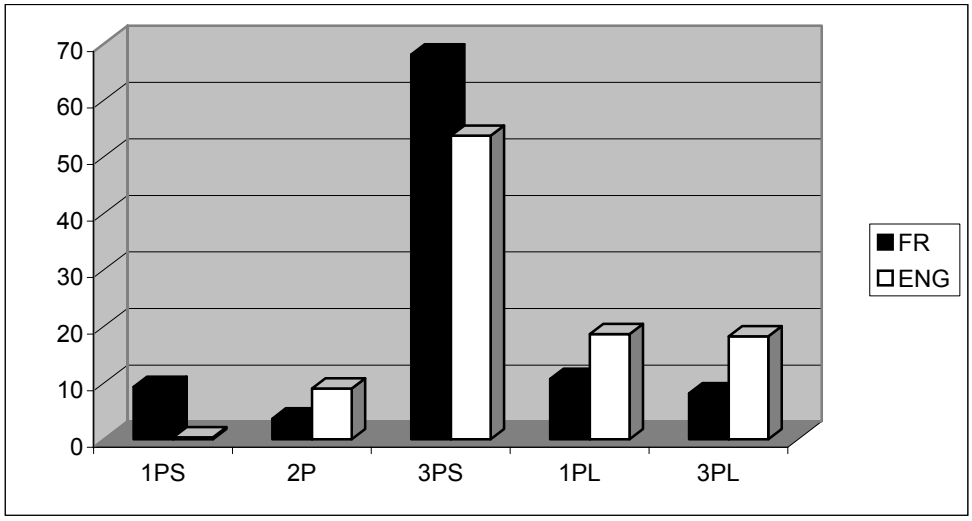
Graphique : Répartition contrastive des classes linguistiques (en %)

Soulignons qu'il est difficile de déterminer précisément ce que le graphique ci-dessus contraste : s'agit-il de différences génériques, discursives ou systémiques ? Cette question reste ouverte, et mériterait d'être validée de manière plus précise.

Le genre français contient ainsi une proportion plus importante de pronoms, de déterminants, de prépositions et d'éléments de langue étrangère (cette dernière catégorie étant biaisée), tandis que l'anglais utiliserait davantage d'adverbes et de connecteurs, d'adjectifs, de verbes, de noms, voire de numéraux.

9.2.3.3. Des personnes

L'examen des systèmes de personne dans les deux langues requiert plusieurs adaptations : l'observation des pronoms *one* et *it* impersonnel est en effet exclu, et la seconde personne ne s'incarne qu'en une forme unique *you*. Par conséquent, les pronoms de 3^{e} personne du singulier (à l'exclusion de *one*, qui n'est simplement pas observable) ont été regroupées au sein d'une même catégorie, de même que les *vous* et *tu* français :

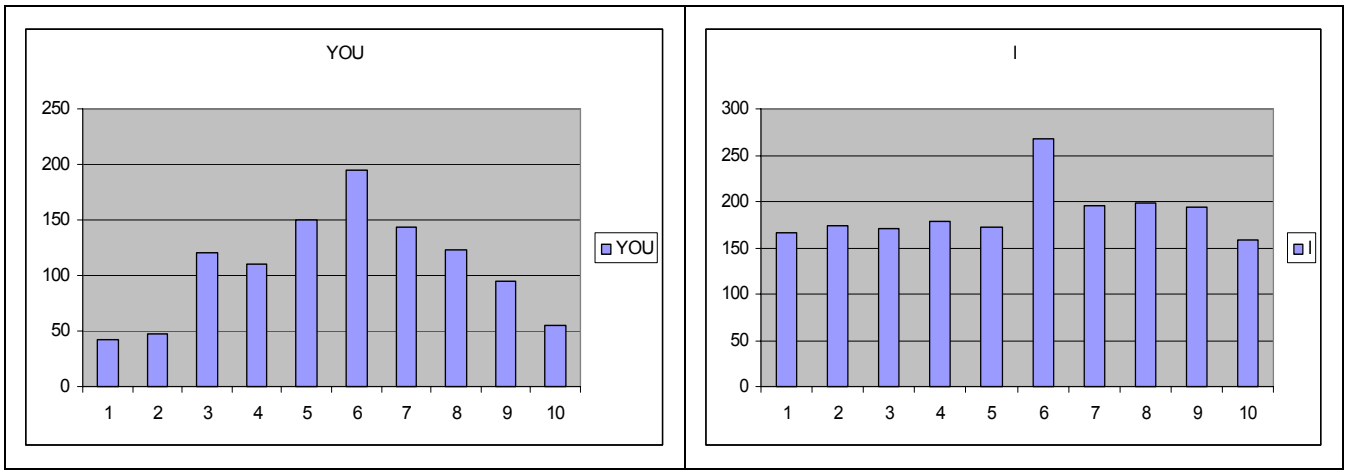


Graphique : Répartition contrastive des pronoms personnels sujets (moyennes absolues pondérées en % par texte)

Comme nous avons déjà pu l'observer, le français recourt massivement à la troisième personne du singulier – et les pronoms *on, il/elle anaphorique* et *il impersonnel* étaient globalement employés en proportions égales (*v.* chapitre 3). Il semble en aller différemment du genre anglais, qui mobilise davantage les pronoms *WE, THEY* et *YOU* que le pronom *I*, qui apparaît comme plus marginal.

On ne relève pas de corrélation significative interprétable du pronom I, à l'exception peut-être d'une corrélation positive avec les jours de la semaine au singulier (+0.24), qui indiquerait un emploi privilégié du pronom dans les textes exemplifiés. En revanche, *you* est nettement corrélé aux exemples, et plus spécifiquement aux corpus oraux mobilisés par certains textes exemplifiés ; le genre de la conversation, qui recourt particulièrement au pronom, est ainsi particulièrement présent dans les textes. *You* est ainsi fortement corrélé aux interjections (+0.5), aux marques de première personne du singulier (formes *am* +0.48 et *me* +0.41), aux questions (points d'interrogation +0.44, *do* et *did* auxiliaires resp. +0.41 et +0.27), aux slashs (+0.41), aux deux points (+0.40), aux titres antéposés comme *Mr.* ou *Prof.* (+0.34), aux points de suspension (+0.31), aux adverbes locatifs de type *alongside* (+0.27) et de type *whenever, wherever* (+0.25) et aux fractions (+0.25). L'ensemble de ces éléments dessinent une dimension *exemplification* comparable à celle que nous avions mise au jour pour le français (présence de questions, d'une dimension interlocutive *I / you*, etc.), et précisent les types de matériel linguistique privilégiés par les textes anglo-saxons.

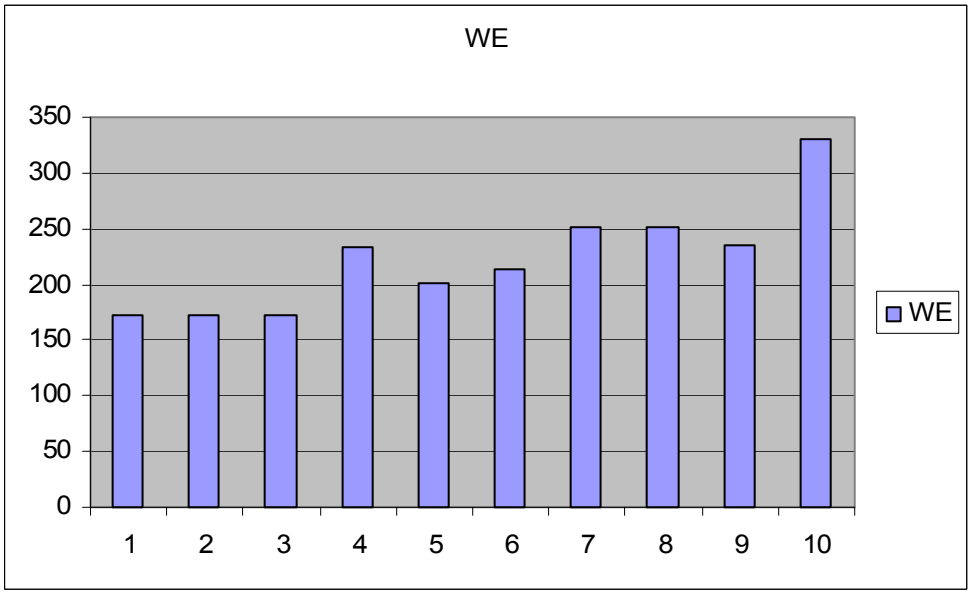
I et *you* sont donc, comme en français, particulièrement employés dans les exemples, ce que confirme l'examen de leurs configurations tactiques : les deux pronoms atteignent ainsi leur maximum en corps d'article, lieu privilégié de positionnement des exemples :



Graphique : Configurations tactiques des pronoms I et YOU

Notons que l'article de linguistique anglo-saxon semble atteindre son pic d'exemples au sixième dixième de son corps ; on admirera la belle symétrie de la courbe ascendante/descendante du pronom *you*, qui nous semble indiquer un usage quasi exclusif du pronom dans les exemples des textes – I semble ainsi plus diversement employé.

Il en va fort différemment du pronom *we*, dont la disposition tactique est bien distincte, et d'ailleurs inverse de celle que nous avions observée avec le *nous* français :



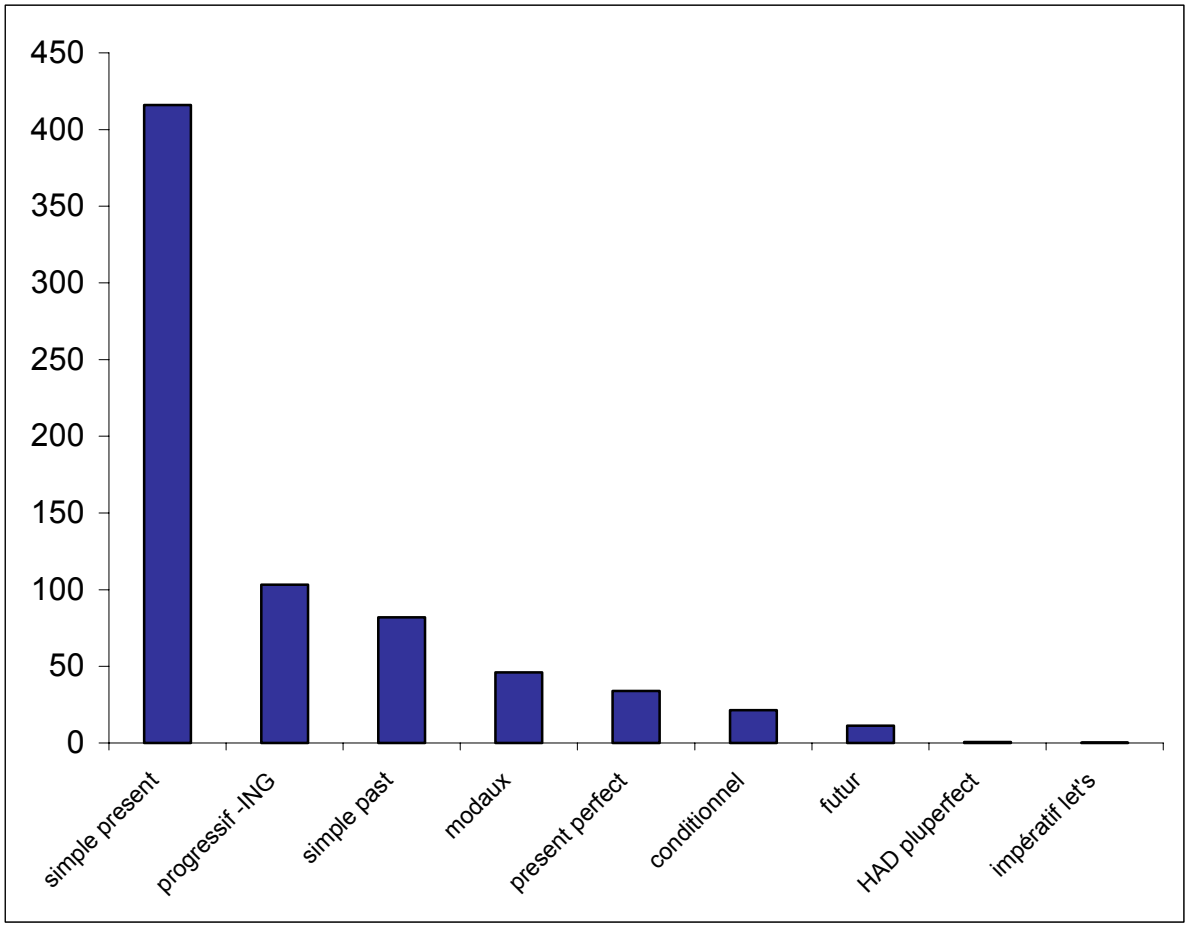
Graphique : Configuration tactique du pronom WE

On observe en effet une courbe ascendante du *we*, avec un maximum atteint en fin d'article, tandis que le pronom français connaissait son apogée en début de texte. Il semble ainsi que le lecteur français soit guidé par *nous* en début d'article, dans le cadre d'une fonction d'annonce de ce qui va suivre (*nous* était d'ailleurs significativement corrélé au futur), tandis que le *we* anglo-saxon semble récapitulatif ; le pronom est d'ailleurs significativement corrélé au *present perfect,* ou plus exactement au *have* auxiliaire (+0.3). Les deux pronoms remplissent ainsi une fonction dialogique distincte selon la langue considérée : dans le cas français, le *nous* est tourné vers le futur du texte, tandis que dans l'article anglo-

saxon, le *we* récapitulerait son passé, voire éventuellement les recherches à venir – ce qui induirait que les français demeurent dans le cadre textuel, tandis que les anglo-saxons s'en écartent et se positionnent dans un cadre plus large englobant le texte et ceux à venir.

9.2.3.4. Des temps verbaux

Soulignons d'emblée que la comparaison des temps français et anglais est loin d'être évidente : la forme progressive anglaise en *–ing* n'a pas exemple pas d'équivalent immédiat en français et les deux systèmes aspectuels ne se recoupent pas ; le parfait / imparfait ne se répartit donc pas de la même manière en français et en anglais :



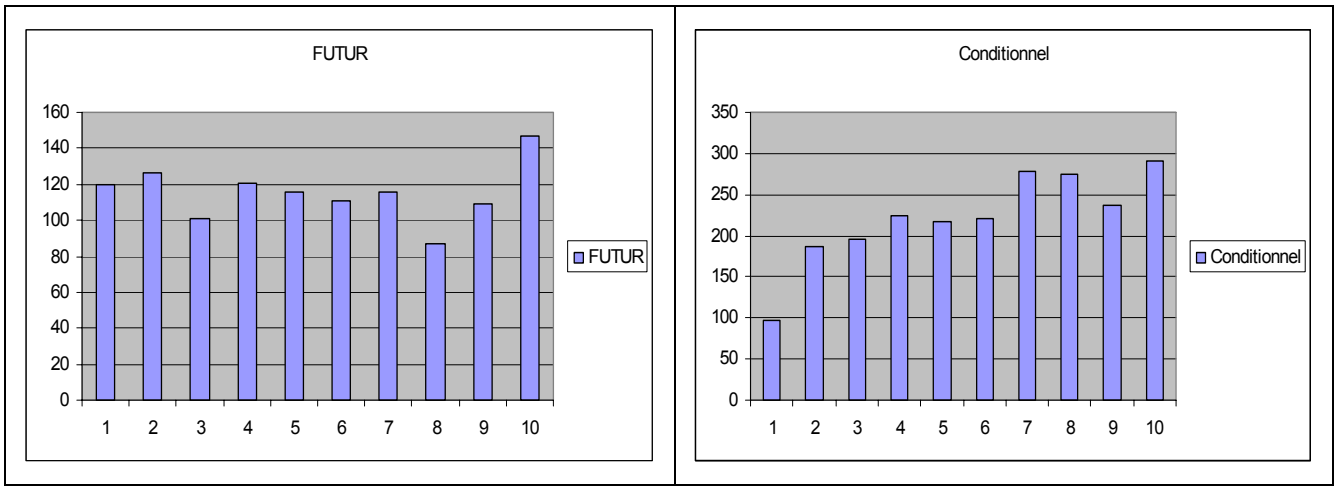
Graphique : Répartition des temps conjugués dans les textes entiers (corpus français et anglais)

Le graphique qui précède permet toutefois d'observer certaines régularités : le présent est bien le temps le plus représenté dans les deux langues, et on observe que l'ensemble des temps du passé occupent le second rang et sont suivis du conditionnel et du futur.

Plusieurs différences substantielles peuvent toutefois être soulignées : le passé composé est ainsi plus employé que le *present perfect* anglais, et on observe un usage plus important du futur en français.

Will et *would* étant généralement appréhendés conjointement par les grammaires comme les deux versants d'une même forme *will*[12], les deux formes nous semblent mériter une observation plus approfondie, ne serait-ce que pour légitimer la distinction que nous avons choisie d'effectuer entre les deux auxiliaires.

Si le futur et le conditionnel – qui sont des temps sémantiquement proches dans la mesure où ils expriment des évènements qui n'ont pas eu lieu – s'opposaient en français sur le plan tactique, les configurations des deux auxiliaires anglais, bien que relativement distinctes, suivent le même mouvement ascendant :



Graphique : Configuration tactique des auxiliaires WILL et WOULD

Les deux auxiliaires se rapprochent ainsi davantage du conditionnel que du futur français sur le plan tactique. Examinons maintenant leurs corrélats textuels : il est d'abord intéressant d'observer que les verbes conjugués au futur dans les deux langues ont les mêmes corrélats négatifs : ils s'opposent tous deux aux dates et aux parenthèses. En revanche, leurs corrélats positifs diffèrent substantiellement : si le futur français était associé à une dimension dialogique prospective (pronom *nous*, impératifs, ordinaux, etc.), il en va différemment de l'auxiliaire futur anglais qui est associé à des éléments distincts, et par ailleurs peu interprétables (subordonnant *as*, connecteur *but*, etc.).

Les corrélats positifs et négatifs de *would* sont également peu significatifs (auxiliaires infinitifs *be* et *have*, locutions prépositionnelles *as for, apart from*, etc. en positif / *have* auxiliaire, conjonctions de coordination, pronom *nous* en négatif). On voit ici toutes les limites d'une perspective émergentiste, et les difficultés que pose l'interprétation d'observables généralistes et non contrôlés – inadaptés ici à la description du discours / du genre qui nous intéresse. Rappelons ainsi qu'avec un ensemble de descripteurs pertinents, nous avions pu mettre au jour une corrélation du conditionnel français au spéculatif (connecteurs de doute et d'opposition, *il* impersonnel) ; faute d'observables, il demeure pour l'heure exclu de valider l'existence d'une telle dimension dans le corpus anglais.

[12] Cela explique leur inclusion au sein d'une même classe par CLAWS – et par l'ensemble des systèmes d'étiquetage de l'anglais.

Précisons notre examen en examinant les co-occurrents des deux auxiliaires *will*[13] et *would* et des deux temps français observés – dans la mesure où le corpus français est deux fois plus étendu, nous avons adopté un seuil de 20 pour le français, et de dix pour l'anglais (fréquence absolue) :

WILL				**WOULD**			
Co-occurrents gauche > 10		**Co-occurrents droit > 10**		**Co-occurrents gauche > 10**		**Co-occurrents droit > 10**	
Forme	**Fréquence absolue**	**Forme**	**Fréquence absolue**	**Forme**	**Fréquence absolue**	**Forme**	**Fréquence absolue**
I	150	be	288	it	101	be	264
we	105	not	55	I	69	have	99
it	68	have	39	we	49	not	64
they	28	show	25	that	39	like	47
that	27	also	19	they	32	seem	24
which	26	see	16	this	31	expect	18
and	25	take	14	which	27	also	17
as	14	argue	13	one	17	appear	12
this	11	need	13	he	16	argue	12
students	11	write	12	and	15	you	12
you	10	give	12	what	13	make	11
		only	11	there	11	need	10
		now	11	she	10	normally	10
		consider	11	students	10		
		focus	11				
		help	10				

Tableau : Co-occurrents droit et gauche des auxiliaires will et would[14] *–ASLA*

Futur				**Conditionnel**			
Co-occurrents gauche > 20		**Co-occurrents droit > 20**		**Co-occurrents Gauche > 20**		**Co-occurrents Droit > 20**	
Forme	**FA**	**Forme**	**FA**	**Forme**	**FA**	**Forme**	**FA**
nous	611	sera	216	ne / n'	266	serait	431
on	467	aura	104	qui	208	saurait	91
je / j'	227	verrons	72	il	139	aurait	87
ne / n'	195	notera	61	se / s'	107	seraient	85
il	128	dira	49	on	77	ferait	31
qui	92	verra	44	je / j'	53	permettrait	31
se / s'	158	seront	41	y	36	correspondrait	22
le	66	permettra	39	en	34		
en	59	fera	38	nous	33		
y	39	remarquera	34	ce	29		

[13] Les formes contractées de l'auxiliaire ('ll) sont 30 fois moins représentées (234 *vs.* 6201) et relèvent nettement des exemples – fait intéressant, on notera qu'ils sont également très majoritairement précédés des pronoms *I* et *We*.

[14] Les formes contractées des auxiliaires ('ll, won't, 'd et wouldn't) sont comparativement 30 fois moins représentées (e.g. 234 *vs.* 6201 pour *'ll* et *will*). Associées à un style peu soutenu et plus oral, ces formes relèvent des exemples des textes, et ont donc été isolées des décomptes.

elle	28	trouvera	25	elle	26
me	28	auront	22	le	26
et	25	parlera	22		
ce	24	reviendrons	22		
tu	24	appellerons	21		
l'	23	viendra	21		

Tableau : Co-occurrents droit et gauche des temps futur et conditionnel –ASLF

Les co-occurrents relevés sont particulièrement stabilisés : la sélection du premier co-occurrent gauche de *will* ramène ainsi 437 formes distinctes *vs.* 288 si l'on extrait les premiers co-occurrents à droite de la forme et seuls 11 co-occurrents gauche de *will* apparaissent au moins 10 fois (43.18% de l'ensemble des co-occurrents relevés), *vs.* 16 pour les co-occurrents droit (50.36%). Il en va d'ailleurs de même de *would* : seuls 14 co-occurrents gauche de *will* apparaissent au moins 10 fois (44.22 % de l'ensemble des co-occurrents relevés), *vs.* 13 pour les co-occurrents droit (59.17%). Les co-occurrents français semblent significativement plus contraints : 85.59% des co-occurrents gauche et 45% des co-occurrents droits des verbes conjugués au futur apparaissent ainsi plus de dix fois dans le corpus – le futur français semble ainsi plus volontiers précédé d'un grammème que d'un syntagme.

Will et *would* et les temps français du futur et du conditionnel partagent de nombreux co-occurrents (en grisé), de rang toutefois différent.

Will et would sont généralement suivis de la forme verbale *be* : le devenir de l'état est privilégié, et w*ill be* est sept fois plus représenté que *will have*, tandis que *would be* est 2.5 fois plus employé que *would have*. On relève ainsi de nombreuses phraséologies introductives et dialogiques de type *will be* + participe passé (*it will be argued, possible, helpful, recalled, shown*, etc.) – l'auteur pouvant bien évidemment se permettre d'exprimer un degré de certitude optimal dans ces cas, puisque le futur n'est autre que celui du texte, tandis qu'une prédiction de type *will + have* est autrement plus aventureuse.

Le conditionnel exprimant un futur bien plus hypothétique (*it would be interesting, reasonable, possible, difficult, accurate*, etc.), il n'est pas surprenant que l'écart entre *be* et *have* soit moins prononcé.

Si *would* préfère d'ailleurs les tournures impersonnelles de type *it would be (interesting, reasonable, possible, accurate, etc.*), *will* est d'abord associé à la première personne du singulier, acteur affirmé de la recherche anglo-saxonne et auteur-guide de la recherche présentée : les verbes de monstration (*to see, to show*) sont ainsi fortement mobilisés, alors que les formes verbales associées au conditionnel sont sémantiquement modérées (*like, seem, expect*…).

Les pronoms *I* et *We* n'agrègent bien évidemment pas les mêmes formes verbales, *We* ayant souvent une valeur inclusive dialogique : *I + would* est d'abord suivi – par ordre de décroissance – des verbes *to like, to argue, to say* et *to suggest*, tandis que *We + would* est plus volontiers employé avec *to expect, to like, to suggest, to need*. Les formes de politesse comme *I (We) would like* sont d'ailleurs bien plus spécifiques à l'anglais qu'au français. Avec will, *I* est plus volontiers suivi des verbes *to argue, to show, to discuss, to describe, to try, to examine, to focus* et *to present*, tandis que c'est avec *to see, to consider, to call, to show* et *to examine* que *We* est employé.

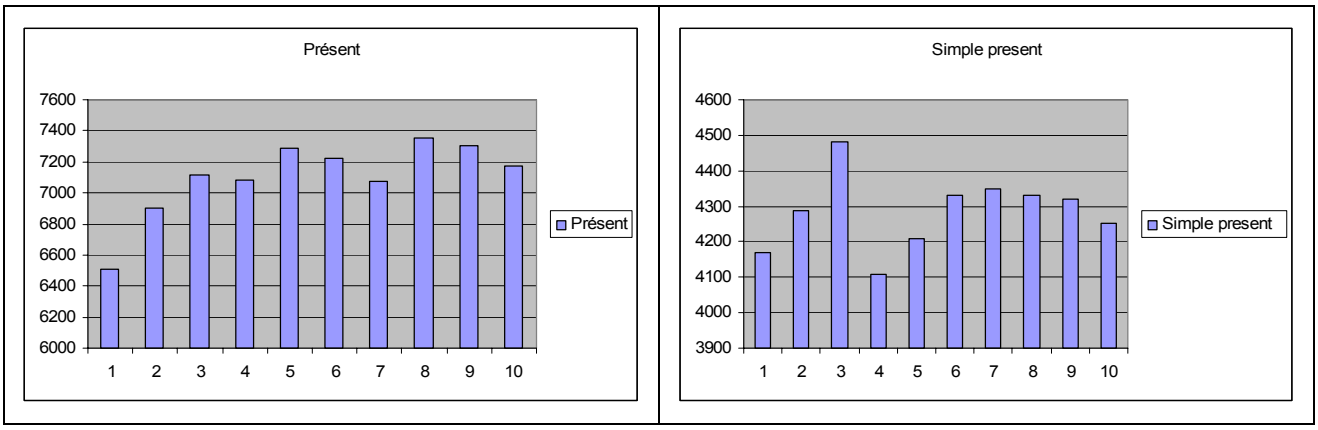
Si l'on confronte ces observations aux données françaises, on constate d'abord que le futur répugne moins à la forme *aura* que l'anglais : *sera* n'est finalement que deux fois moins employé que *aura* – il serait ici bien entendu important d'isoler les emplois narratifs du futur. Le conditionnel français semble plus proche du futur anglo-saxon sur ce plan, dans la mesure où *serait* est cinq fois plus employé qu'*aurait*.

Les deux futurs sont associés aux marques de première personne – au *nous* pouvons-nous adjoindre le pronom *on* français, à valeur potentiellement inclusive étant donné son indétermination, tandis que le conditionnel préfère les marques de troisième personne dans les deux cas.

A noter que le pronom *je,* pourtant bien plus représenté en français qu'en anglais, demeure six fois moins employé que *nous* au futur, le français privilégiant largement *nous* et *on* : comme il l'a déjà été amplement remarqué dans l'ensemble des études du courant ESP, l'acteur de la recherche est au centre du genre de l'article anglo-saxon, ce qui motive l'emploi important de verbes comme *to show* ou *to argue*, absents des relevés français qui leur préfèrent des verbes d'observation ou de parole comme *noter, remarquer, dire* ou *parler*. En outre, on a pu voir que le genre de l'article français semble globalement plus attaché à l'annonce du développement de l'article en son début, tandis que le genre anglo-saxon s'attacherait davantage à la récapitulation en fin d'article.

Dans la mesure où l'on observe de plus nombreuses similitudes entre *will* et le futur français, et *would* et le conditionnel français, la distinction effectuée entre les deux auxiliaires semble légitime ; si l'on reconsidère les configurations tactiques des deux futurs anglais et français (graphique 201), on peut dire que l'article anglo-saxon s'oriente davantage vers un futur postérieur au texte – et s'intéresse plus spécifiquement aux retombées de la recherche présentée : il s'inscrirait ainsi dans un processus de recherche plus global, dans lequel il n'est qu'inclus, tandis que le genre français serait plus autonome, voire autosuffisant. Ces observations nous semblent donc permettre d'objectiver de manière originale les caractères opposés (théorique *vs*. empirique) des deux conceptions française et anglo-saxonne de la Recherche.

Notons pour finir que le *simple present* anglo-saxon semble avoir une disposition tactique plus irrégulière qu'en français ; comparons ainsi les configurations des deux temps présent :



Graphique : Configurations tactiques du présent français et du simple present anglais

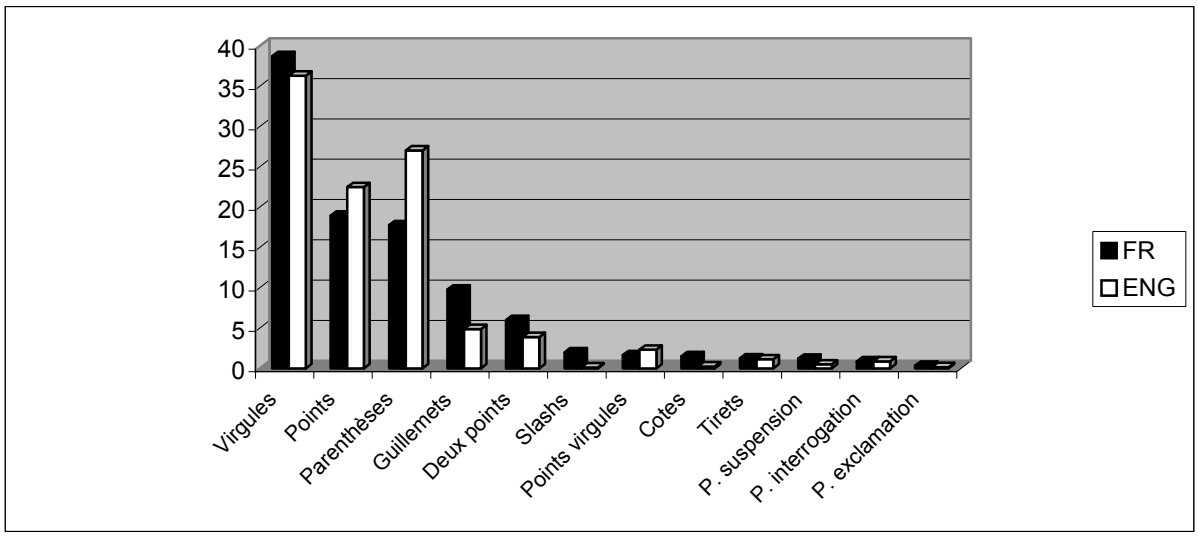
Si l'on observe une représentation progressive continue du présent, le genre anglo-saxon semble marqué par une rupture au quatrième dixième de son corps. Ce phénomène est peut-

être lié à la structure plus normée des textes (IMRAD ou structure approchante), et il serait intéressant de corréler la répartition des temps verbaux au séquençage des textes.

9.2.3.5. Des ponctuations

Les ponctuations sont inégalement représentées d'une langue à l'autre : si les virgules demeurent les premières ponctuations employées, l'anglais paraît recourir davantage aux parenthèses, qui supplantent les points dans le classement.

L'article français contient dans l'ensemble plus de virgules, de guillemets, de deux points, de slashs et de guillemets simples (rappelons que pour ces deux dernières ponctuations, les chiffres sont globalement peu fiables), tandis que le genre anglo-saxon contiendrait davantage de points, de parenthèses et de points virgules :



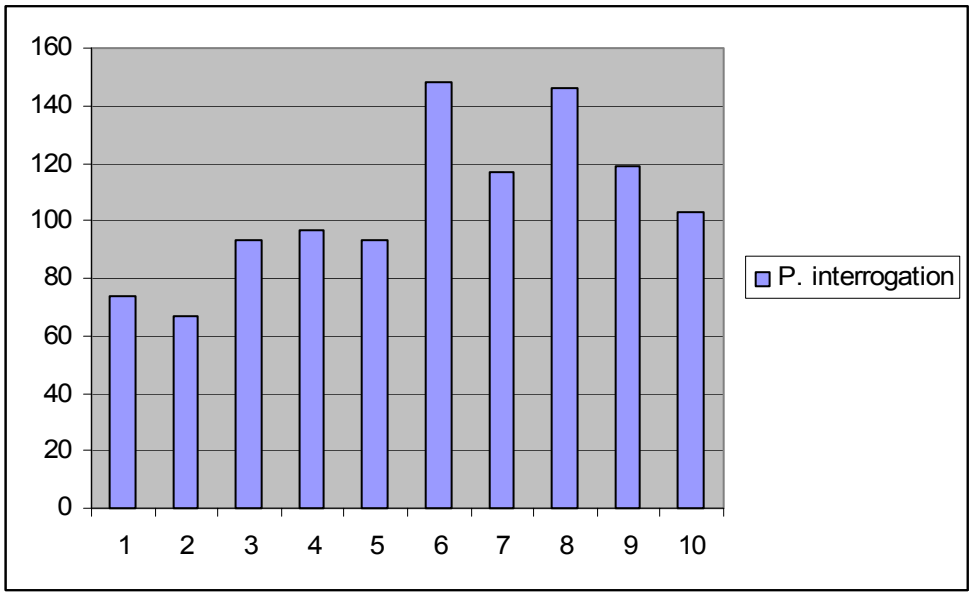
Graphique : Répartition contrastive des pronoms personnels sujets (moyennes absolues pondérées en % par texte)

Certains ponctuations sont significativement corrélées aux exemples des textes dans les textes anglo-saxons : les *deux points,* les *slashs,* les *points de suspension*, *d'interrogation* et *d'exclamation* semblent ainsi associés aux exemples oraux. Les quatre premiers sont significativement corrélés aux marqueurs de seconde personne (resp. +0.41 / +0.31 / +0.44,), tandis que les deux points et les points d'exclamation et d'interrogation sont associés aux interjections (resp. +0.6 / +0.4 / +0.55) et aux mois (+0.39 / +0.26 / +0.28). Les points de suspension sont quant à eux significativement corrélés à la forme *am* (+0.35).

Soulignons que les points d'interrogation sont corrélés aux deux points et aux points de suspension, alors que les slashs et les points d'exclamation sont intercorrélés.

Les genres français et anglais manifestent donc un usage similaire des ponctuations les plus expressives dans les exemples des textes ; on peut donc légitimement penser qu'elles soient intercorrélées avec d'autres descripteurs comme *I* ou *you* au sein d'un éventuel pôle *exemplification* dans la structure générique du genre anglo-saxon.

Si nous avions pu observer une courbe descendante des points d'interrogation dans les textes que nous avions associée à la mise en place de l'hypothèse et des questionnements de départ, on observe un phénomène quasi-inverse dans les textes anglais :



Graphique : Configuration tactique des points d'interrogation

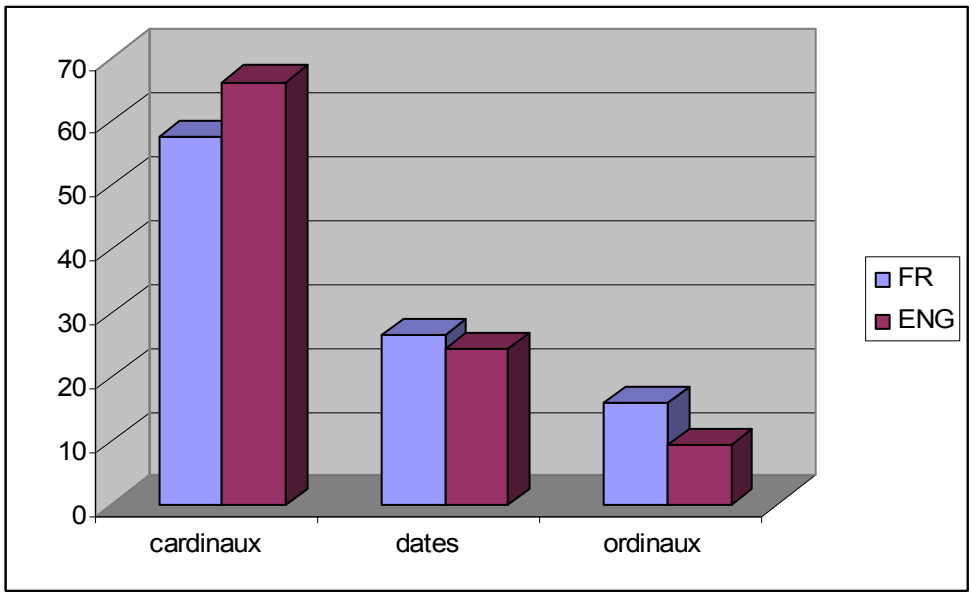
Le point d'interrogation atteint ainsi son apogée dans la deuxième partie de l'article – et toujours dans le sixième décile, qui semble décidément représenter l'un des lieux de prédilection de l'exemple. Il semble donc que le début d'article anglo-saxon est comparativement moins interrogatif, hypothèse qui serait à approfondir avec davantage de matériel.

Bien que les deux genres semblent ici distincts, on relève de nombreux usages communs : outre un emploi privilégié des ponctuations expressives dans les exemples, on remarque que les guillemets simples sont fortement associées aux éléments de langue étrangère dans les deux langues (*fr.* : + 0.17 / *eng.* : +0.66). De surcroît, virgules et parenthèses sont également en concurrence dans les textes anglais (-0.71) et on observe la même corrélation des parenthèses aux marqueurs formels (symboles **ZZ1**, numéraux, etc.), tandis que les virgules sont significativement corrélées à l'ensemble des adverbes et des connecteurs (classe 2), aux relatifs *whom* et aux tirets (vraisemblablement d'incise). Notons d'ailleurs que les guillemets s'opposent aux parenthèses (-0.44). Ces éléments permettent déjà de supposer que les textes plus appliqués ou plus formalisés ont un profil morphosyntaxique spécifique, comme ce que nous avions observé pour le français.

Parmi les éléments distinguant les deux genres, mentionnons de surcroît que dans les textes anglo-saxons les points-virgules s'opposent aux chiffres (-0.35) en s'associant aux dates (+0.37) ; s'il existe une dimension historico-narrative en anglais, les points-virgules y seraient donc associés, ce qui n'était pas le cas en français.

Fait notable, les points sont enfin significativement associés aux formes verbales conjuguées au *simple past* (*were*, *was* et **VVD**) tandis qu'ils s'opposent au *simple present* (**VV0** et **VVZ**) : les textes contenant plus de points, et vraisemblablement plus de phrases simples, privilégieraient donc le *simple past* au *simple present*.

9.2.3.6. Des numéraux



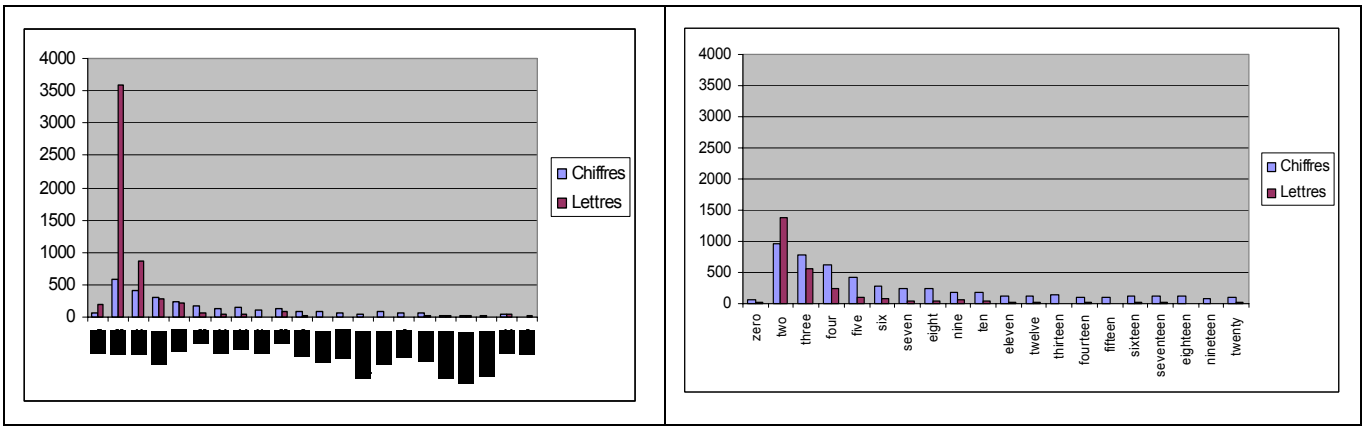
Graphique : Répartition contrastive des pronoms personnels sujets (moyennes absolues pondérées en % par texte)

A. Numéraux cardinaux

On observe d'abord un intérêt plus prononcé des anglo-saxons pour la quantification ; si les moyennes absolues obtenues par texte sont équivalentes (63.53 numéraux relevés dans les textes français en moyenne *vs.* 53.42 dans les textes anglo-saxons), le coefficient de variation obtenu est trois fois plus élevé dans les articles français (0.77, *vs.* 0.22 en anglais[15]) ; la présence de numéraux dans le versant anglo-saxon de l'article est ainsi stabilisée et donc *caractéristique* du genre.

A fortiori, si nous avions pu relever 22 nombres représentés en lettres en français, on en comptabilise deux fois plus en anglais (51) et on n'observe pas un emploi aussi important de la forme orthographiée du chiffre 2 – qu'on ne saurait d'ailleurs associer à un travail de quantification :

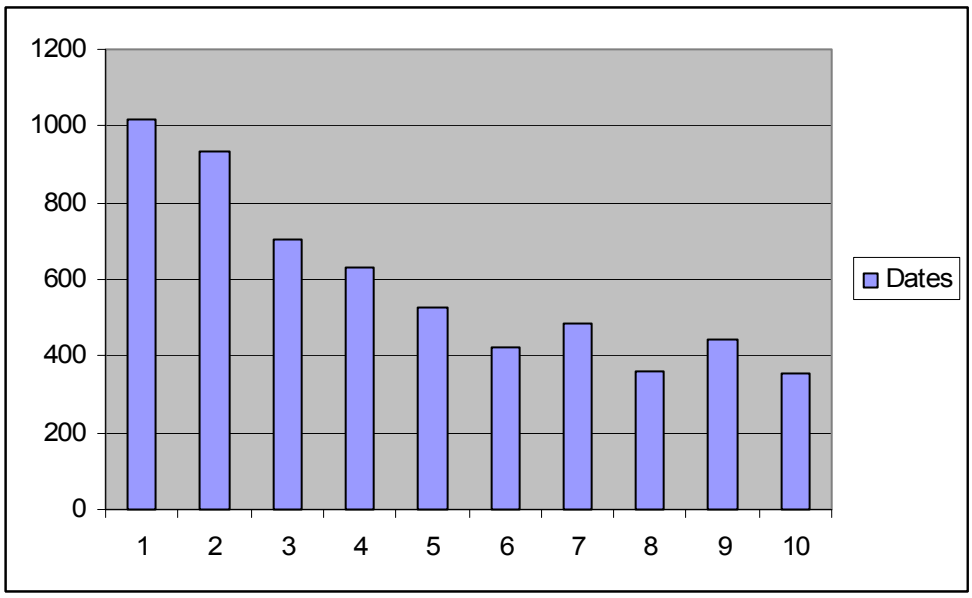
[15] Soulignons de surcroît que la variance obtenue est de 2379.36 en français *vs.* 141.41 en anglais, ce qui constitue un écart pour le moins remarquable.



Graphique : Formes lettres et chiffres des nombres (chiffres absolus) en français et en anglais

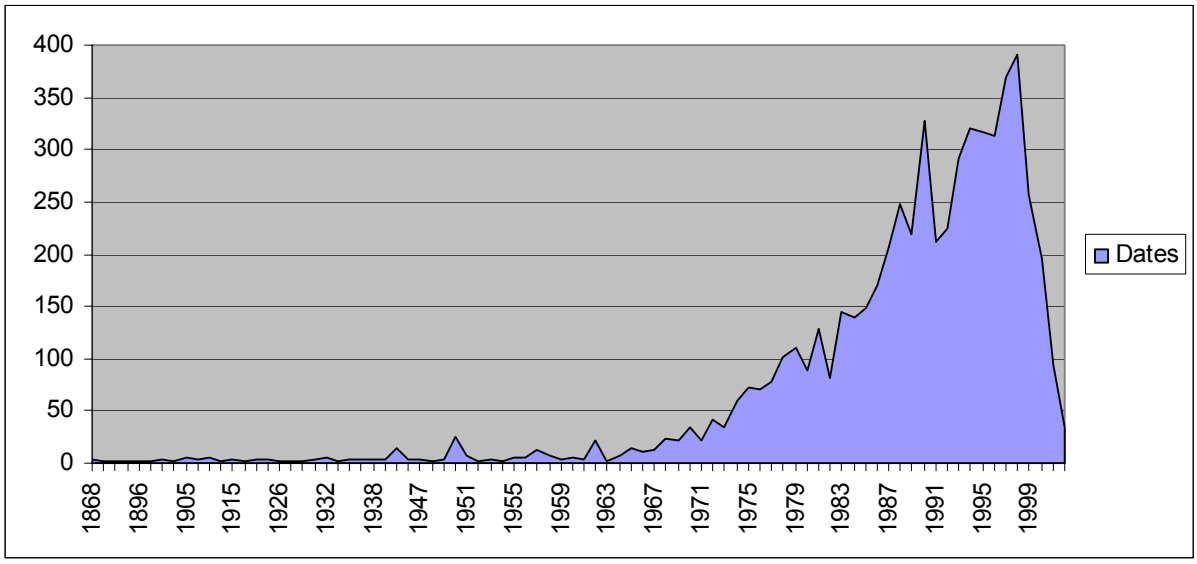
L'intérêt des anglo-saxons pour les chiffres et le quantifiable est ici manifeste, et ce phénomène semble encore une fois illustrer deux conceptions bien différentes de la recherche.

B. Dates

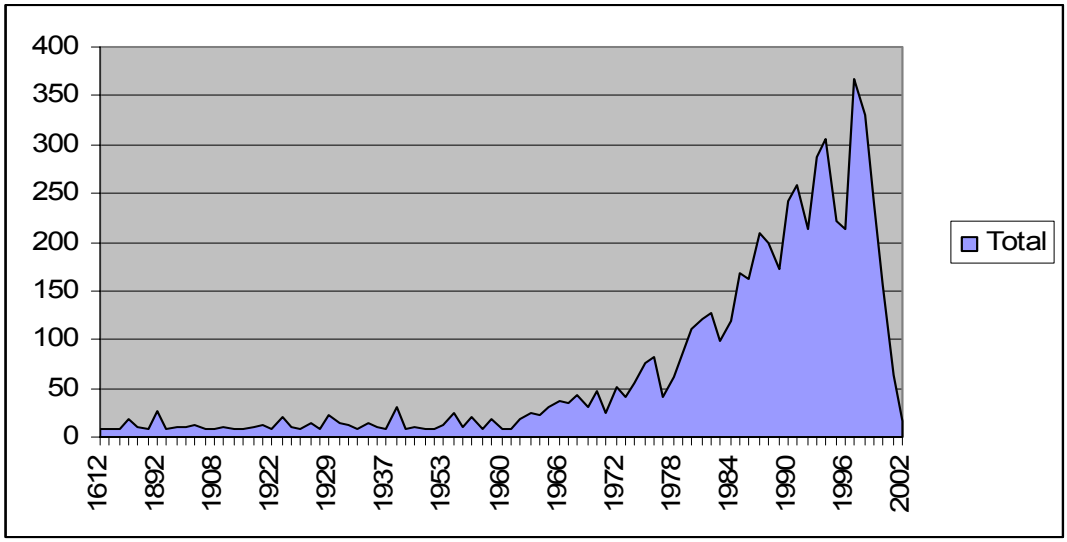


Graphique : Configuration tactique des dates

Si l'on remarque dans les deux corpus une décroissance significative des dates au fil du texte, avec un maximum atteint en début d'article, lieu de positionnement du chercheur dans (l'histoire de) son champ, les courbes de répartition des dates diffèrent sensiblement dans les deux genres :



Graphique : Répartition des dates apparaissant plus d'une fois dans le corpus anglo-saxon (chiffres absolus)



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

Bien que nous n'ayons écarté que les dates hapax du corpus anglo-saxon, les courbes ont une forme similaire : dans les deux cas, on note une croissance – irrégulière – des dates de 1960 à 2000, qui présuppose un empan citationnel, voire une durée de vie des textes scientifiques, de quarante années dans les deux communautés.

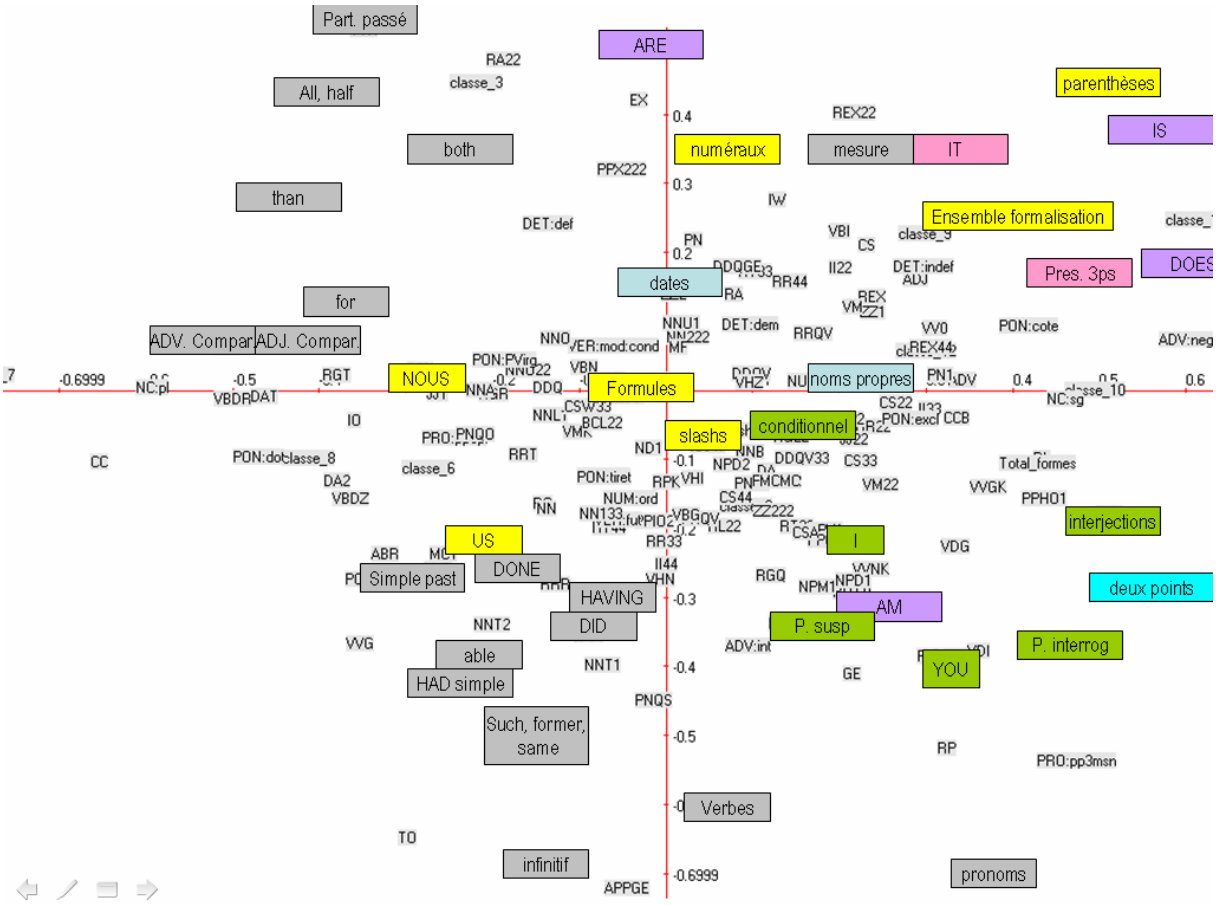
Fait notable, peu de dates-clés émergent avant 1960 dans les textes anglo-saxons, tandis que le fond citationnel français est stabilisé entre 1612 et 1960 : il conviendrait de réitérer l'expérience dans quelques dizaines d'années, afin d'apprécier la pérennisation des références anglo-saxonnes.

9.3. Axes d'organisation du genre de l'article en anglais

Afin d'être en mesure de comparer l'organisation morphosyntaxique des deux genres, nous avons mené une ACP sur les textes anglo-saxons.

9.3.1. Restriction des descripteurs

En dépit des modifications apportées, le système de description CLAWS 7 que nous avons adopté s'est avéré globalement problématique dans la section précédente : les descripteurs et leurs corrélations étaient souvent peu interprétables et de nombreuses variables se sont avérés génériquement marginales. *A fortiori*, les 236 observations prises en compte dessinent une structure difficilement comparable à celle du genre français, comme le montre le graphique qui suit :



Graphique : Positionnement des variables sur les deux premiers axes factoriels

On observe que la lecture du graphique est polluée par les nombreuses catégories adoptées, dont la pertinence descriptive est souvent discutable. Par exemple, les distinctions opérées au niveau des auxiliaires (*do, have,* etc.) multiplient les descripteurs et complexifient considérablement l'interprétation des données. Il en va de même des prépositions (distinction *for, if, with/without*, etc.) ou des temps verbaux (*are, is, am*, etc.) qui se distribuent en de nombreuses catégories entravant encore une fois l'interprétation et la comparaison des données.

C’est pour ces raisons que nous avons choisi d’effectuer une restriction des données : les observations ont été réduites à un ensemble de 80 descripteurs, que nous décrivons dans le tableau suivant :

Classe	Distinctions (conservées ou élaborées)	Contenu
1. Formalisation	ZZ1, ZZ221, FO	
2. Adverbes et connecteurs	ADV	Ensemble des adverbes
	CCB	Ensemble des connecteurs hors conjonctions de coordination
	ADV :int, ADV :neg, CC	
3. Adjectifs	ADJ	Ensemble des adjectifs
4. Pronoms	PRO :indef	PNX1 + PPX121 + PPX122 + PPX221 + PN + PN1 + PN121 (122) + PNQV + DB + DB2
	PRO :rel	DDQ + DDQGE + PNQO + PNQS
	Pronoms personnels	PRO :pp1sn, PRO :pp2, IT, PRO:pp3msn, PRO:pp1pl, PRO:pp3pl
	Pronoms clitiques	ME, HIM/HER, US, THEM
	Réflexifs	Refl:sg, Refl:pl
	PRO:poss (PPGE), EX (*there*)	
5. Verbes	VER :inf	TO + VVI + VBI + VHI + VDI
	VER :pres	VBM + VBZ + VBR + VD0 + VVZ + VDZ + VV0
	VER :aux :pper	VHZ + VH0
	Simple Past	VVD + VBDR + VBDZ + VDD
	Progressif	VVG + VBG + VDG + VHG
	VER :pper	VVN + VBN + VHN + VDN + VVNK
	VER :cond, VER :mod :cond, VER :fut, GoingTo (VVGK), HAD (VHD), VER:mod, VER:imp	
6. Déterminants	DET :indef	DET :indef + DA2
	DET :def, DET :dem, DET :poss (APPGE)	
7. Noms	NC :sg	NC :sg + NNA + NNB + NNL1 + NNT1 + NNU1 + NPD1 + NPM1
	NC :pl	NC :pl + NNO2 + NNT2 + NNU2 + NPD2
	NN	NN + ND1 + NN131 + NN21 + NN221 + NNO
	NP	

8. Prépositions	PREP	IF + II + IO + IW + GE
	PREP :1st	II21 + II31 + II41
9. Composite	FU	
10. Eléments de langue étrangère	FGW	
11. Ponctuations	Aucune modification	
12. Subordonnants	SUB	Ensemble des subordonnants
13. Interjections	INT	
14. Numéraux	Aucune modification	

Tableau : Système de descripteurs final

9.3.2. Analyse en Composantes Principales

9.3.2.1. Diagramme des valeurs propres

Le tableau ci-dessous décrit les 40 premiers facteurs de l'ACP : la première valeur propre, qui est ici comprise entre 1 et 80 est égale à 8.92. Le pourcentage d'inertie du premier facteur est de 11.16% (soit 8.92 variables) tandis que les quatre premiers facteurs en rendraient compte de 25.58 (pourcentage d'inertie de 31.98) :

Nb	Valeur propre	% d' inertie	% cumulé	
1	8.92	11.16	11.16	**********
2	6.36	7.95	19.12	**********
3	5.41	6.77	25.88	**********
4	4.88	6.10	31.98	**********
5	3.83	4.80	36.78	**********
6	3.07	3.84	40.62	**********
7	2.76	3.46	44.08	**********
8	2.61	3.27	47.35	**********
9	2.36	2.95	50.30	**********
10	2.20	2.76	53.06	**********
11	2.12	2.65	55.71	**********
12	2.01	2.52	58.23	**********
13	1.92	2.41	60.64	**********
14	1.83	2.29	62.93	**********
15	1.76	2.21	65.13	**********
16	1.70	2.13	67.27	**********
17	1.58	1.99	69.25	**********
18	1.51	1.90	71.15	**********
19	1.38	1.73	72.88	**********
20	1.31	1.64	74.52	**********
21	1.26	1.58	76.10	**********
22	1.22	1.53	77.63	**********
23	1.09	1.37	79.00	**********
24	1.01	1.27	80.27	**********
25	0.92	1.15	81.42	**********
26	0.91	1.14	82.56	**********
27	0.87	1.10	83.66	**********
28	0.83	1.04	84.70	**********
29	0.82	1.03	85.72	**********
30	0.77	0.96	86.69	**********
31	0.74	0.93	87.62	**********
32	0.67	0.84	88.46	**********
33	0.62	0.78	89.24	**********
34	0.60	0.76	90.00	**********
35	0.58	0.73	90.73	**********
36	0.55	0.70	91.43	**********
37	0.52	0.66	92.08	**********
38	0.50	0.64	92.72	**********
39	0.46	0.58	93.30	**********
40	0.43	0.54	93.84	**********

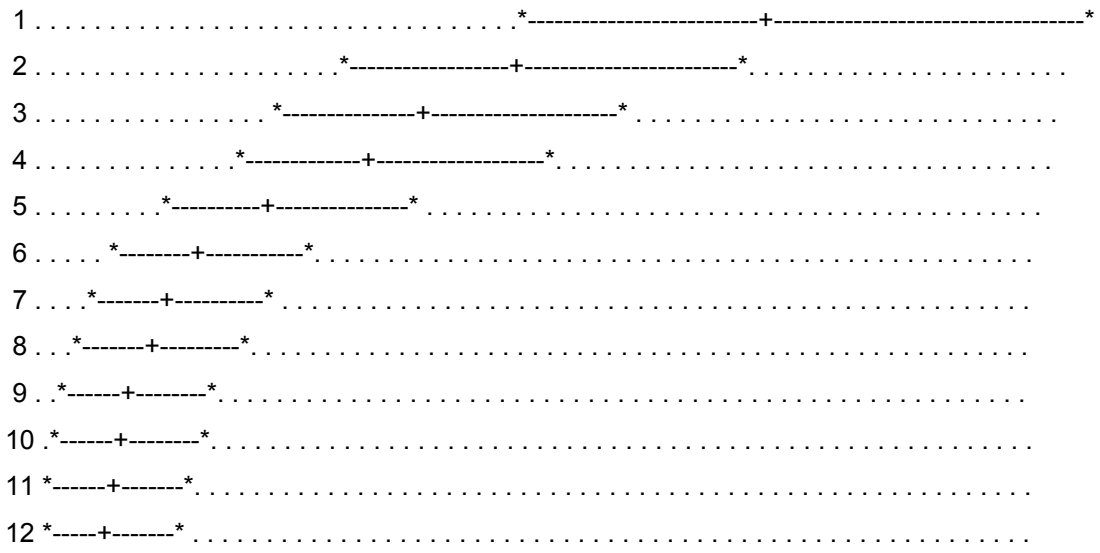
Tableau : Diagramme des 40 premières valeurs propres (sortie DTM)

Bien que les valeurs recueillies soient sensiblement plus élevées que celles que nous avions obtenues sur le corpus français, la décroissance obtenue est significativement plus régulière, ce qui traduit un nuage plus sphérique et présage d'un intérêt plus limité des facteurs.

L'observation des intervalles de confiance d'Anderson des quatre premières valeurs propres montre d'ailleurs qu'aucun facteur n'est significativement individualité – contrairement à ce que nous avions observé sur le corpus français, dans lequel le premier facteur s'individualisait de manière significative :

	Borne inférieure	**Valeur propre**	**Borne supérieure**
vp1	6.78	8.92	11.74
vp2	4.83	6.36	8.37
vp3	4.11	5.41	7.12
vp4	3.71	4.88	6.42

Tableau : Intervalles de confiance des quatre premières valeurs propres

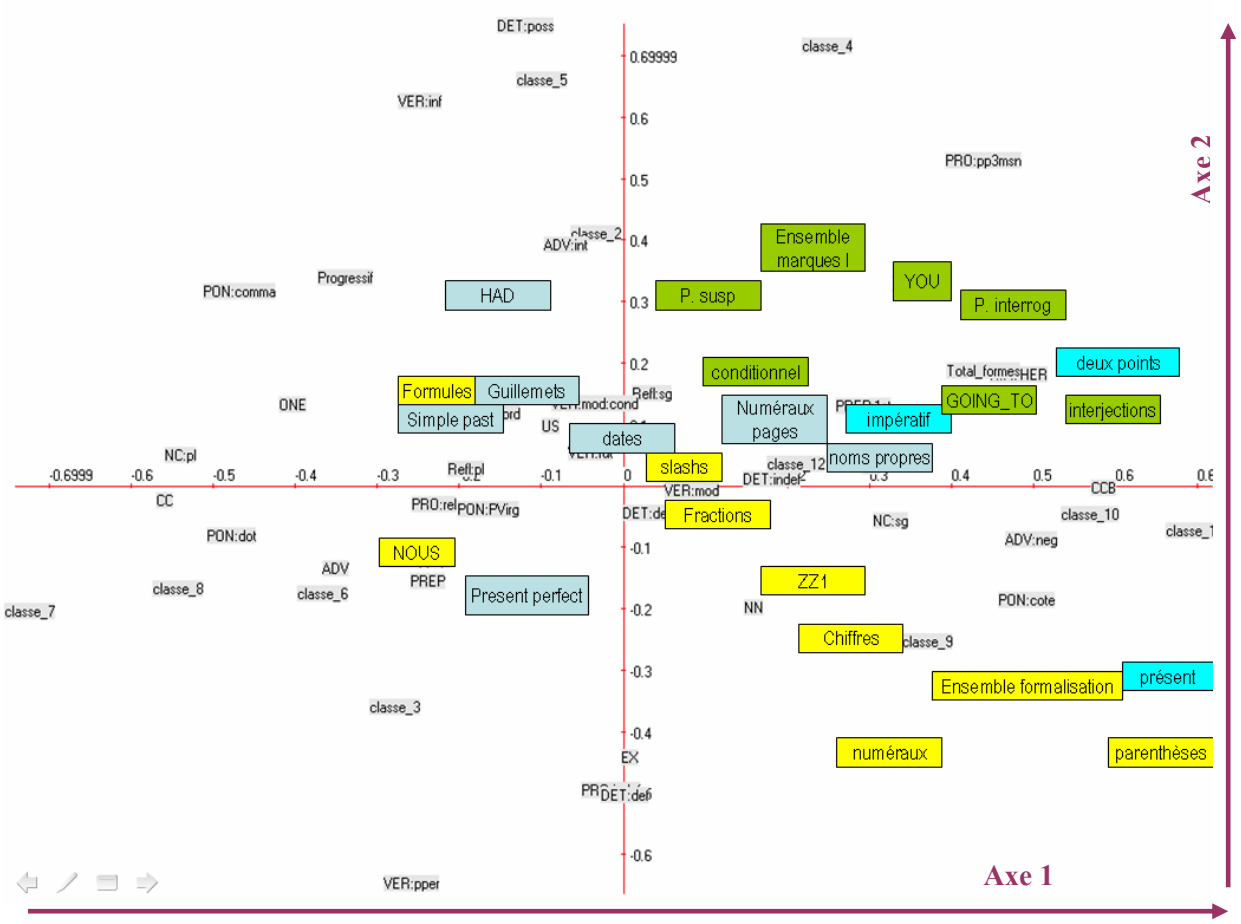


Graphique : Position relative des intervalles

Malgré ces résultats finalement peu encourageants quant à l'existence d'une structure générique du versant anglo-saxon du genre, on analysera les premiers facteurs obtenus qu'on tentera de contraster aux axes génériques français ; comme nous l'avions déjà souligné, il serait dommage de rejeter avec des critères statistiques des facteurs qui semblent interprétables (Escofier & Pagès, 1998).

9.3.2.2. Analyse des facteurs principaux

Examinons d'abord le premier plan factoriel obtenu – pour faciliter la comparaison, les groupements de descripteurs mis au jour au sein du genre français sont restitués :



Graphique : Positionnement des variables sur les deux premiers axes factoriels

L'axe 1 oppose d'abord le passé au présent, et distingue les descripteurs associés au *matériel linguistique et non linguistique* (corpus, formalisation) présent dans les articles ; en d'autres termes, le premier facteur est positivement corrélé aux éléments relevant finalement du *hors-texte*, tandis qu'il est négativement corrélé aux descripteurs du corps du texte – et décrivant d'ailleurs ce hors-texte.

Fait intéressant, on retrouve comme en français un groupement de marqueurs relevant d'un style plus oral et par conséquent des exemples des textes (en vert) qui s'oppose sur le deuxième axe aux marques de formalisation.

En revanche, deux dimensions sont globalement absentes : si le passé s'oppose au présent – de manière d'ailleurs bien plus nette en anglais qu'en français -, on ne saurait parler de pôle historico-narratif, dans la mesure où les noms propres, les dates, les indices de pagination (de citation) et les temps du passé ne s'isolent pas au sein d'un groupement intercorrélé. La dimension rhétorique regroupant l'ensemble des descripteurs spécifiques au discours scientifique (présent, *on,* deux points, impératif, etc.) semble également absente de l'organisation générique anglo-saxonne.

On retrouve la corrélation que nous avions observée entre *we* et le *present perfect* sur l'axe, bien que les descripteurs aient été réduits ; contrairement à ce que nous avions observé en français, *we* n'est pas associé, mais s'oppose au contraire aux descripteurs de formalisation.

Les descripteurs s'opposant aux marqueurs de formalisation et d'exemplification sont d'ailleurs intéressants : on observe ainsi un usage plus intensif du progressif, des points-virgules, des guillemets et du pronom *one* en corps de texte.

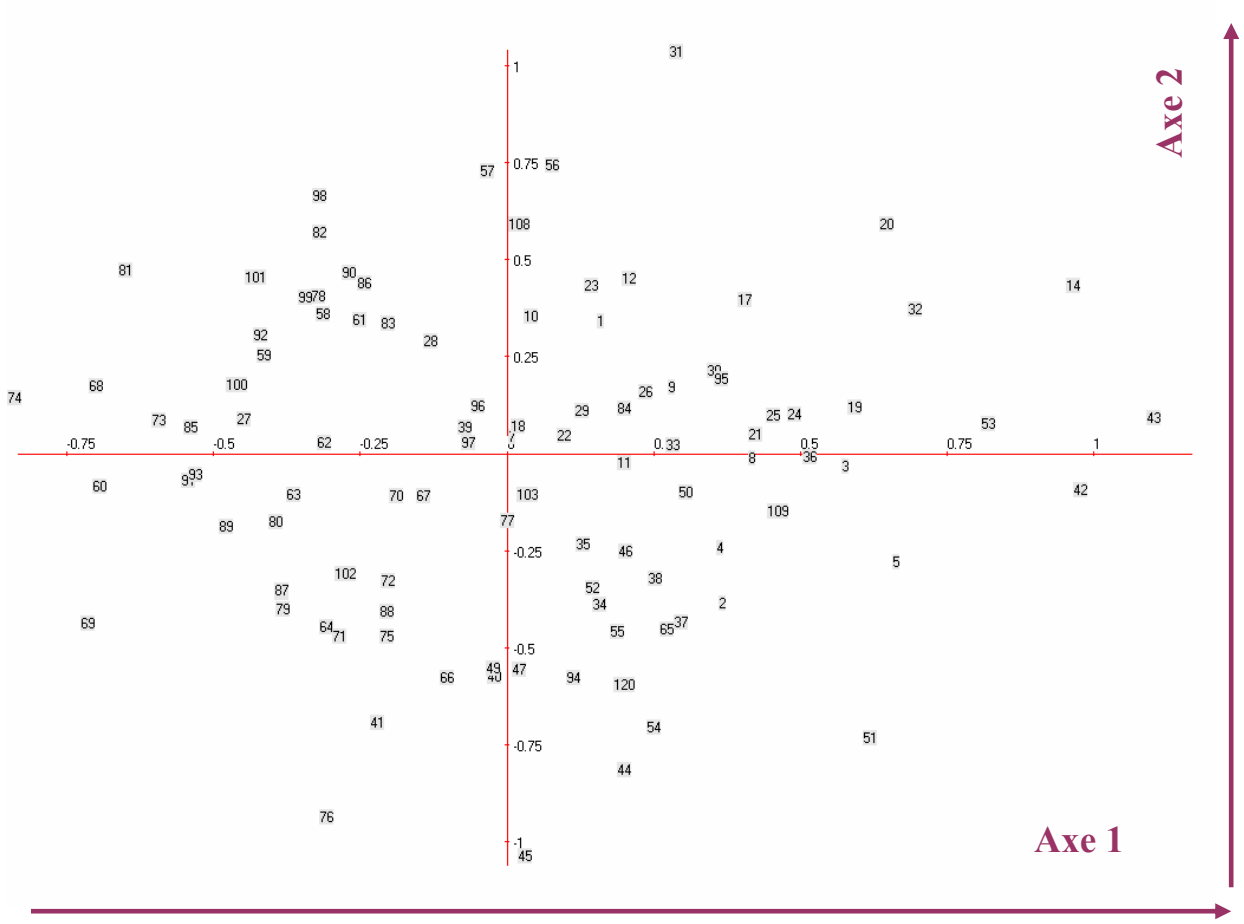
Les observations émises sont validées par l'examen de la carte de Kohonen obtenue sur les descripteurs (représentation des variables en 25 neurones, 5*5), qui permet de préciser les groupements de variables :

classe_12 VER:pres VER:mod NC:sg IT DET:indef CCB			US Refl:sg PON:tiret PON:guil NUM:tirets NUM:ord FO	 classe_3 VER:aux:pper PRO:pp1pl PON:comma ADV
VER:cond Total_formes PREP:1st DET:dem	classe_2	VER:inf DET:poss		Refl:pl Progressif PRO:rel
PON:susp PON:slash NUM:frac	classe_4	classe_5 ADV:int	VER:mod:cond ONE	THEM PRO:pp3pl NC:pl CC
classe_13 PRO:pp2 PON:int PON:excl ME	PRO:pp3msn	ZZ222 NUM:dat	PON:PVirg HAD	 classe_7 Simple_past PON:dot
classe_9 classe_10 VER:imp PON:cote PON:colon NUM:car NN HIM/HER GoingTo	 classe_11 classe_1 PON:par	 classe_14 ZZ1 NP ADV:neg	 EX	 classe_8 classe_6 VER:pper PRO:indef PREP DET:def

Graphique : Carte de Kohonen

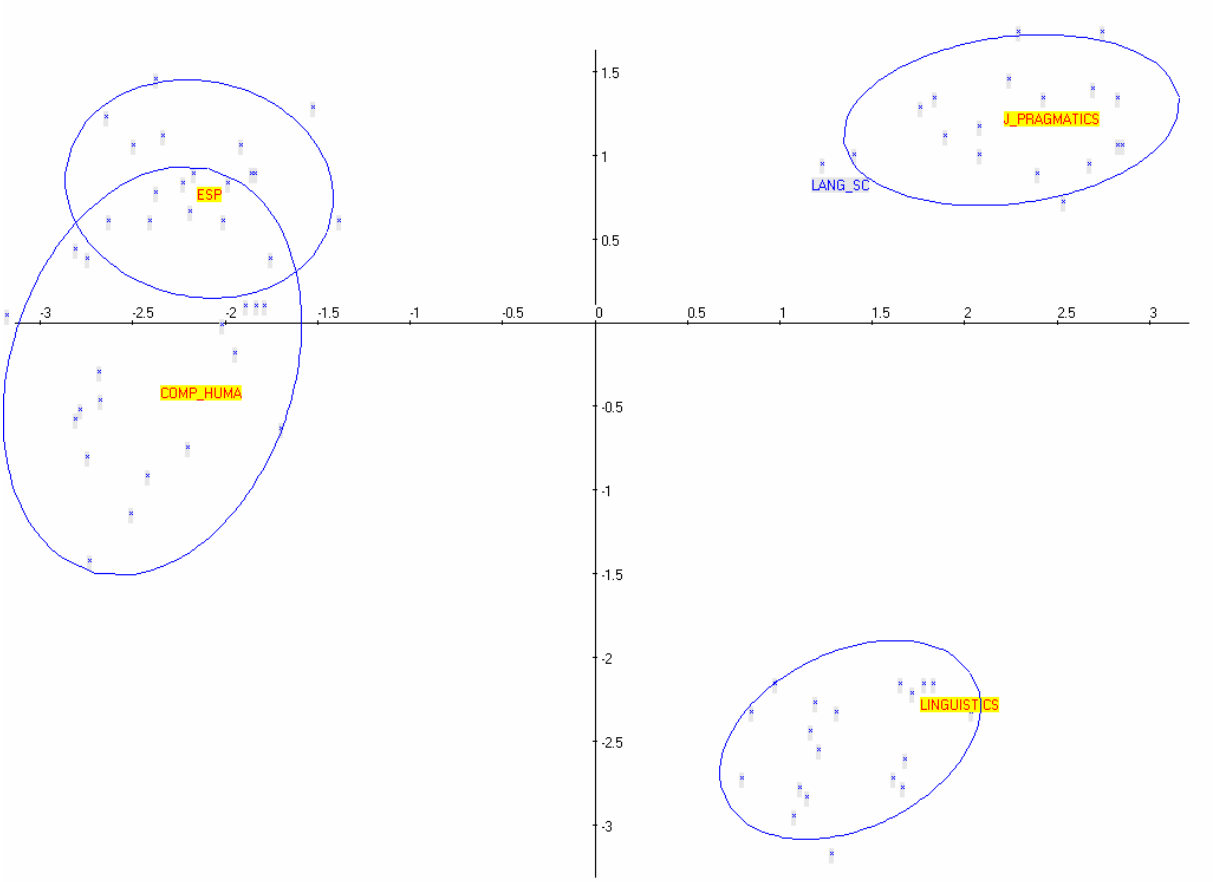
On observe bien un neurone contenant interjections, points d'interrogation et d'exclamation, pronoms *you* et *me*, tandis que les temps du passé sont diamétralement opposés au présent, qui parait ici bien plus opposé aux marques de formalisation qu'il ne l'était sur le plan factoriel.

Si l'on examine maintenant le positionnement des individus sur le premier plan factoriel, on n'observe pas d'éléments frontaliers comme en français. Le nuage est en effet bien plus homogène, ce qui laisse déjà penser que le genre anglais est plus normé :



Graphique : Positionnement des individus sur les deux premiers axes factoriels

Si nous avions pu observer en français une orientation des revues plus historique sur le versant négatif de l'axe 1 (et le versant positif de l'axe 2), on observe un phénomène comparable, et même plus manifeste en anglais : comme le montre le graphique qui suit, les quatre revues sont significativement distinctes, malgré le recouvrement partiel de *Computers and the Humanities* et *ESP* :



Graphique : Ellipses de confiance des revues sur le premier plan factoriel

Les revues *ESP* et *Computers and the Humanities* sont ainsi positionnées sur le pôle plus historique mis au jour précédemment, tandis que *Journal of Pragmatics* se situe au niveau du pôle oral : la revue s'intéresse en effet particulièrement au genre de la conversation, à l'instar de l'ensemble du courant pragmatique anglo-saxon. Les articles de *Linguistics* dont nous disposons semblent comparativement plus formalisés comme l'indique leur emplacement.

Plus spécialisées et plus délimitées, les revues anglo-saxonnes semblent peut-être plus identifiables que les revues françaises sur le plan morphosyntaxique, ce qui serait à valider ultérieurement sur une collection textuelle plus large.

9.3.3. Synthèse

Malgré les problèmes d'équivalences posés (systèmes d'étiquetage, textes avec/sans exemples, etc.), les descripteurs employés et leur traitement par ACP nous ont permis de mettre au jour des lieux de contraste et d'équivalence qui n'auraient pu l'être par d'autres biais : le genre de l'article s'articule dans les deux langues entre trois pôles morphosyntaxiques principaux reflétant des courants plus historiques, plus appliqués et plus formalisés. Ce résultat nous semble particulièrement intéresser la recherche d'information et la description des genres scientifiques et nous envisageons de préciser et de poursuivre l'analyse dans nos recherches futures.