Chapitre 2 : Méthodologie : mise en place de l'observatoire de genre

Le chapitre suivant est dédié à la mise en place méthodologique de notre observatoire du genre : après avoir décrit les corpus mobilisés par notre entreprise (2.1.), on présentera les observations morphosyntaxiques qui fonderont nos descriptions et nos analyses (2.2.).

2.1. Corpus

Dédiée à l'observation du genre de l'article scientifique de revue linguistique, la présente thèse mobilise six corpus, soit 586 textes au total :

- un corpus génériquement homogène d'articles scientifiques français de revues linguistiques (désormais corpus ASLF) composé de 224 textes ;
- et cinq corpus de comparaison :
  - ✓ un corpus « Auteurs » destiné à examiner les variations stylistiques du genre de l'article de linguistique (122 articles) ;
  - ✓ un corpus « Mécanique » voué à observer les variations domaniales du genre de l'article (49 textes) ;
  - ✓ deux corpus « Présentations de revues » et « Comptes rendus » destinés à explorer les variations morphosyntaxiques et lexicales des textes scientifiques d'un genre à l'autre dans le domaine linguistique (53 comptes rendus et 45 présentations) ;
  - ✓ un corpus génériquement homogène d'articles scientifiques anglo-saxons de revues linguistiques (dorénavant ASLA) et voué à observer les variations génériques de l'article en français et en anglais (191 textes constitués, 103 textes exploités).

On soulignera que les deux premiers corpus de comparaison ont été construits dans le cadre de projets collaboratifs relativement indépendants de la présente thèse : « Mécanique » est la propriété de V. Clavier (Gresec, Grenoble), tandis que « Auteurs » a été constitué et collecté par F. Rinck (Lidilem, Grenoble) et moi.

On présentera la constitution et les propriétés du corpus ASLF avant de détailler les cinq corpus de contraste.

2.1.1. Corpus ASLF

2.1.1.1. Critères de sélection

De manière générale, la constitution du corpus est soumise aux objectifs de la recherche : notre démarche est globalement inductive, et elle vise à explorer l'organisation du genre de l'article. Par conséquent, le corpus ASLF se doit d'être *génériquement homogène*, et *représentatif* du genre de l'article de revue linguistique. Malgré le périmètre particulièrement délimité de notre objet (homogénéité discursive, générique et domaniale), cette contrainte de représentativité demeure délicate, voire illusoire, notamment parce que la linguistique se

scinde en une diversité de sous-domaines et d'écoles probablement régis par des normes linguistiques spécifiques que nous maîtrisons encore mal à l'heure actuelle.

La représentativité d'un corpus doit donc être pensée en tant que principe de constitution, et comme le rappelle Habert (2000), on se doit d'éviter les deux types généraux d'erreurs statistiques menaçant les généralisations (Biber, 1993) : « l'incertitude » (*random error*) et la « déformation » (*bias error*). Il serait ainsi peu pertinent de prétendre décrire un genre avec un échantillon de dix textes, ou d'observer l'article de revue linguistique à partir de textes issus du seul sous-domaine syntaxique.

Ce sont les critères de sélection combinés à l'histoire de la récolte qui nous permettront d'estimer la représentativité, et les biais éventuels de notre corpus.

Plutôt qu'une sélection de chaque article un à un[1], il a été choisi un ensemble de revues accréditées dans les deux champs scientifiques linguistiques francophone et anglophone, ce qui garantit d'abord l'attestation et la conformité de chacun des articles du corpus. Puisqu'ils sont généralement thématiques, les numéros de revues ont été conservés dans leur intégralité (hors articles de langue étrangère bien entendu), ce qui nous permettra d'évaluer l'impact d'un thème sur les pratiques rédactionnelles. Le genre s'observant d'abord en synchronie, c'est l'année de publication 2000 qui été arrêtée.

Si nous avions présélectionné un ensemble de 23 numéros de revues distinctes en amont de la collecte, à l'aide du *Répertoire des revues francophones en sciences du langage* et de l'avis éclairé de plusieurs experts du champ, il nous a rapidement fallu modifier le panel de départ en fonction des fruits de la récolte : en effet, nous n'avons pas obtenu les droits d'exploitation et les archives de l'ensemble des revues choisies[2], tandis que de nombreuses archives de revues que nous ne pensions pas inclure au départ nous sont parvenues.

Le choix du corpus est donc apparu en partie soumis à des critères pratiques de disponibilité, voire d'*existence d'archives* des revues : de nombreux périodiques, comme *La linguistique* par exemple, n'ont en effet pas d'archives numériques, ce qui nous a fortement interpellée quant à la pérennité du patrimoine scientifique linguistique français. Par conséquent et à sa manière, notre entreprise contribue à la sauvegarde de ce capital numérique, ce que nous n'avions pas mesuré au départ.

Les données ayant été rassemblées par des biais différents[3] (prise de contact avec le responsable d'un numéro ou l'éditeur d'une revue, etc.), les revues recueillies se sont au final trouvées très inégalement représentées : par exemple, nous disposons des archives complètes des *Cahiers de Praxématique*, de la revue *LINX* de Paris X-Nanterre ou de la *Revue de Sémantique et Pragmatique* d'Orléans, tandis que nous n'avons pu obtenir qu'un seul exemplaire de la revue *Langage*.

De ce fait, nous avons dû sélectionner certains numéros au détriment d'autres, afin d'atteindre une représentation satisfaisante des revues ; l'année de publication a été

[1] Les articles d'une revue ayant déjà été sélectionnés par des experts, il serait peu pertinent de mettre en œuvre une seconde procédure de sélection.

[2] Sans compter les délais de réception des textes : par exemple, nous attendons toujours la revue TAL, pour laquelle nous avions obtenu tous les accords et promesses possibles…

[3] Je remercie d'ailleurs tout particulièrement F. Rastier et G. Bergounioux, à qui je dois la grande majorité du corpus.

déterminante dans cette sélection, et ce sont les numéros datés autour de 2000 qui ont été prioritairement élus – dans la mesure du possible, car certaines revues, comme les *Cahiers du CIEL* ont disparu peu avant.

Certains sous-domaines linguistiques ont d'emblée été écartés de la sélection : nous avons ainsi pris la décision d'exclure les textes et les numéros trop formalisés, étant donné leur sémiotique particulière, qui ne pourrait être pleinement appréhendée dans notre cadre descriptif (*v.* Herreman, 1997). Les domaines phonologique, phonétique ou de logique mathématique sont ainsi peu, voire non représentés dans le corpus, d'autant qu'à l'inverse d'autres domaines linguistiques, ils sont globalement internationalisés (les articles majeurs sont publiés en anglais).

Etant donné que nous souhaitions observer cet axe de manière précise et dans le cadre d'un chapitre autonome, nous avons tâché de *neutraliser* la variation stylistique *Auteur* du corpus : avec le jeu des co-auteurs, 226 linguistes sont représentés dans les 224 textes du corpus. Si la très grande majorité des auteurs n'apparaissent qu'une fois dans le corpus, on relève quatre textes de Combettes (l'un d'entre eux co-écrit avec Prévost), trois de Cortès (dont un article co-écrit avec Szabo), trois de Crévenat-Werner, trois de François (dont un texte co-publié avec Manguin) et trois de Paillard (dont une co-écriture avec Kisseleva).

Au total, ce sont donc 32 numéros de 11 revues de linguistique française que nous avons retenus, soit un ensemble de 224 textes.

2.1.1.2. Composition

Onze revues sont représentées dans le corpus ASLF :

- ✓ Les *Cahiers de Praxématique* (Montpellier) ;
- ✓ les *Cahiers du CIEL* (Paris 7) ;
- ✓ *Histoire, Epistémologie, Langage* (SHESL, CNRS, Paris 7) ;
- ✓ *Langages* (Larousse) *;*
- ✓ *Langue française* (Larousse) ;
- ✓ *LINX* (Paris X-Nanterre)
- ✓ La *Revue de Sémantique et Pragmatique* (Orléans) ;
- ✓ *Scolia* (Strasbourg) *;*
- ✓ *Sémiotiques* (ILF) *;*
- ✓ *Syntaxe et Sémantique* (Caen) ;
- ✓ *Verbum* (Nancy).

La plupart sont des périodiques universitaires, de visée et d'ouverture d'ailleurs distinctes : les *Cahiers du CIEL*, qui constituent la publication du Centre Interlangue d'Etudes en Lexicologie (C. Cortès, Paris 7) ou *LINX*, la revue des linguistes de l'université Paris X – Nanterre, publient souvent, voire exclusivement pour *CIEL*, les membres de leurs laboratoires de rattachement respectifs. En revanche, la *Revue de Sémantique et Pragmatique* ou les *Cahiers de Praxématique*, publiés à Orléans et Montpellier, ont opté pour une spécialisation linguistique, élargissant le champ des contributeurs au niveau national.

Toutes les revues sélectionnées relèvent du domaine linguistique. Si certaines revues sont très générales, d'autres sont plus spécialisées et ne concernent qu'une branche (lexicologie, sémantique, syntaxe, etc.) ou qu'une théorie particulière de la linguistique. Ces spécialisations

ont peu affecté la sélection ; le fait que la revue soit *attestée* dans le champ est effectivement plus déterminant[4]. Nous avons cependant cherché à éviter, dans la mesure du possible, la sur-représentation d'un sous domaine, ou d'un courant linguistique spécifique.

Le tableau qui suit présente les caractéristiques des 32 numéros de revues sélectionnés – les références des 224 textes (identifiants, auteurs et titres) sont présentées en annexe 1 :

ID	Revue	Editeur	Thème	Référence	Lieu Publication	Nb Articles
1	Cahiers de Praxématique	Kanellos	Sémantique de l'intertexte	N°33, 1999	Montpellier	6
2	Cahiers de Praxématique	Bosredon, Tamba et Petit	Linguistique de la dénomination	N°36, 2001	Montpellier	8
32	Cahiers de Praxématique	Schnedecker et Theissen	Topicalisation et partition	N°37, 2001	Montpellier	6
3	Cahiers du CIEL	Cortès	Théories et Pratiques du Lexique	1994-5	Paris	9
4	Cahiers du CIEL	Cortès	Problèmes de classement des unités lexicales	Déc. 1996-7	Paris	7
7	HEL	Lallot	Horizons de la grammaire alexandrine	Vol. 22, Fasc. 1, 2000	Paris	3
8	HEL	Archaimbault et Léon	Le dialogue, un objet d'étude?	Vol. 22, Fasc. 1, 2000	Paris	4
9	HEL	Lallot	Horizons de la grammaire alexandrine	Vol. 22, Fasc. 2, 2000	Paris	3
10	Langage	Arnavielle	Participe présent et gérondif	N°149, 2003	Paris	7
11	Langue française	Bergounioux	La parole intérieure	N°132, 2001		7
12	Langue française	Franckel	Le lexique entre identité et variation	N°133, 2001		8
13	Langue française	Cuxac	La langue des signes: enjeux institutionnels et linguistiques	N°137, 2003		5
14	LINX	Trévise	L'hypothétique	N°41, 1999	Paris	8
15	LINX	Gadet et Wachs	Approches sociolinguistiques du plan phonique	N°42, 2000	Paris	11
16	LINX	Sörés et Marchello-Nizia	Invariants et variables dans les langues. Études typologiques	N°45, 2001	Paris	12
17	LINX	Anis et Kleiber	Du sens au sens	N°47, 2002	Paris	6
5	RSP	Bergounioux	Approches	N°6, déc.	Orléans	7

[4] Certaines revues sont effectivement plus *attestées* que d'autres par les linguistes du champ, sans égard au domaine ou au courant traité.

			Sémantiques des prépositions	1999		
6	RSP	Bergounioux	Non thématique	N°8, déc. 2000	Orléans	5
18	RSP	Bergounioux	Les connecteurs entre langue et discours	N°5, juin 1999	Orléans	4
19	RSP	Bergounioux	Non thématique	N°7, juin 2000	Orléans	7
20	Scolia	Reichler-Béguelin	Problèmes de sémantique et de relations entre micro- et macro-syntaxe	N°5, 1995	Strasbourg	15
21	Scolia	Schmoll	Contexte(s)	N°6, 1996	Strasbourg	12
22	Sémiotiques	Habert	Dépasser les sens iniques dans l'accès automatisé aux textes	N°17, déc. 1999		5
23	Sémiotiques	Blès et Samain	Incidences de l'impossible dans le langage	N°18&19, déc. 2000		6
30	Syntaxe et Sémantique	Cordier, François et Victorri	La sémantique du lexique verbal	N°2, 2001	Caen	11
31	Syntaxe et Sémantique	Ledegen et Rossi-Gensane	Colloque Cologne, Les Grammaires du français et les "mots-outils"	N°3, 2002	Caen	7
24	Verbum	Benninger, Lagae et Carlier	Autour du futur	N°22-3, 2000	Nancy	5
25	Verbum	Cornish	Référence discursive et accessibilité cognitive	N°22-1, 2000	Nancy	5
26	Verbum	Schnedecker et Bianco	Référence (pro) nominale plurielle, aspects linguistiques et psycholinguistiques	N°22-4, 2000	Nancy	5
27	Verbum	Péry-Woodley	Cohérence et relations de discours à l'écrit	N°23-1, 2001	Nancy	4
28	Verbum	Tyvaert	Sémantique des verbes, Nouvelles approches	N°23-4, 2001	Nancy	5
29	Verbum	Charolles, Le Goffic et Morel	Y a-t-il une syntaxe au-delà de la phrase ?	N°24-1-2, 2002	Nancy	11

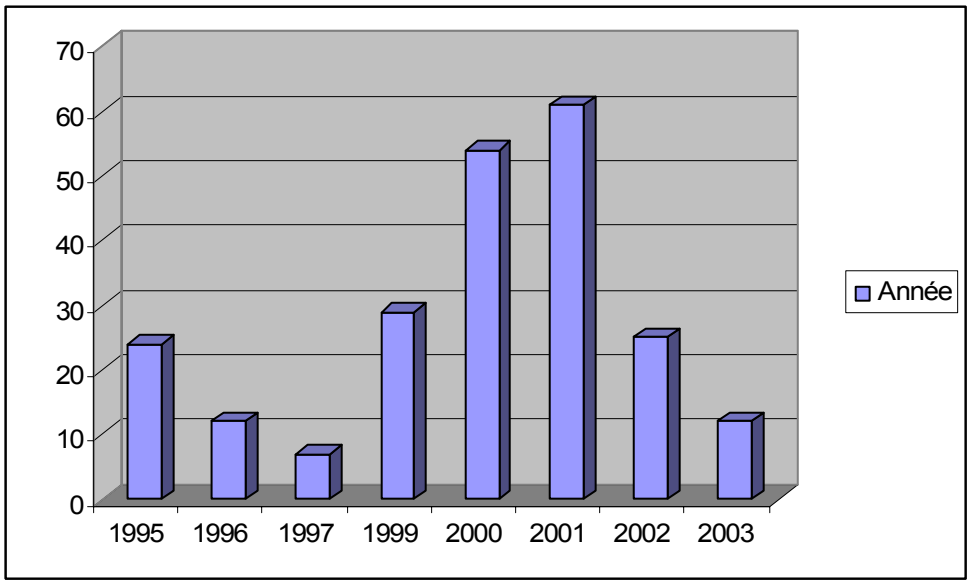
Tableau : Constitution du corpus par numéro de revue

Soulignons la présence d'une revue constituée non pas d'articles, mais d'*articles dans des actes* (ID 31) : étant donné qu'il n'est pas certain que les deux catégories soient linguistiquement distinctes, les articles dans des actes étant soumis à des procédures de sélection équivalentes et n'étant normalement pas plus « oralisés » que les articles, le numéro

a été adjoint au corpus sans réserves particulières. On observera néanmoins son caractère éventuellement distinct dans le chapitre 3.

Par ailleurs et bien que nous ayons veillé à éviter toute sur-représentation d'un domaine linguistique, force est de constater que notre corpus comporte une majorité d'articles de *sémantique*. La branche sémantique française étant bien développée, il n'est pas impensable de supposer que le courant linguistique dans son ensemble est plus orienté vers ce domaine – ce qui est peu vérifiable.

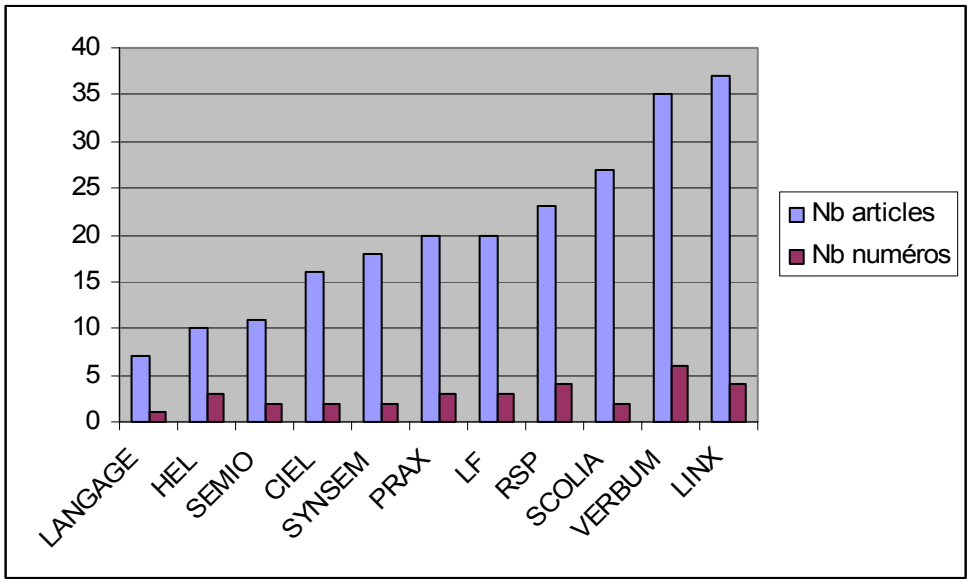
Comme nous l'avons déjà évoqué, c'est l'année 2000 qui a été privilégiée lors de la constitution du corpus : si 2001 est finalement l'année la plus représentée, on note que 75% des textes du corpus sont bien datés entre 1999 et 2002. L'année 1995 est honorablement représentée dans le corpus, du fait de deux numéros (identifiants 3 et 20) particulièrement denses :



Graphique : Représentation des années de publication

On s'attachera ultérieurement à vérifier l'impact éventuel de ce facteur sur les résultats (*v.* chapitre 4).

Certaines revues sont enfin plus représentées que d'autres ; la pondération des revues et des numéros s'est avéré délicate, d'une part parce que nous ne disposions pas des mêmes masses textuelles pour chaque périodique, et d'autre part parce que les numéros contiennent un nombre pour le moins inégal de textes :



Graphique : Répartition des articles et des numéros de revue

LINX est ainsi particulièrement dense, avec une moyenne de 9.25 textes par numéro, tandis que les fascicules d'HEL contiennent le plus petit nombre d'articles (3.33 en moyenne).

Le caractère significativement distinct des revues sera évalué au cours du chapitre 3.

Examinons enfin la provenance géographique des textes en isolant 47 articles, pour lesquels nous ne disposons pas de l'information : Paris est de manière non surprenante la ville la plus représentée du corpus (53 articles).

Parmi les pays les plus représentés, la Suisse, berceau du structuralisme, est le pays le plus représenté du corpus (18 articles), suivie du Canada (5) et de l'Allemagne (4). Peu d'articles sont issus de pays non francophones : on ne relève ainsi que trois articles d'auteurs américains et anglais.

La position dominante de Strasbourg (17 textes et deuxième rang) est corrélée à la revue strasbourgeoise Scolia, qui fédère les deux tiers des articles d'auteurs strasbourgeois relevés.

Paris étant la ville la plus représentée, nous avons tenu à préciser cette donnée en observant les villes parisiennes les plus figurées, ce qui est éclairant quant à la délimitation institutionnelle parisienne de la communauté linguistique française : trois institutions se détachent nettement : Paris VII (18 textes), X (13) et III (7). La position dominante de Paris VII est en partie due aux *Cahiers du CIEL*, qui sont fortement associés à Paris VII, et dans une moindre mesure à Paris III, tandis que LINX participe à la représentation importante de Paris X-Nanterre dans le corpus.

Le corpus est ainsi géographiquement marqué, du fait des revues sélectionnées ; cela ne nous semble pas problématique pour autant, ces déterminations étant inévitables. Il est en revanche important de les mentionner afin de circonscrire la portée et l'inscription sociale du corpus, d'autant que ces informations permettent de saisir certaines délimitations géographiques et institutionnelles de la communauté linguistique française.

ASLF étant exploré de manière progressive et détaillée au fil de la thèse, nous n'approfondirons pas davantage sa description.

2.1.1.3. (Pré-)traitement

A. Normalisation des textes

Si les outils de traitement automatique des langues autorisent des types de format différents (.TXT, .DOC, .RTF, etc.), c'est le format le moins enrichi et le plus accepté .TXT que nous avons d'abord adopté.

Bien qu'il nous semblât au départ évident de conserver l'intégralité des textes, certaines composantes du péri-texte de l'article ont été écartées en raison de leurs propriétés spécifiques et somme tout autonomes du corps, ou du *body* de l'article : l'inclusion de la bibliographie affecte par exemple considérablement le profil morphosyntaxique du texte. A l'issue d'une analyse pilote effectuée sur plusieurs textes, il s'avère qu'elle modifie non seulement le nombre de noms propres du texte, mais encore le nombre de paragraphes, de ponctèmes, de chiffres, etc. En outre, le nombre de références varie largement d'un texte à l'autre, à l'instar des notes de bas de page, qui ont également été écartées : les notes nous semblent en effet mériter une étude à part entière, qui ne sera malheureusement pas menée dans le cadre de cette thèse, mais que nous envisageons de conduire par la suite ; il serait ainsi intéressant de déterminer le statut des notes de linguistique, et d'observer si elles dissimulent d'occultes batailles académiques (Grafton, 1999).

Les résumés (et *abstracts* en anglais) ont également été supprimés, dans la mesure où la très grande majorité des textes sélectionnés n'en contenait aucun : composant avéré de l'article anglo-saxon, le résumé ne semble pas – encore – constituer une partie établie et caractéristique du genre français, ou de son péri-texte : la revue LINX rassemble ainsi les résumés des textes au sein d'une section isolée de la revue, ce qui illustre bien leur statut non péri-textuel.

Le développement probable de bases d'indexation des revues françaises et francophones devrait pourtant entraîner son essor, le résumé représentant la « vitrine » de l'article, et pouvant éventuellement préciser, voire se substituer à certains éléments de son introduction.

Bibliographies, notes de bas de page et résumés ont été systématiquement isolés des corps d'articles, dont nous avons par contre cherché à préserver l'intégrité. Ainsi, les citations et les exemples des textes ont vraisemblablement une linguistique toute spécifique (*v.* chapitre 5), qui pervertirait les caractéristiques du corps de texte. Comme les notes de bas de page et les bibliographies, leur nombre varie d'un texte à l'autre ; dans la mesure où elles représentent des configurations optatives qui ne sont pas positionnellement définies (notes et bibliographies sont particulièrement identifiables, du fait de leur format et de leur place dans l'article), nous avons choisi de les baliser plutôt que de les supprimer, ce qui nous permettra de contraster les textes avec / sans ces configurations.

B. Balisage

Comme on le verra *infra*, l'annotation morphosyntaxique a été convertie en XML ; conquise par la TEI (Text Encoding Initiative), nous avons systématiquement appliqué des balises conformes à ses recommandations (TEI P4), afin de garantir la portabilité, l'échange et à terme, la pérennité du corpus.

Un ensemble de balises et un en-tête (*header*) ont été constitués (*v.* annexe 5) : les éléments de structuration du document (<front>, <body>, <back>, titres et niveaux de division, paragraphes, etc.), de même que les composants de l'article (références, notes,

tableaux, figures, exemples et citations) ont été étiquetés, ainsi que le formatage initial des textes (récupéré dans Word) : italiques, gras, souligné, etc., et les éléments de langue étrangère, particulièrement représentés dans les articles.

Si certains éléments sont semi-automatiquement étiquetables (utilisation d'expressions régulières[5] dans Word et dans d'autres éditeurs de texte, de scripts Perl, etc.), la procédure est coûteuse, et n'a pu être appliquée à l'ensemble des textes du corpus : seul 1/5^{e} des textes environ a été balisé de la sorte.

En outre s'est posé le problème de l'intégration de l'annotation morphosyntaxique au balisage TEI, fusion problématique qui a demandé un développement important à S. Loiseau (fusion des annotations morphosyntaxique et XML-TEI, Loiseau, 2006).

Etant donné leur importance et leur impact sur les caractéristiques morphosyntaxiques des textes, seuls les exemples et les citations ont été balisés dans l'ensemble des textes ; les balises ont été posées manuellement sur les textes annotés (*v.* chapitre 5).

2.1.2. Corpus de comparaison

2.1.2.1. Corpus « Auteurs »

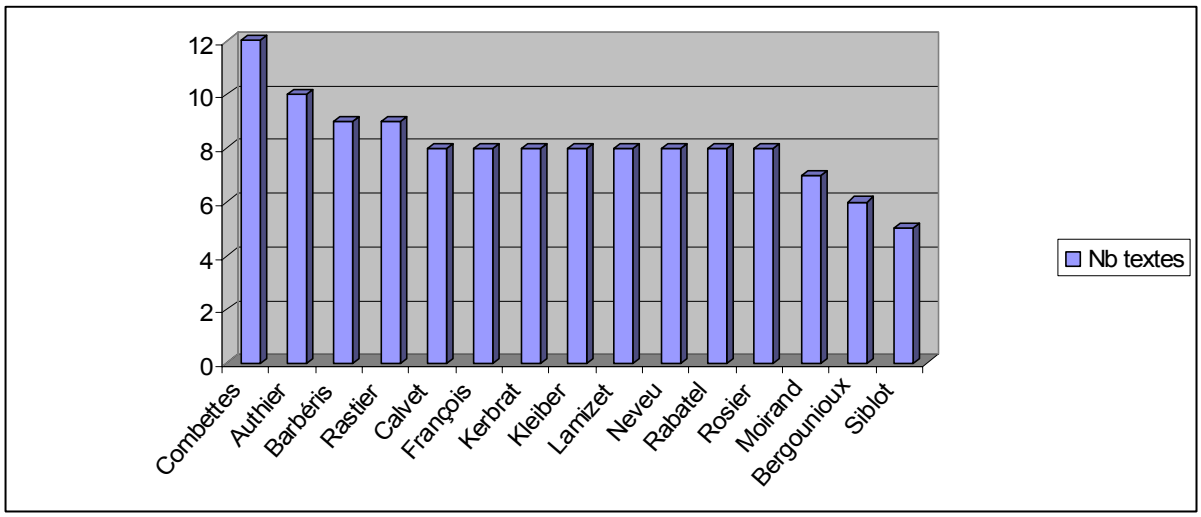
Comme il a déjà été évoqué, le corpus « Auteurs » a été construit dans un cadre collaboratif. Le corpus a été collecté à la suite de demandes à un ensemble de linguistes du champ ; Rinck et moi-même disposant déjà d'une collection importante d'articles de linguistique, ce sont les auteurs les plus représentés dans nos corpus respectifs qui ont d'abord été la cible de nos requêtes. Observer les styles d'auteur scientifique revient en effet à examiner les modalités de construction stylistique d'une communauté, et il nous semblait important dans cette perspective de réunir un ensemble de linguistes *accrédités* dans le champ – *v.* Rinck (2006) pour un contraste des textes avec des productions de doctorants en linguistes.

Ayant arrêté une date de clôture du corpus afin de lancer les analyses rapidement – et dans les cadres temporels de nos doctorats respectifs, ce sont les 15 auteurs ayant répondu à l'appel le plus promptement qui ont été pris en compte – notons que le nombre de textes dont nous disposions (toutes archives reçues comprises) était suffisant pour certains auteurs, comme Siblot par exemple.

Les quantités d'articles obtenues variant d'un auteur à l'autre[6], nous avons restreint les données à une moyenne d'environ 8 textes par auteur. Le corpus est donc constitué d'un ensemble de 122 textes :

[5] Les expressions régulières ou rationnelles (en anglais *regular expressions* dont l'abrégé est **regexp** ou **regex**, parfois traduites par **expressions régulières**) sont une famille de notations compactes et puissantes pour décrire certains ensembles de chaînes de caractères. Ces notations sont utilisées par plusieurs éditeurs de texte et utilitaires (particulièrement sous Unix), par exemple Vim, Emacs, sed et awk, pour parcourir de façon automatique des textes à la recherche de morceaux de texte ayant certaines formes, et éventuellement remplacer ces morceaux de texte par d'autres. (source : Wikipédia, http://fr.wikipedia.org/wiki/Expression_rationnelle)

[6] Rabatel nous a par exemple fait parvenir 19 textes, tandis que Moirand nous en a envoyé 7.



Graphique : Représentation des auteurs par nombre d'articles

Certains auteurs sont plus représentés que d'autres : on relève ainsi deux fois plus d'articles de Combettes que de textes de Siblot. Cela est en partie lié au fait que plusieurs des textes envoyés étaient déjà compris dans nos corpus respectifs (intersection de 18 articles entre « Auteurs » et ASLF). Les textes du corpus ASLF ont donc été adjoints aux 15 styles observés.

Les références des textes du corpus « Auteurs » sont livrées en annexe 2. On soulignera que nous ne disposons malheureusement pas des références exactes de chaque article, les auteurs ne nous les ayant pas toutes remises.

Dans la mesure où le corpus « Auteurs » est particulièrement exploré dans le chapitre 7 qui lui est dédié, nous n'approfondirons pas davantage sa description.

2.1.2.2. Corpus « Mécanique » et « Genres »

A. « Mécanique »

L'étude des variations du genre de l'article d'une discipline scientifique à l'autre nécessitant le choix d'une, ou de plusieurs disciplines, et l'application d'une méthodologie d'analyse similaire, c'est le domaine mécanique que nous avons décidé d'observer et de contraster ; le choix peut paraître singulier dans la mesure où les deux domaines scientifiques n'ont pas d'affinités frappantes : au contraire, ils s'opposent fortement (sciences dures / sciences humaines, discipline théorique / appliquée, etc.).

L'adoption du domaine mécanique s'est donc imposée de fait, et non de droit, dans le cadre d'une collaboration avec V. Clavier, qui disposait du corpus et d'experts pour le valider et l'interpréter. Dans la mesure où la mécanique, comme nombre de disciplines appliquées, est internationalisée, peu de revues contiennent des *articles*, qui sont d'abord publiés en anglais dans des revues internationales.

Les textes contrastés au domaine linguistique sont donc des *articles dans des actes*, et non des *articles*, *i.e.* des textes de genres *a priori* différents. Malgré les réserves que cette décision pourrait éventuellement impliquer, nous les considérerons comme équivalents, puisqu'ils ont finalement valeur d'articles dans la communauté mécanique française.

Cette mise en contraste exploratoire nous permettra de préciser certaines spécificités de la structure générique linguistique, ce qui intéresse hautement notre étude.

Le corpus « Mécanique » est constitué de 49 textes extraits du XV[e] Congrès Français de Mécanique, conduit sous l'égide du groupe « thématiques transverses » AUM de l'Association Française de Mécanique (AFM).

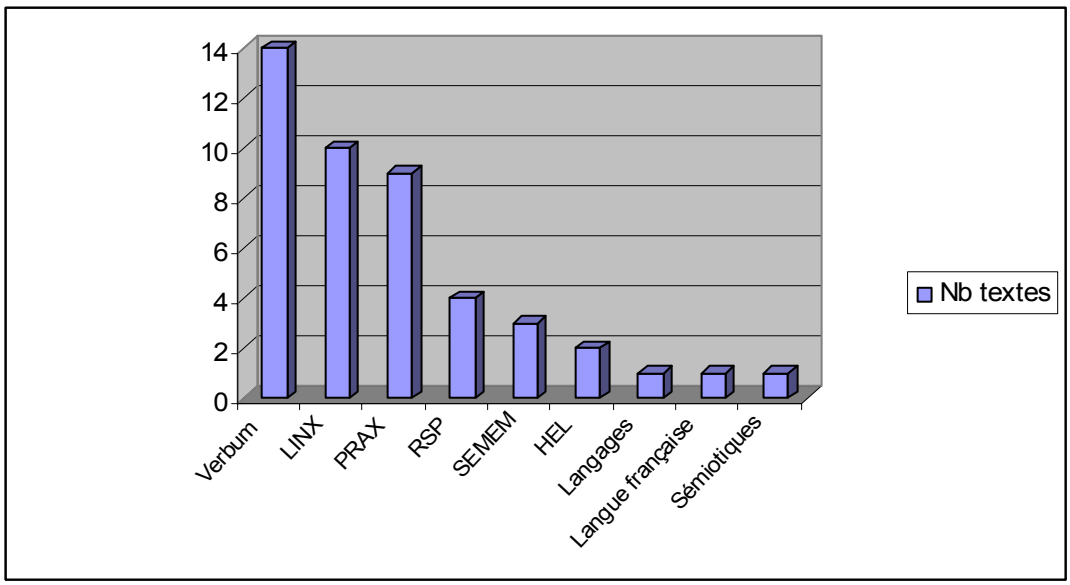
Sept domaines, sur les 26 représentés dans les actes du congrès, sont figurés dans le corpus :

- ✓ Acoustique (3 textes) ;
- ✓ Aérodynamique, Hydrodynamique (6 textes) ;
- ✓ Biomécanique (8 textes) ;
- ✓ Conception Production (8 textes) ;
- ✓ Dynamique des structures et des machines (7 textes) ;
- ✓ Ecoulements polyphasiques (8 textes) ;
- ✓ Endommagement – rupture – fatigue (4 textes) ;
- ✓ Environnement, Milieux poreux (1 texte).

B. « Genres »

Les revues françaises contenant un nombre plus restreint de genres scientifiques que les revues anglo-saxonnes : au final, deux genres, dont la présence est loin d'être systématique, ont pu être relevés dans l'ensemble des archives qui nous ont été confiées : la *présentation de revue* et le *compte rendu* (d'ouvrage ou de conférence[7]).

45 présentations de revues ont ainsi été extraites de 45 numéros de revues différents ; dans la mesure où toutes les revues ne sont pas précédées d'une présentation, les textes ont souvent été extraits de numéros, voire de revues non représentés dans ASLF comme *Semen* :

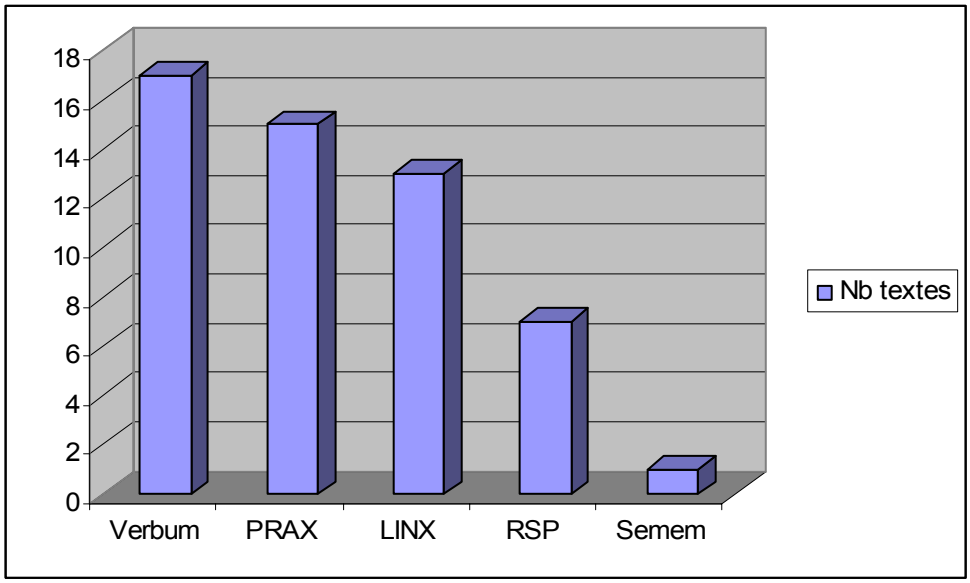


Graphique : Présentations de revue

[7] La très grande majorité des textes du corpus sont des comptes rendus d'*ouvrages*, on ne relève qu'un compte rendu de conférence.

La variable *Revue* est ainsi trop inégalement représentée pour être observable et corrélable à des caractéristiques linguistiques.

Il en va de même des comptes rendus, dont la présence et l'existence varie d'une revue à l'autre. Au final, nous avons extrait 53 comptes rendus de cinq revues : *LINX, Cahiers de praxématiques, RSP*, *Semen* et *Verbum* :



Graphique : Comptes rendus

Plus qu'à leurs déterminations éditoriales ou sociales, c'est aux structures génériques des deux genres et à leurs lieux de contraste ou de similarité avec l'article que l'on s'intéressera.

2.1.2.3. Corpus ASLA

La constitution du corpus ASLA suit la même procédure que celle du corpus ASLF : un ensemble de revues linguistiques anglo-saxonnes a d'abord été sélectionné et validé par plusieurs experts du champ. L'opération s'est avérée plus délicate, dans la mesure où les communautés linguistiques et anglo-saxonnes interfèrent relativement peu ; de surcroît, la communauté française est beaucoup plus restreinte. Le monde linguistique anglo-saxon englobe en effet de très nombreuses communautés nationales, drainant de nombreux auteurs non natifs. De nombreuses revues anglo-saxonnes sont ainsi publiées aux Pays-bas et il va de soi qu'une revue dite 'internationale' ne renvoie pas à la même réalité selon qu'elle est publiée en France, en Grande-Bretagne ou aux Pays-Bas : les articles d'une revue internationale française sont généralement rédigés en français, mais comprennent un résumé en langue anglaise (ce qui est peu commun dans les revues nationales françaises) et admettent plusieurs articles rédigés en anglais par des non natifs[8]. Une revue internationale originaire des Pays-Bas imposera l'usage de l'anglais, à l'instar des britanniques ou des américaines, ce qui concourt à l'élargissement de la communauté anglophone.

De manière générale, les pratiques rédactionnelles et de sélection des articles sont bien distinctes d'une communauté à l'autre ; si le corpus français est majoritairement constitué de numéros de revues élaborés par cooptation, cette procédure est moins courante dans les

[8] Le nombre d'articles rédigés dans d'autres langues que le français est d'ailleurs généralement limité.

périodiques anglo-saxons, qui recourent davantage à des modalités de sélection anonyme : la communauté est effectivement plus étendue et les soumissions spontanées ou par appel plus nombreuses qu'en France, où les linguistes sont rapidement amenés à se connaître.

Ainsi, *Applied Linguistics* insiste lourdement sur l'anonymat de la sélection ; à l'instar du *Journal of Linguistics* et de *Language*, la revue exige l'envoi de manuscrits anonymes. Le nom de l'auteur doit être inscrit sur un feuillet détachable, et ne doit pas apparaître dans l'article (les revues exigent d'y substituer la mention 'author'). *Applied Linguistics* et *Journal of Linguistics* demandent en outre, et dans la mesure du possible, de ne pas pouvoir l'identifier à partir des remerciements et des références du texte. Ces exigences sont susceptibles d'influencer le discours, et leur impact devra être étudié. Les trois revues sont nettement anglo-saxonnes, dans la mesure où elles sont affiliées à plusieurs sociétés linguistiques américaines, anglaises, voire internationales.

Un parcours des consignes de rédaction exigées par un panel de revues françaises et anglo-saxonnes (indépendamment des revues comprises dans le corpus) fait émerger des différences frappantes :

- ✓ Les consignes imposées dans les revues françaises sont d'ordre essentiellement technique et régulent le format des textes (notes de bas de page, bibliographie, titres, etc.), mais la plupart des revues ne se permettent pas de formuler des consignes stylistiques. Au mieux note-t-on certaines exigences relatives à la structuration des textes : le *Bulletin de la Société Linguistique de Paris* insiste par exemple sur la *lisibilité* des articles, dans la mesure où « tout article de plus de six pages doit être divisé par des intertitres ou, à défaut, par une numérotation » ;
- ✓ En revanche, les revues anglo-saxonnes prêtent une attention moindre au formatage de l'article mais soulignent qu'en contrepartie, les auteurs devront veiller au style employé. Le style *reader-friendly* est ainsi exigé dans *Language* et *Journal of Linguistics*, *Applied Linguistics* recommandant aux auteurs de se conformer au style de la revue. Certaines consignes n'apparaissent pas dans les revues françaises : elles ont trait au sexisme (les auteurs soumettant un article au *Journal of Pragmatics* sont ainsi renvoyés au *Guidelines for Nonsexist Use of Language*) ou au copyright ; certaines revues rappellent en outre qu'il est interdit de soumettre un article ayant déjà été publié et se réservent un droit d'exclusivité de l'article, ce qui est peu répandu en France.

Le respect des consignes imposées par la revue joue un rôle important quant à l'homogénéité de la revue et la régulation de la pratique rédactionnelle. La *Revue de linguistique romane* mentionne ainsi peu de consignes, ce qui nous semble contribuer à l'hétérogénéité prononcée des articles.

Ce rapide examen des consignes imposées par les revues françaises et anglo-saxonnes illustre la complexité de la sélection ; dans tous les cas, il n'est pas envisageable de constituer un corpus de revues *équivalentes* dans les deux langues (alors qu'une étude contrastive interlangue de journaux ne poserait pas les mêmes difficultés). En outre, l'observation des consignes laisse présager d'une plus grande homogénéité du style rédactionnel anglo-saxon – et d'une plus grande normalisation linguistique du genre.

Si c'est un corpus de 32 numéros de revues, soit 7 revues et 190 textes qui a été constitué, nous n'en présenterons que les 103 articles qui ont pu être exploités dans le cadre de cette thèse.

Quatre revues sont représentées dans le corpus ASLA :

- ✓ *Journal of Pragmatics* (mensuel, Elsevier), spécialisé en pragmatique, au sens anglo-saxon du terme.

  Web : http://www.elsevier.com/wps/find/journaldescription.cws_home/505593/description#description

- ✓ *English for Specific Purpose* (trois livraisons par volume annuel, Elsevier), spécialisé en didactique de l'anglais de spécialité.

  Web : http://www.elsevier.com/wps/find/journaldescription.cws_home/682/description#description

- ✓ *Linguistics* (six par volume annuel, Mouton de Gruyter), revue interdisciplinaire et non spécialisée de sciences du langage.

  Web : http://www.degruyter.de/rs/6468_407_ENU_h.htm

- ✓ *Computers and the Humanities* (quatre par volume annuel, Springer)

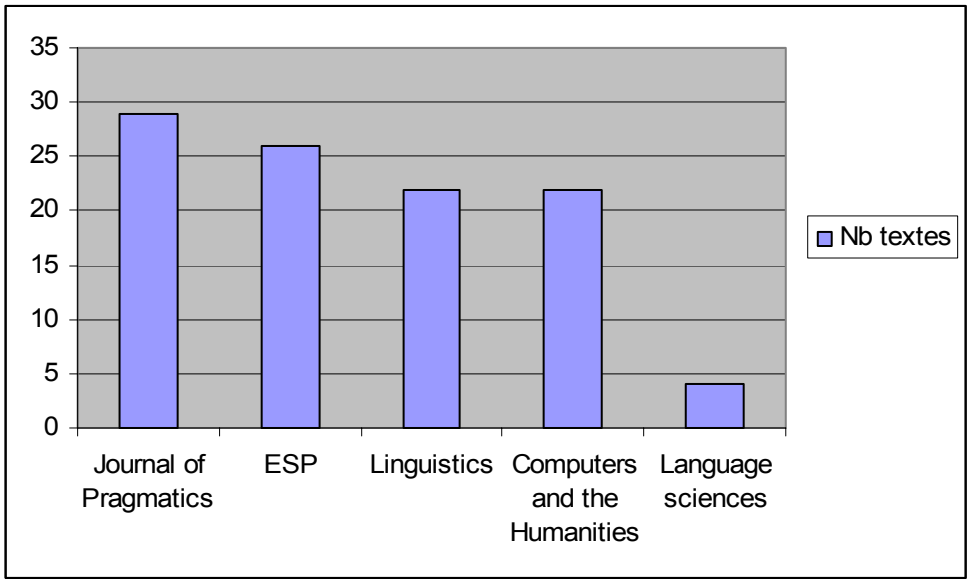
  Web : http://springerlink.metapress.com/(koghhe3uhiueiezxuguqunur)/app/home/journal.asp?referrer=parent&backto=linkingpublicationresults,1:100251,1

- ✓ *Language Sciences* (six par volume annuel, Elsevier), revue interdisciplinaire et non spécialisée de sciences du langage.

  Web : http://www.elsevier.com/wps/find/journaldescription.cws_home/867/description#description

Linguistics et *Language* sont des revues généralistes tandis qu'*ESP, Computers and the Humanities* et *Journal of pragmatics* sont spécialisées : il sera intéressant d'observer leur position sur les axes d'organisation générique du genre anglo-saxon, et le caractère distinct de leurs objets d'étude privilégiés respectifs.

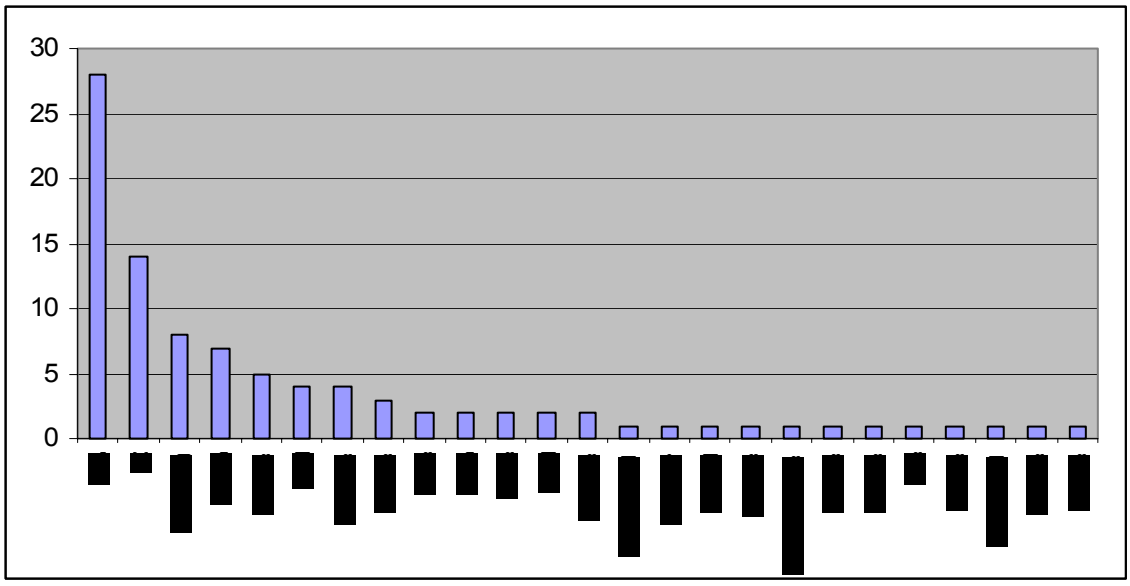
Les quatre premières revues de la liste sont également représentées, tandis que nous n'avons inclus que 4 articles de *Language sciences*, ce qui interdira la prise en considération de la revue :



Graphique : Représentation des revues dans le corpus ASLA

Afin de faciliter la comparaison d'ASLF et ASLA, la variation stylistique individuelle est également neutralisée dans le corpus anglo-saxon : 121 chercheurs sont représentés dans 103 textes, et seul un auteur, D. Eddington apparaît à deux reprises dans la collection.

La communauté linguistique anglo-saxonne étant internationale et géographiquement plus étendue que son homologue français, c'est par pays plutôt que par région ou département qu'a été observée l'origine géographique et institutionnelle des textes du corpus – précisons que seule la provenance de 8 auteurs n'a pu être déterminée[9] :

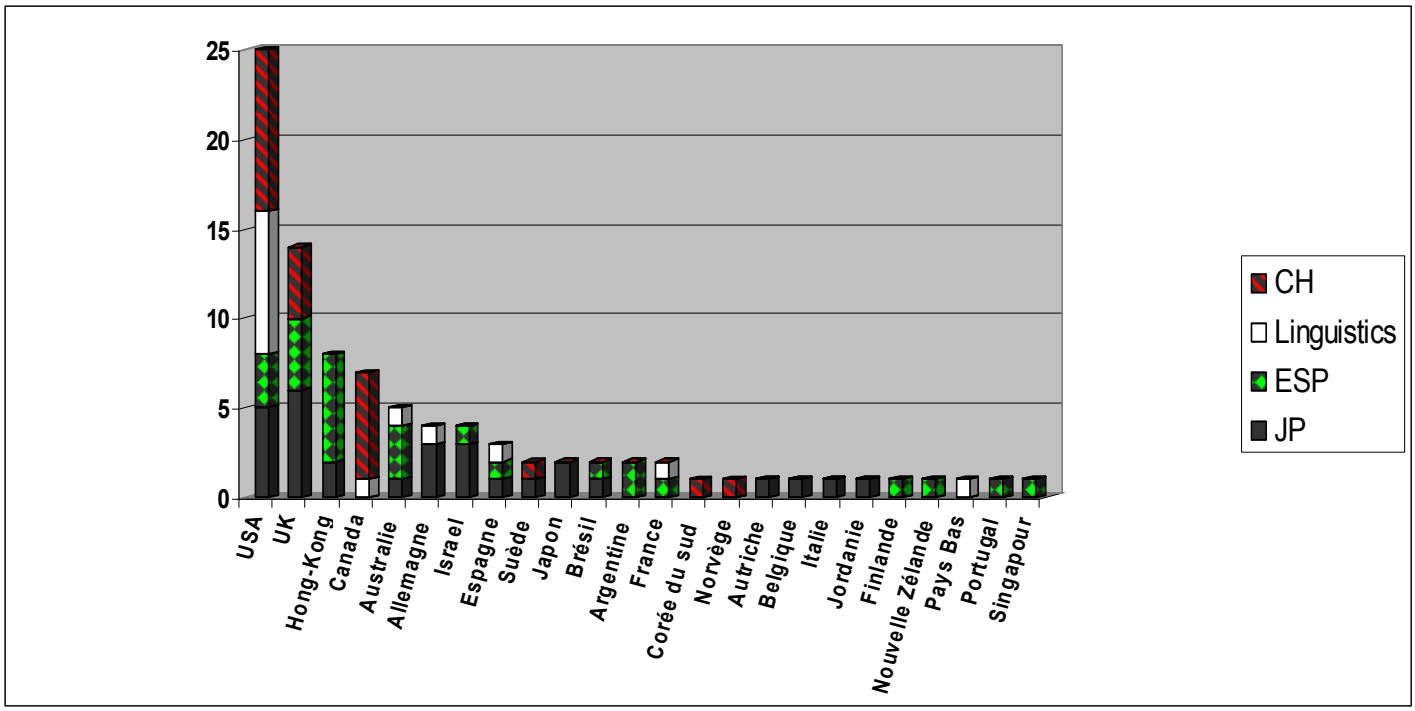


Graphique : Pays représentés dans ASLA (nombre d'articles par pays)

[9] Chiffre nettement moins élevé que pour le corpus français : les affiliations institutionnelles sont en effet plus systématiquement présentes dans les articles anglo-saxons que dans les textes français. Les 8 auteurs « indécidables » sont issus de la revue généraliste *Linguistics*, d'ailleurs internationale anglophone plutôt qu'anglo-saxonne (Editeur de Gruyter, Berlin). Nous avons pu récupérer certaines affiliations par le biais du Web, la plupart des chercheurs anglo-saxons assurant leur visibilité par le biais de pages web, pratique encore peu répandue en France.

Comme le montre le graphique 10, les Etats-Unis et la Grande-Bretagne demeurent les pays les plus représentés, suivis de pays (en partie ou en totalité) anglophones comme Hong-Kong, le Canada ou l'Australie.

Précisons cette remarque en observant les représentations par pays des revues d'ASLA :



Graphique : Pays représentés dans ASLA

Le graphique est éloquent : les pays sont inégalement représentés. Hong-Kong et l'Australie sont essentiellement corrélées à *ESP* (rappelons que le courant est né en Australie), au contraire du Canada, essentiellement représenté dans *Computers and the Humanities*.

Les communautés nationales semblent ainsi avoir développé des pôles de spécialisation qui encouragent leur sur- (sous-) représentation dans certaines revues. Ces résultats seraient certes à affiner et à valider dans des corpus plus importants, mais ils esquissent déjà certaines frontières scientifiques.

Les caractéristiques linguistiques du corpus seront explorées au sein du chapitre 9, dédié au contraste d'ASLF et ASLA, et plus généralement, à l'analyse contrastive des genres.

2.2. Annotation

2.2.1. Evaluation et exploitation des outils d'annotation morphosyntaxique disponibles[10]

2.2.1.1. De l'inadéquation des outils d'a »nnotation et des systèmes d'étiquetage disponibles

Etant donné la taille du corpus, un étiquetage manuel des textes ne saurait être envisagé. Les étiqueteurs morphosyntaxiques ou encore assignateurs automatiques de catégories (*taggers*) ont été largement développés ces dernières années, et il est communément admis que les résultats qu'ils obtiennent sont tout à fait honorables.

Aussi avions-nous dans un premier temps adopté le logiciel Cordial, de la société Synapse développement. Outre sa qualité d'annotation, Cordial propose un grand nombre de variables (plus de 200) qui ont prouvé leur efficacité dans diverses entreprises de validation de typologies textuelles et d'études en corpus (e.g. Malrieu et Rastier 2001, Beauvisage 2001). Malgré ses qualités et à la suite d'analyses menées à l'aide du logiciel (Poudat 2003, Poudat et Loiseau 2005), Cordial s'est finalement avéré peu approprié aux spécificités de notre étude, dans la mesure où les variables proposées n'étaient globalement pas adaptées à la description du discours scientifique, et a fortiori à celle de l'article de revue linguistique : l'étiqueteur propose ainsi de nombreuses catégories sémantiques ambiguës et peu transparentes (*e.g.* noms 'abstraits' / 'concrets', 'humanoïdes', etc.), et considère les paragraphes précédés de tirets comme indices de 'dialogue', ce qui illustre l'orientation du logiciel vers la description des textes littéraires.

Outre les diverses erreurs d'étiquetage récurrentes relevées (confusions imparfait / conditionnel, étiquetage problématique des noms propres[11], etc.), deux types d'erreurs spécifiquement liées aux caractéristiques de nos textes ont pu être distinguées :

(i) celles entraînées par une segmentation[12] ou une identification incorrecte de mots ou d'unités, liés à certaines caractéristiques du *discours scientifique en général*, comme par exemple :

- ✓ Les *marques de formalisation* (*e.g.* symboles ou indices formels) de type 'X' ou 'SNprep', mal découpés et étiquetés comme noms propres ou comme abréviations dans le meilleur des cas ;
- ✓ Les *marqueurs de structuration* (*e.g.* titre, liste) comme '1.1.' traités comme de simples cardinaux suivis de points ;
- ✓ De nombreuses *séquences en langue étrangère* (citations, exemples, etc.) entraînant un bruit considérable.

[10] Cette entreprise a été menée dans le cadre d'un contrat de bourse Marie Curie à *Bergen Advanced Training Site in Multilingual Tools* (BATMULT) – Laboratoire *AKSIS* (Department of Culture, Language, and Information Technology) – Octobre-décembre 2003.

[11] Les auteurs de linguistique sont naturellement peu connus du logiciel, qui semble avoir pourtant assimilé un nombre important de noms d'hommes politiques ou d'auteurs littéraires…

[12] Ou *tokenisation*.

(ii) celles entraînées par un découpage ou une identification incorrecte de mots ou d'unités, liés à certaines caractéristiques du *discours scientifique linguistique* : les *marqueurs d'acceptabilité linguistique* de type '?' ou '!' sont systématiquement étiquetées comme ponctuations, ce qui augmente substantiellement le nombre de points d'interrogation et d'exclamation de certains textes.

On observera également que de nombreux phénomènes qui intéresseraient par contre la description des textes scientifiques sont absents du système d'étiquetage proposé : les pronoms *il impersonnel* et *on* ne sont par exemple pas distingués, pas plus que les passifs ou les modaux.

Soulignons bien entendu que ces problèmes ne sont aucunement l'apanage de Cordial, et qu'il n'existe à notre connaissance aucun outil d'étiquetage automatique spécifiquement adapté à l'observation du discours scientifique ; *robustesse*, *efficacité*, *précision*, *adaptabilité* et *ré-emploi* sont les qualités que doit présenter un étiqueteur (Cutting et al., 1992) et c'est le plus souvent sur de tels critères que les outils sont évalués et développés. Les taggers sont donc développés dans une perspective généraliste : ils sont supposés étiqueter « en langue » et doivent par conséquent pouvoir s'appliquer à tout texte, bien que (Biber, 1993) et (Illouz, 1999) aient montré que la qualité des sorties variait substantiellement d'un (type de) texte et d'un outil à l'autre.

Ces options généralistes ont de sérieuses conséquences, tant sur le jeu d'étiquettes développé, qui doit être robuste[13], que sur le corpus et les hypothèses d'entraînement de l'outil, qui doivent être généralistes.

Si ces outils généralistes présentent un intérêt applicatif indéniable en termes de classification et de validation de typologies textuelles (e.g. Kessler, Karlgren & Cutting), leur potentiel descriptif reste faible ; l'étude d'un discours, d'un genre ou d'un corpus requiert en amont un ensemble d'hypothèses descriptives qui ne sauraient être contenues dans un jeu de variables d'une cinquantaine d'étiquettes trop générales, a fortiori lorsque la démarche adoptée est exploratoire et porte sur un corpus relativement homogène. Dans ce cas de figure, l'adoption d'un tel outil requiert l'étiquetage ultérieur – et (semi-)manuel – de variables additionnelles. En effet, la caractérisation d'un corpus ne peut être que différentielle : dans le cas d'un corpus hétérogène, un ensemble de variables générales peut constituer un noyau différentiel permettant de contraster les sous-corpus entre eux et de caractériser certaines de leurs différences. En revanche, un corpus aussi homogène que le nôtre – tant discursivement, que génériquement ou domanialement – limite le pouvoir discriminant de variables trop générales et nécessite un ensemble de variables plus fines, plus adaptées aux caractéristiques des textes et plus interprétables.

Dans cette perspective, nous avons envisagé un temps d'évaluer les outils en l'état et la pertinence des variables morphosyntaxiques disponibles relativement à nos objectifs[14]. L'entreprise s'est toutefois avérée trop coûteuse en termes de temps et de moyens. En effet, bien qu'on ne vise à évaluer les étiqueteurs que relativement à une tâche donnée, les entreprises d'évaluation (e.g. Grace, Eagles) demandent des investissements importants (élaboration d'un système de mesure efficace et pertinent, constitution d'un corpus de

[13] C'est-à-dire offrir un nombre restreint d'étiquettes renvoyant aux grandes parties du discours et aux traits linguistiques les plus identifiables formellement (genre, nombre le plus souvent). Ainsi, le système d'annotation du projet Penn Treebank, le plus utilisé à l'heure actuelle, ne comporte que 36 étiquettes, cf. annexe 7.

[14] *V.* annexe 7, qui reprend (Poudat, 2004) sur Texto !

http://www.revue-texto.net/Corpus/Publications/Poudat_Taggers.html

référence, développement d'un système d'étiquetage de référence fédérant les caractéristiques des systèmes évalués, travail quantitatif de comptage des erreurs, etc.[15]) pour un résultat finalement mitigé, dans la mesure où nous serions tenue d'adopter l'étiqueteur ayant obtenu les meilleures performances quelles qu'elles soient. Il pourrait éventuellement être envisageable d'évaluer la qualité des variables et de ne retenir que celles ayant obtenu les meilleurs scores ; l'exploitation de plusieurs sorties d'étiqueteurs est cependant problématique et nécessiterait un développement important[16].

La solution d'évaluation a donc été abandonnée, d'autant que les outils existants ne réalisent pas leurs meilleures performances sur les textes qui nous intéressent, dans la mesure où ils ont été entraînés et développés sur d'autres types de corpus : en effet, les problèmes de disponibilité des corpus et le coût élevé d'une entreprise d'annotation manuelle contraignent les concepteurs à entraîner leurs étiqueteurs sur les corpus annotés existants, et le plus souvent sur le corpus Penn Treebank pour l'anglais. Bien que particulièrement intéressant en termes de taille (4.5 millions de mots), le Penn Treebank est composé d'un ensemble hétérogène de textes (Dept. of Energy abstract, Dow Jones Newswire stories, MUC-3 messages, WBUR radio transcripts, Brown corpus, etc.) dont la représentativité est discutable : ainsi, le corpus Brown est composé de textes datant de 1961. On se heurte à des problèmes similaires en français, dans la mesure où les outils sont bien souvent entraînés sur des corpus textuels hétéroclites dans lesquels des romans du XIX^{e} côtoient des articles du journal *Le Monde* du XXIe siècle…

C'est d'ailleurs pour cette raison que des variables non ambiguës dans notre corpus ne peuvent être étiquetées automatiquement, dans la mesure où elles seraient ambivalentes dans d'autres genres et types de discours : ainsi, l'on sait que dans le discours scientifique, la phraséologie « il semble » indique ordinairement la présence d'un impersonnel. Or, dans d'autres genres comme le roman par exemple, le « il » renverra davantage à un anaphorique, ce qui rend l'énoncé ambigu. On peut donc légitimement supposer qu'un étiqueteur entraîné sur des corpus scientifiques, et a fortiori, sur un sous-ensemble de notre corpus, obtiendrait des résultats bien supérieurs en sortie.

En raison de l'inadéquation des systèmes d'étiquetage et des performances somme toute passables que les outils disponibles en l'état obtiennent sur notre corpus, nous avons décidé d'exploiter les possibilités d'entraînement qu'offraient certains étiqueteurs. En effet, plusieurs des outils recensés se sont avérés « entraînables », c'est-à-dire qu'ils permettent de générer un outil d'annotation automatique à partir d'un corpus d'articles manuellement étiqueté. Le système d'annotation et la langue de départ sont libres, ce qui convient particulièrement bien à notre étude.

2.2.1.2. La solution d'entraînement

A. Des étiqueteurs entraînables

Parmi les étiqueteurs recensés (*v.* annexe 8), quatre systèmes, fondés sur des hypothèses linguistiques et computationnelles distinctes, offrent une possibilité d'entraînement :

[15] (Santos & Gasperin 2002) et (Padro & Marquez 2002) ont d'ailleurs souligné les problèmes que posent les entreprises d'évaluation et remis en question leur efficacité sur certains points.

[16] Que l'on pourrait rapprocher aux travaux menés dans le cadre du projet AMALGAM.

1. Brill tagger

Développé par Eric Brill en 1993 et fondé sur les travaux de Bloomfield et Harris, Brill infère des règles d'étiquetage à partir d'un corpus annoté puis procède à une analyse distributionnelle pour réduire les erreurs d'étiquetage.

Les mots sont étiquetés en *deux étapes*:

- un ensemble de règles est inféré d'un corpus d'entraînement pour prédire l'étiquette la plus probable d'un mot (par exemple, un mot se terminant par –ed est très susceptible d'être un verbe au *simple past*). Soulignons que toutes les occurrences d'un mot dans le corpus recevront le même tag ;
- l'étiquetage est affiné à l'aide de règles contextuelles (*ex.* : changer l'étiquette d'un mot taggé 'verbe' en 'nom' si le mot précédent est étiqueté 'déterminant').

Le tagger ne segmente pas le texte en *tokens* : tous les textes utilisés dans la phase d'entraînement doivent donc être pré-segmentés (une phrase par ligne). Les marques de ponctuation doivent être séparées des mots par une espace.

Le *corpus d'entraînement* (étiqueté manuellement) doit être divisé en deux corpus (un programme fourni dans le package permet de le répartir au hasard, phrase par phrase) : le premier sera utilisé pour générer les règles d'étiquetage des mots inconnus, le second les règles contextuelles.

La phase d'apprentissage est relativement longue, dans la mesure où le programme est écrit en Perl, langage interprété beaucoup plus lent qu'un langage compilé comme C par exemple. Il faut environ *trois jours* pour générer les premières règles à partir d'un corpus de 250 000 mots.

La procédure d'entraînement du corpus est très bien détaillée dans le fichier fourni par Eric Brill.

2.. MBT tagger

Le tagger MBT (ILK, W. Daelemans) est fondé sur des techniques d'apprentissage et de classification, et concrètement sur le logiciel de classification linguistique TiMBL (également distribué gratuitement par les mêmes concepteurs).

Les méthodes d'apprentissage basées sur la mémoire (*memory-based learning*) fonctionnent sur l'idée qu'un raisonnement par *analogie* est tout aussi efficace, voire plus approprié qu'un raisonnement fondé sur des *règles* (comme Brill). Dans cette perspective est stocké un ensemble de représentations de situations antérieures (les exemples) destiné au traitement d'éléments nouveaux à partir d'une mesure de similarité.

Générer un tagger s'effectue à l'aide d'un fichier organisé en deux colonnes séparées par une espace : dans la première sont consignés les mots ou les marques de ponctuation, et les étiquettes dans la seconde.

La procédure d'entraînement semble relativement aisée : MBT génère deux bases à partir du texte d'entrée, une base pour les mots connus et une autre destinée au traitement des mots inconnus. Soulignons que l'utilisateur doit fournir des informations concernant le contexte (contextes droit et gauche, portée) et la forme des mots à étiqueter (position, accents, caractères numériques, etc.) : un certain nombre de tests doit donc être effectué afin de générer le tagger le plus pertinent.

3. TreeTagger

TreeTagger (TC Project, H. Schmid) se rapproche des taggers *n-gram* traditionnels mais utilise un arbre de décision binaire pour calculer la taille du contexte à utiliser afin d'estimer les probabilités de transition.

Il fonctionne à partir de deux programmes : *train-tree-tagger*, qui génère un fichier paramètre à partir d'un lexique et d'un corpus manuellement balisé et *tree-tagger*, qui prend un fichier paramètre et un fichier texte en arguments et qui permet d'étiqueter les textes.

L'entraînement de TreeTagger s'effectue avec la commande suivante :

```
train-tree-tagger <lexicon> <open class file> <input file> <output file>
```

Quatre fichiers sont donc nécessaires :

- **<lexicon>** est un fichier contenant un lexique pleine forme. On retrouve un format similaire à celui requis par les autres étiqueteurs : une forme par ligne, chaque occurrence d'un mot étant suivie par une tabulation et un ensemble de paires tag-lemme, elles-mêmes séparées par des espaces.

Exemple:

aback RB aback

abacuses NNS abacus

abandon VB abandon VBP abandon

abandoned JJ abandoned VBD abandon VBN abandon

- **<open class file>** est un fichier contenant une liste d'étiquettes susceptibles d'être affectés aux mots inconnus (mots non inclus dans le lexique).

Exemple (pour le système d'annotation du projet Penn Treebank):

FW JJ JJR JJS NN NNS NP…

- **<input file>** est un fichier contenant des données balisées (un mot par ligne). Chaque ligne contient un token et un tag, séparés par une tabulation. Les marques de ponctuation sont considérées comme des tokens et doivent être étiquetées comme tels.
- **<output file>** est le fichier dans lequel les paramètres résultants du tagger sont enregistrés.

D'autres paramètres – optionnels – sont également proposés.

4. TnT tagger

TnT (Trigrams'n'Tags) tagger est un étiqueteur statistique constitué d'un ensemble de méthodes de lissage (*smoothing*)[17] et de traitement des mots inconnus, implémenté sur un algorithme fondé sur les modèles de Markov (Brants, 2000).

Son entraînement requiert un format de fichier identique à celui de MBT : un token par ligne et deux colonnes, l'une contenant les mots, l'autre les tags.

[17] Ajustement de l'ensemble des données au modèle, ou à la courbe.

TnT est facilement entraînable, il suffit d'enregistrer un corpus étiqueté en format .TT dans le répertoire MODELS et de l'appeler en paramètres de l'étiquetage *via* la commande *tnt-para*.

5. Synthèse

Le tableau qui suit synthétise les caractéristiques des quatre procédures d'entraînement :

Tagger	Brill Tagger	MBT Tagger	TnT Tagger	TreeTagger
Entrée	Corpus manuellement annoté	Corpus manuellement annoté	Corpus manuellement annoté	Corpus manuellement annoté Lexique
Format	Pré-tokenisation, une phrase par ligne, format Mot[/]Tag	Format Mot[SP]Tag, un token par ligne	Format Mot[SP]Tag, un token par ligne	*Corpus* : Mot[tab]Tag, un token par ligne *Lexique* : Mot[tab] paire(s) tag-lemme, un token par ligne
Réserves	Implémentation du module en Perl = lenteur de la procédure	Paramétrage compliqué du contexte et de la forme des mots		Lexique requis en entrée de la procédure

Tableau : Synthèse des procédures d'entraînement

B. Procédure

La procédure d'entraînement est incrémentale : on étiquette manuellement un corpus d'entraînement de taille restreinte qui permet de générer un premier étiqueteur ; ce dernier permet d'annoter un nouveau corpus qui sera manuellement corrigé et qui sera fusionné avec le premier corpus d'entraînement. On poursuit le processus jusqu'à ce que l'on obtienne un étiqueteur satisfaisant ; dans le cas présent, l'opération a été reconduite jusqu'à annotation complète des textes du corpus ASLF et des corpus de comparaison « Styles », « Genres » et « Mécanique » ; il serait bien entendu intéressant d'évaluer les performances de l'étiqueteur généré de manière objective (taux de précision et de rappel).

1. Corpus d'entraînement : correction (semi-) manuelle en contexte

Le jeu d'étiquettes établi (présenté *infra*), un sous-ensemble de vingt articles français, soit 136 936 mots, a été sélectionné pour tenir lieu de corpus d'entraînement. Les textes sont issus de trois revues linguistiques également représentées : les *Cahiers du CIEL* (Centre Interlangue d'Etudes en Lexicologie), les *Cahiers de Praxématique* et la *Revue de Sémantique et Pragmatique*.

L'annotation manuelle de 136 936 mots étant difficilement envisageable, nous avons opté pour une correction manuelle en contexte des résultats obtenus par un étiqueteur[18].

[18] L'utilisation de plusieurs sorties d'étiqueteurs poserait en effet de nombreux problèmes et constituerait un travail de recherche à part entière (cf. Projet AMALGAM).

C'est finalement la version française de TreeTagger qui a été retenue, les autres outils ayant été écartés pour différentes raisons : MBT et TnT ne permettent pas l'étiquetage du français et la version de Brill développée par l'Inalf n'est pas pourvue d'un découpeur (*tokenizer*). Cordial fournit en sortie un fichier complexe, et difficile à exploiter en l'état, tandis que les étiquettes qu'il propose sont globalement éloignées du système d'annotation choisi.

Au contraire, les étiquettes de TreeTagger sont plus proches du résultat souhaité : c'est donc à partir des sorties de TreeTagger que nous avons constitué le corpus d'entraînement. La plupart des éléments à modifier n'étaient pas automatisables, ou nécessitaient un développement plus important que leur annotation manuelle. La correction de différents phénomènes a toutefois pu être semi-automatisée à l'aide d'expressions régulières.

La procédure d'annotation s'est au final avérée particulièrement coûteuse en temps – pour information, l'annotation du corpus de référence de l'évaluation Grace (140 000 mots) a nécessité 200 heures.

2. Choix de TnT tagger

Une fois le corpus d'entraînement constitué, nous avons dû choisir un étiqueteur parmi les quatre outils potentiels. TreeTagger a été rapidement exclu en raison du lexique requis en entrée de l'entraînement, que nous n'avons jamais été en mesure de récupérer.

Nous avons donc entraîné Brill, MBT et TnT et c'est ce dernier que nous avons finalement retenu (un texte étiqueté par les 3 outils après entraînement est proposé en annexe 7) : MBT est particulièrement complexe d'utilisation et requiert un paramétrage important[19]. L'étiquetage d'un texte après entraînement et sans paramétrage nous a fourni une sortie décevante, et majoritairement constituée de mots inconnus.

La sortie de Brill était globalement satisfaisante, mais les temps d'entraînement du tagger nous ont semblé très problématiques ; il a ainsi fallu une nuit pour entraîner l'outil sur un corpus limité de 20 textes, sur une machine aux performances très honorables, ce qui nous a inquiétée pour la suite (entraînement sur des collections de plus en plus importantes).

TnT s'est avéré particulièrement souple d'utilisation, et très rapide à entraîner (une minute, voire quelques secondes suffisent pour un corpus constitué de plusieurs centaines de textes). L'étiqueteur est en outre un modèle génératif qui obtient de meilleurs résultats sur des corpus d'entraînement de taille restreinte comme le nôtre que les modèles discriminants (Clark et al., 2003) et qui a obtenu les meilleurs taux de précision dans diverses études (e.g. Zavrel et Daelemans 2000, Sjobergh, 2003).

3. Incrémentation : traitement de l'ensemble du corpus

Le corpus ASLF a été progressivement étiqueté avec TnT ; entraîné au départ sur une vingtaine de textes, l'étiqueteur généré a permis d'annoter un autre sous-corpus, qui une fois corrigé a été adjoint au premier corpus d'entraînement, ce qui a généré un nouvel étiqueteur, etc. Nous avons répété la procédure jusqu'à étiquetage de la totalité du corpus.

On précisera que si les corpus « Mécanique » et « Genres » ont bien été annotés, les textes n'ont pas été adjoints au corpus d'entraînement en raison de leur caractère génériquement ou

[19] Bien que ses concepteurs nous aient proposé de la paramétrer, nous avons finalement pris la décision de l'exclure, d'une part parce que nous aurions été amené à le reparamétrer, la procédure d'entraînement étant incrémentale, et d'autre part parce que TnT nous donnait entièrement satisfaction.

domanialement distinct. L'étiqueteur final généré est donc entraîné sur un total de 366 articles scientifiques de linguistique.

Certaines catégories ont été systématiquement vérifiées et corrigées manuellement : TnT étant un étiqueteur 3gram qui ne considère donc qu'un contexte très limité de 3 tokens, les étiquettes qui nécessiteraient la prise en compte d'une fenêtre plus large pour être correctement identifiées nécessitent vérification. Il en va ainsi de la distinction entre *il* impersonnel et anaphorique : les séquences de type *il + est + ADJ* ou *il + se + trouve* entravent par exemple la désambiguïsation de *il*.

2.2.2. Du jeu d'étiquettes retenu

Notre système d'annotation doit remplir deux critères : il doit être adapté aux caractéristiques du discours scientifique et suffisamment robuste pour être pris en charge par les outils. Cette dernière restriction bannit l'utilisation de catégories trop fines qui ne pourraient être correctement étiquetées, mais autorise la prise en compte de variables sémantiques comportant des régularités formelles au niveau du *token*. En effet, un étiqueteur est en mesure d'annoter tout type de variable (e.g. morphologique, phonologique, morphosyntaxique) à condition que celle-ci soit associable à une(des) forme(s) de *token*(s) particulière(s).

De manière générale, nous avons cherché à construire un système linguistiquement cohérent adapté à la description du genre de l'article de linguistique.

En ce sens, nos hypothèses sont plus discursives que grammaticales : certaines des variables adoptées dépassent donc le niveau morphosyntaxique et relèvent davantage du plan sémantique.

Le système d'annotation fédère deux types de variables : un ensemble de catégories morphosyntaxiques « de langue », incluant les grandes parties du discours et leurs attributs traditionnels (nombre, temps et modes verbaux, etc.) et un ensemble de variables supposées caractéristiques du genre de l'article de revue linguistique, comprenant des spécificités du discours scientifique et du domaine scientifique linguistique.

2.2.2.1. Parcours des systèmes d'annotation existants

Afin d'avoir une idée des types de catégories morphosyntaxiques qu'il serait pertinent de sélectionner, nous avons d'une part examiné les recommandations d'Eagles (1996) pour l'annotation morphosyntaxique des corpus, et d'autre part documenté les systèmes d'annotations utilisés par les étiqueteurs recensés (*v.* annexe 9).

Eagles distingue 13 catégories morphosyntaxiques distinctes :

1.	N[noun]	2.	V[verb]	3.	AJ[adjective]
4.	PD[pronoun/determiner]	5.	AT[article]	6.	AV[adverb]
7.	AP[adposition]	8.	C[conjunction]	9.	NU[numeral]
10.	I[interjection]	11.	U[unique/unassigned]	12.	R[residual]
13.	PU[punctuation]				

Si les catégories *N, V, AJ, AV, C, NU, I*, et *PU* nous semblent légitimes, la distinction entre *PD* et *AT*, motivée par les problèmes de délimitation entre pronoms, déterminants et articles paraît plus discutable, dans la mesure où différencier les déterminants des pronoms nous paraît linguistiquement plus justifié que de les distinguer des articles.

Nous ne retiendrons ni la classe *AP*, peu adaptée aux spécificités du français, ni la catégorie résiduelle *R*, dédiée aux éléments exclus des parties du discours traditionnelles comme les mots de langue étrangère ou les formules mathématiques, qui nous semblent devoir être étiquetées.

Les 13 catégories proposées par Eagles demeurant très générales, observons leurs spécifications à travers l'exemple de deux systèmes d'annotation morphosyntaxique du français :

1. le système d'étiquetage développé par l'INaLF pour Brill à partir de la base Frantext (Lecomte et Paroubek, 1996) ;
2. le jeu de descripteurs morphosyntaxiques du français de TreeTagger (http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html);

Le tableau qui suit synthétise les catégories mobilisées par les deux systèmes (*v.* tableaux 1 et 2[20], annexe 9, pour le détail) :

Légende

Codes: correspondent aux systèmes d'annotation documentés dans les tableaux 1 et 2 :
[1] système d'annotation de l'Inalf pour Brill
[2] système d'annotation de TreeTagger

CG: catégorie grammaire traditionnelle le plus souvent, pas de sous-distinction.
XX: catégorie inexistante

POS ou catégorie	Français
ADJECTIFS	[1] Trait nombre (sg/pl) [2] Catégorie générale + distinction des adjectifs numéraux
ADVERBES	CG
COORDONNANTS	[1] Coordonnants + expressions (*soit...soit*) [2] CG
DETERMINANTS	[1] distinction non contractés/ contractés (DU et DES) Trait nombre (sg/pl) [2] déterminants définis/indéfinis adjectifs démonstratifs, possessifs (catég. *pronom*) adjectifs indéfinis
INTERJECTIONS	[1] CG [2] XX
NUMERAUX/ CARDINAUX	[1] nombres cardinaux (chiffres /lettres), dates [2] XX
PARTICIPES PASSES	[1] distinction auxiliaires ETRE et AVOIR/autres contextes Trait nombre (sg/pl) [2] CG
PARTICULARITES	[1] catégorie préfixes détachés, isolés entre deux blancs (*ex.* : micro...)

[20] Tandis que les tableaux 3 et 4 sont consacrés à l'anglais (une synthèse comparée des quatre systèmes d'étiquetage est également proposée).

	[2] XX
PARTICULES	[1] éléments non autonomes et non regroupés dans des expressions figées [2] XX
PONCTUATIONS	[1] syntaxe [ponctuation] / [ponctuation] (*ex.* : !/!, ?/?, *etc.*). [2] distinction des points/virgules/ autres marques de ponctuation
PREPOSITIONS	[1] prépositions simples [2] CG
PRONOMS	[1] distinction pronoms relatifs/non relatifs supportés/non supportés par le V Trait nombre (sg/pl) [2] pronoms démonstratifs pronoms indéfinis distinction pronoms personnels clitiques/conjugués pronoms relatifs
RESIDUS	[1] abréviations, mots étrangers et symboles [2] abréviations et symboles
SUBORDONNANTS	[1] distinction des conjonctions et locutions de subordinations/QUE [2] CG
SUBSTANTIFS	[1] distinction noms communs /propres Trait nombre (sg/pl) [2] regroupement des NC et NP, pas de trait nombre
VERBES	[1] distinction auxiliaires AVOIR/ ETRE/AUTRE VERBE distinction formes conjuguées/non conjuguées (gérondif & participe présent/infinitif) Trait nombre (sg/pl) [2] étiquetage des temps verbaux simples, mais pas des formes composés (auxiliaires)

Tableau : Synthèse des 2 systèmes d'étiquetage du français sélectionnés

A l'exception des adverbes, les catégories mobilisées diffèrent considérablement :

- ✓ Certaines classes sont simplement absentes : ainsi, les temps composés du français, qui posent des problèmes certains de reconnaissance automatique, ne sont pas pris en charge par le système d'annotation du français de TreeTagger. Encore plus problématique est le système développé par l'INaLF pour Brill, qui ne distingue pas les temps verbaux, mais leur forme composée et leur « nombre » (étiquetage des « formes conjuguées au singulier/pluriel »).
- ✓ Certaines catégories sont peu, voire non linguistiques : on note la présence d'une catégorie résiduelle dans les jeux d'étiquettes de l'INaLF. Soulignons qu'il est bien entendu difficile de demeurer à un niveau strictement morphosyntaxique : certains phénomènes morphosyntaxiques posent problème, tandis que d'autres types d'informations n'en posent aucun.

L'examen des deux systèmes d'annotation nous a permis de considérer des éléments que nous n'aurions pas naturellement inclus : il en va ainsi des préfixes, nombreux dans le discours scientifique linguistique, et des particules comme par exemple le –t de *semble-t-il*.

Les variables devant être calculées relativement à leurs classes morphosyntaxiques d'appartenance[21], c'est un ensemble de 15 classes morphosyntaxiques qui a été élaboré, dans lequel ont été intégrés les descripteurs.

[21] Il serait peu pertinent de calculer le poids des virgules dans un texte relativement à l'ensemble des *tokens* le constituant : les variables nous semblent devoir être calculées au sein de leur système linguistique d'appartenance, comme le propose d'ailleurs Cordial *via* ses fichiers .sta.

2.2.2.2. Classes morphosyntaxiques générales

La constitution des classes s'est effectuée progressivement et en corpus : les catégories ont été graduellement créées et adaptées aux caractéristiques des textes, afin d'obtenir un système de description satisfaisant.

Voici les quinze classes morphosyntaxiques que nous avons finalement élaborées :

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 (formellement délimités par un tiret)
8. Prépositions et amalgames	

Tableau : Classes morphosyntaxiques du système d'étiquetage du français

2.2.2.3. Spécification des classes

On ne détaillera ici que les classes dont le contenu mérite un développement ; les adjectifs et les noms sont ainsi peu problématiques, dans la mesure où nous n'avons distingué que leur nombre ; la classe 8 englobe les amalgames[22] et les prépositions, que nous avons différenciées des locutions prépositionnelles en ajoutant un attribut positionnel de type *afin* PREP :1st *de* PREP :2[nd23] ; la classe 12 rassemble les subordonnants simples et de type locution, également discriminés par un trait positionnel. Enfin et étant donné leur contenu non équivoque, on ne détaillera pas les classes 9, 13 et 15.

A. Classe 1 : formalisation

La classe 1 englobe les symboles, les sigles et les abréviations. La catégorie des symboles a dû être affinée étant donné sa grande hétérogénéité : il nous a semblé pertinent de différencier les symboles spécifiques au métalangage linguistique (*e.g.* SN, SV, ?, * ou GN) des symboles logico-mathématiques (*e.g.* +, -, X ou Y).

Nous avons en outre distingué les symboles d'acceptabilité linguistique (*, ?, ?? ou parfois #) des sigles et des abréviations linguistiques (*SN* ou *prep.*), et des morphèmes linguistiques de type *–ant* ou *–wise*. Cette dernière catégorie s'est également avérée servir l'annotation des éléments non attestés, présents dans les tests d'acceptabilité grammaticale comme :

[22] Que nous aurions également pu inclure au sein des déterminants, ou éventuellement comptabiliser deux fois : une fois dans la classe des déterminants et une autre dans celle des prépositions.

[23] Plutôt que de rassembler les locutions prépositionnelles au sein d'un même token, nous avons choisi d'assigner un attribut positionnel à chacun de leurs constituants.

Dans les phrases suivantes Pierre change, mange, range, *cange, *fange, *pange, *tange…, l'existence des actions de *canger, *fanger, *panger, *tanger est présupposée au même titre que celle de changer, manger ou ranger. (007)[24]

Enfin, la catégorie *SYM :s* est destinée aux précisions de type *(s)* ou *(x)*, spécifiques à l'écrit, que l'on relève tant dans les corps d'article que dans les exemples :

(66) pour aborder ce sujet i - il convient de + de s'attacher à à deux aspects euh je dirai(s) dans un premier temps la richesse de de la langue française et dans un deuxième temps sa ses faiblesses et et son déclin (223, exemple)

B. Classe 2 : adverbes et connecteurs

Il nous a paru pertinent de différencier les adverbes des connecteurs, bien que les deux catégories soient généralement appréhendées conjointement. *A fortiori*, nous avons choisi de distinguer des types de connecteurs (opposition, conséquence, etc.) afin de saisir certains aspects de la rhétorique des textes scientifiques et de leur fonctionnement argumentatif.

Si de nombreuses études sont dédiées aux connecteurs, force est de constater qu'il nous a été difficile d'en obtenir un relevé typologique plus ou moins exhaustif. La plupart des travaux recensés s'intéressent en effet au fonctionnement d'un (type de) connecteur spécifique – le nombre de travaux consacrés au connecteur *mais* est par exemple frappant.

Bien que nous admettions que notre entreprise soit quelque peu élémentaire, puisqu'un connecteur ne saurait être réduit à *un* type de relation logique, elle nous semble plus satisfaisante qu'une inclusion des connecteurs avec les adverbes, et qu'une absence de spécification.

Nous nous sommes d'abord fondée sur les travaux de la grammaire du texte (Adam, 1992) et sur une typologie pédagogique des relations logiques[25] ; une première classification a été esquissée, qui a été enrichie et modifiée par l'observation du corpus et l'extraction systématique des connecteurs représentés.

Quinze types de connecteurs ont finalement été dégagés, chaque étiquette englobant une liste finie de formes :

Type de connecteur	Etiquette	Contenu
Addition, énumération	CON :add	et, de plus, de surcroît, par ailleurs, en plus, aussi, d'une part…d'autre part, en outre, ainsi que, de même, non seulement (mais encore)
Causalité	CON :cau	car, du fait (que, de), en ce que, en raison de, parce que, puisque, eu égard à, aussi (début de phrase)
Certitude	CON :cer	tout à fait, certainement
Concession	CON :ces	à la limite, à la rigueur, du moins, en tout cas, quoi qu'il en soit, bien entendu, bien sûr, certes, toujours est-il que, quand même, quoique

[24] Identifiant du texte (*v.* annexe 1).

[25] http://www.lettres.net/pdf/methodes/relations-logiques.PDF

Conclusion	CON :ccl	en fin de compte, en somme, enfin, finalement, après tout, in fine, en définitive, en conclusion
Conséquence	CON :csq	alors, donc, d'où, en conséquence, par conséquent, à ce titre, du coup, de fait (en début de phrase)
Disjonction	CON :dis	ou, soit, ni
Doute, approximation	CON :dou	peut-être, en quelque sorte, grosso modo, sans (nul) doute, probablement, à peu près, en gros
Exemplification/Typification	CON :exe	par exemple, notamment
Justification/Explication	CON :jus	à juste raison, à juste titre, ainsi, d'ailleurs, de facto, effectivement, en ce sens, en effet, en fait, en réalité, ipso facto, par ailleurs, de fait
Opposition/Restriction	CON :opp	a contrario, au contraire, contrairement (à), par contre, en revanche, cependant, néanmoins, pourtant, toutefois, bien que, malgré tout, mais, or, outre, à l'inverse
Présupposition	CON :pre	a priori, à première vue, d'emblée, en principe, en théorie, en vérité, a posteriori, à raison, à tort, apparemment, a fortiori, en règle générale, en général
Reformulation, précision	CON :ref	à savoir, c'est-à-dire, en d'autres termes, autrement dit, bref, en bref, pour ainsi dire, voire, en l'occurrence, en particulier, soit, en termes modernes, pratiques, en clair, en raccourci, en un mot, en d'autres mots
Séquentialité, temporalité	CON :tem	maintenant, puis, d'abord, de prime abord, ensuite, tout à l'heure, primo, secundo, tertio...
Spatialité	CON :spc	ici, là, ci-dessous, dessus, après, ci-contre, en bas (plus bas), en haut (plus haut), infra, supra

Tableau : Typologie des connecteurs

La catégorie des adverbes englobe au final quatre types : les locutions adverbiales, distinguées par un attribut positionnel, les adverbes interrogatifs, les adverbes de négation et les adverbes simples, suffixés par *-ment*.

D. Classe 4 : pronoms

Préfixés par *PRO*, deux types de pronoms ont été distingués : les pronoms contenant un trait /personne/ des pronoms relatifs, indéfinis, démonstratifs, réflexifs et enclitiques (*y* et *en*).

Le premier type de pronoms inclut d'abord les sept pronoms personnels français : *je, tu, il impersonnel, on, nous, vous* et *ils*. Si le pronom *on* ne pose aucun problème d'identification –

ce qui rend son absence des systèmes d'annotation traditionnels surprenante -, la désambiguïsation du pronom *il* (anaphorique et impersonnel) a demandé un travail de vérification au cas par cas considérable.

Afin d'observer les personnes dans les textes, nous avons adjoint un trait /personne/ aux pronoms possessifs, clitiques (*me, te, nous, vous*) et disjoints (*toi, moi, vous, nous*). On observera que le réflexif *se* fait l'objet d'une catégorie PRO :refl, ce qui est probablement discutable.

C. Classe 5 : verbes

L'ensemble des temps verbaux (simples et composés, conjugués / non conjugués, etc.) ont été annotés. Les temps composés ont été étiquetés de la manière suivante :

auxiliaire VER :aux :[temps simple correspondant] + participe passé VER :pper

soit par exemple *il avait* [VER :aux :impf] / *aurait* [VER :aux :cond] considéré [VER :pper].

Au final, les formes suivantes ont été prises en considération :

- ✓ Temps conjugués simples, mode indicatif : présent, futur, imparfait, passé simple ;
- ✓ Temps conjugués composés, mode indicatif : passé composé, plus-que-parfait, passé et futur antérieur ;
- ✓ Mode subjonctif : présent, imparfait, passé – aucune forme au subjonctif plus-que-parfait (*que j'eusse considéré*) n'a été relevée dans le corpus.
- ✓ Mode impératif : présent – aucun verbe conjugué à l'impératif passé (*aie considéré*) n'a pu être observé ;
- ✓ Mode conditionnel : présent et passé ;
- ✓ Infinitifs présent (*considérer*) et passé (*avoir considéré*) ;
- ✓ Participes présent et passé.

A ces formes verbales nous avons adjoint une catégorie sémantique modale regroupant les verbes *devoir, falloir, paraître, pouvoir, sembler* et *vouloir*, à l'origine destinée à la comparaison interlangue français / anglais. L'information s'est avérée pertinente pour la description des textes, ces verbes étant particulièrement représentés.

D. Classe 6 : déterminants

Les déterminants définis, indéfinis, démonstratifs et possessifs ont été distingués. Les possessifs se distinguent par un attribut /personne/ (*e.g.* DET :poss :pp1sn).

E Classe 10 : éléments de langue étrangère

Les éléments de langue étrangère, particulièrement représentés dans les textes scientifiques, et *a fortiori* dans les textes de linguistique (138.46 éléments en moyenne, soit 2% de l'ensemble des formes relevées par texte) se sont avérés problématiques : étant donné la diversité potentielle des langues traitées, nous avons opté pour une catégorie générale, toutes langues confondues. Ce choix a bien entendu affecté l'identification de la catégorie par l'étiqueteur, étant donné l'ambiguïté des formes d'une langue à l'autre.

Les éléments étrangers renvoyant le plus souvent à des exemples ou des citations, ils apparaissent généralement par séquences au sein de l'article, ce qui a fait l'objet d'un travail manuel de vérification important : l'identification correcte de l'ensemble d'une séquence est en effet exceptionnelle, bien que les résultats se soient considérablement améliorés avec le processus d'apprentissage progressif. Au final, l'étiqueteur s'illustre dans l'identification de certaines langues particulièrement représentées, comme l'anglais ou l'allemand, mais obtient des résultats moins brillants lorsqu'il s'agit d'identifier des séquences en espagnol ou en italien, qui partagent de plus nombreuses formes avec le français.

F. Classe 11 : ponctuations

Quatorze marques de ponctuations ont pu être observées dans les textes : accolades, barres obliques et barres obliques inverses (*slash* et *antislash*), deux points, virgules, guillemets doubles et simples, crochets, points, points d'interrogation, d'exclamation et de suspension, parenthèses, points virgules et tirets d'incises.

G. Classe 14 : numéraux

Parce qu'ils sont particulièrement représentés dans les textes scientifiques, nous avons accordé une attention toute spécifique aux numéraux ; outre les cardinaux et les ordinaux, généralement distingués par les étiqueteurs, les dates, indices important d'intertextualité et de référence, ont été annotées.

Les indices de structuration textuelle comme les marqueurs de liste ou de titre (de type 1.2.3.), très présents dans les textes scientifiques, ont été étiquetés et regroupés dans la catégorie.

2.2.2.4. Système de descripteurs final

Le système adopté contient un ensemble de 145 observables dont voici une synthèse :

Tag	Description	Tag	Description
ABR	Abréviation, surtout nom personne	ADV	Adverbe simple
ADV :1st/2nd	Locution adverbiale	ADV:int	Adverbe interrogatif
ADV:neg	Adverbe négation	ADJ :sg	Adjectif singulier
ADJ :pl	Adjectif pluriel	CON:add	Connecteur d'addition
CON:cau	Connecteur de causalité	CON:ccl	Connecteur conclusif
CON:ces	Connecteur concessif	CON:cer	Connecteur de certitude
CON:csq	Connecteur de conséquence	CON:dis	Conn. de disjonction
CON:dou	Conn. de doute	CON:exe	Conn. d'exemplification
CON:jus	Conn. de justification	CON:opp	Conn. d'opposition
CON:pre	Conn. de présupposition	CON:ref	Connecteur reformulatif
CON:spc	Conn. Spatial	CON:tem	Conn. temporel
DET:def	Déterminant défini	DET :indef	Déterminant indéfini
DET:dem	Déterminant démonstratif	DET :poss:[pers]	Déterminant possessif
DTC:sg	Déterminant contracté sg	DTC:pl	Déterminant contracté pl
FGW	Elément étranger	INT	Interjection
LS	Indice liste, titre	NC : sg	Nom commun sg
NC :pl	Nom commun pl	NP	Nom propre
NUM :car	Numéral cardinal	NUM :dat	Numéral date
NUM :ord	Numéral ordinal	NUM :par	Renvoi
PON:accol	Accolade	PON:antslash	Barre oblique inverse
PON:colon	Ponctuation deux points	PON:comma	Virgule
PON:cote	Guillemet simple	PON:croch	crochet
PON:dot	Point	PON:excl	Point d'exclamation
PON:guil	Guillemet	PON:int	Point d'interrogations

PON:par	Parenthèse	PON:Pvirg	Point virgule
PON:slash	Barre oblique	PON:tiret	Tiret de phrase
PREF	Préfixe	PREP	Préposition simple
PREP:1st/2nd	Locution prépositionnelle	PRO:dem	Pronom démonstratif
PRO:clit	En, y, le, lui, la, leur	PRO:clit[pers]	Clit. de personne
PRO:indef	Pronom indéfini	PRO:poss	Pronom possessif
PRO:disj[pers]	Pronom forts	PRO:pp1sn	JE
PRO:pp2sn	TU	PRO:pp3msn	IL, ELLE anaphorique
PRO:pp3sn	ON	PRO :pp3isn	IL impersonnel
PRO:pp1pl	NOUS	PRO:pp2pl	VOUS
PRO:pp3pl	ILS, ELLES	PRO:refl	Pronom réfléchi
PRO:rel	Pronom relatif	SUB (1st)	Subordonnant
PUL	-t	VER:aux:subp	Auxiliaire subjonctif
VER:aux:pres	Auxiliaire présent	VER :aux :fut	Auxiliaire futur
VER :aux :cond	Auxiliaire conditionnel	VER :aux :impf	Auxiliaire imparfait
VER :aux :inf	Auxiliaire infinitif	VER:aux:simp	Auxiliaire passé simple
VER :aux :pper	Auxiliaire passé (été)	VER :cond	V conjugué conditionnel
VER :fut	V conjugué futur	VER :impf	V conjugué imparfait
VER :inf	Infinitif	VER :subp	Verbe subjonctif
VER :pres	V conjugué présent	VER :simp	V conjugué passé simple
VER :subi	Verbe subjonctif impft	VER :mod :cond	Modal conditionnel
VER :mod :fut	Modal futur	VER :mod :impf	Modal imparfait
VER :mod :inf	Modal infinitif	VER : mod :pper	Modal participe passé
VER :mod :subp	Modal subjonctif	VER :parpres	Participe présent
VER:mod:pres	Modal présent	SIG:ling	Acron. (SN) & abbr. ling (prep.)
VER :pper	Participe passé	SYM	Symboles
SIG	Acronymes et sigles	SYM :gram	Morphèmes linguistiques
SYM :ling	Symb. Linguistiques (*, ?, ??, etc.)	SYM :s	Lettre s indicateur du pluriel

Tableau : Système de description adopté

2.2.2.5. Aspects techniques

A. Segmentation

Dans la mesure où TnT ne procède pas à une pré-segmentation des textes et requiert en entrée des textes tokenisés, nous avons dû implémenter un analyseur lexical spécifique (*tokeniser*) en C++ : les outils de segmentation existants sont en effet peu adaptés aux spécificités de nos textes et de nos descripteurs, car centrés pour l'essentiel au niveau peu pertinent du mot (Vergne et Guiguet, 1998).

De manière générale, les ponctuations sont considérées comme des séparateurs par l'ensemble des outils disponibles, ce qui entrave la segmentation des abréviations et des numéraux de type 2.2.3. ou (1) : le programme élaboré ne dispense malheureusement pas d'une vérification des textes segmentés, ce qui est en partie lié au manque de systématicité des auteurs des textes. Les numérotations des titres ou des exemples sont par exemple souvent peu rigoureuses et endossent des formes distinctes de type 2.2.3[pas de point], 2.2.3. ou 2-2-3… / (1a), 1a-, (1a-), (1a.) ou (1.a) Il en va de même des abréviations de type *e.g.* ou *i.e.* qui comprennent inégalement un point final. A l'inverse, les phrases se terminant par un chiffre ou un symbole suivi d'un point demeurent ambiguës.

B. Encodage et constitution des tables de données

L'ensemble du corpus annoté a été balisé en XML selon les recommandations de la TEI P4 (www.tei-c.org/P4X) pour le codage grammatical des corpus. Chaque paire élément-étiquette est annoté <*w*> (word), avec un attribut global *ana* spécifiant sa catégorie morphosyntaxique.

L'adoption d'XML comme format de représentation a permis de mettre le corpus à portée des méthodologies contemporaines d'extraction et de quantification de corpus : nous avons donc principalement[26] utilisé le logiciel d'extraction CR (*CorpusReader*), développé par Sylvain Loiseau pour constituer nos tables de données et procéder à nos extractions :

> CR exploite les enjeux scientifiques d'un format comme XML. Ce format propose en effet une représentation unifiée de phénomènes sémiotiquement hétérogènes, et donc des corpus empiriquement riches. L'objectif de CR est dès lors d'exploiter toute cette richesse pour faire des sélections de sous-corpus, des quantifications, des tables de données, etc. (Loiseau, 2006).

[26] Au départ, nous avions, avec l'aide de L. Martin, que je remercie particulièrement, programmé une macro en Visual Basic pour remplir notre table de données Excel à partir des sorties de TnT, et procédions à nos extractions au moyen de feuilles de styles et de logiciels comme Cooktop ou XMLWrench : CR, disponible sur http://panini.u-paris10.fr/~sloiseau/CR/, nous a considérablement simplifié la tâche, et je remercie vivement Sylvain Loiseau de ses conseils d'utilisation.