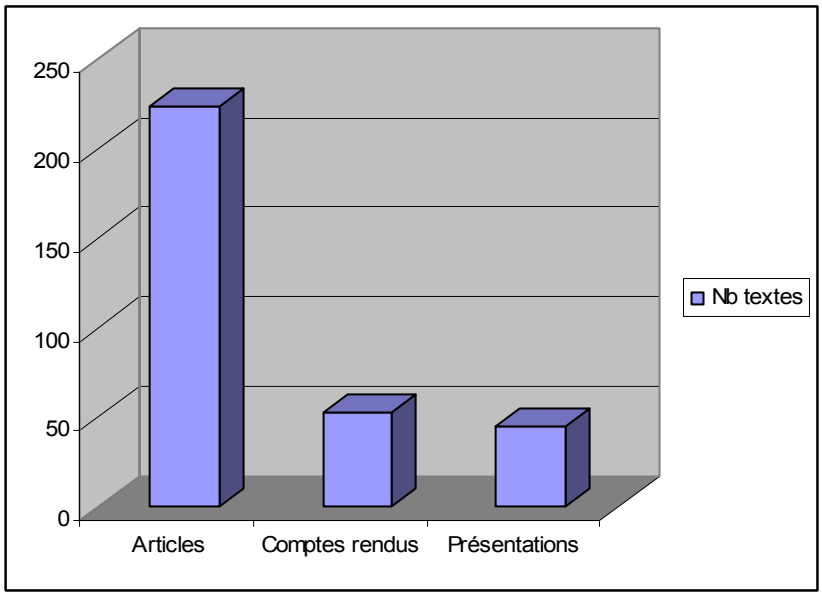
Chapitre 8 : Genres et domaines

8.1. Genres en contraste : les trois genres de la revue de linguistique

Force est d'abord de constater que les revues françaises comprennent un nombre bien plus restreint de genres scientifiques que les revues anglo-saxonnes[1]. Outre l'article, seuls deux genres, dont la présence est loin d'être systématique, ont pu être relevés dans le corpus : *l'article introductif de numéro thématique* et le *compte rendu* (d'ouvrage ou de conférence).

Le genre de l'article a donc été contrasté à ces deux genres[2], ce qui nous a permis de mettre en évidence certaines de ses caractéristiques au sein du domaine scientifique linguistique.

Dans la mesure où les revues constitutives du corpus contenaient peu ou prou de comptes rendus et de présentations, les textes ont souvent dû être extraits d'autres numéros, voire d'autres revues de linguistique. Au final, 53 comptes rendus et 45 présentations de revue ont pu être rassemblés :



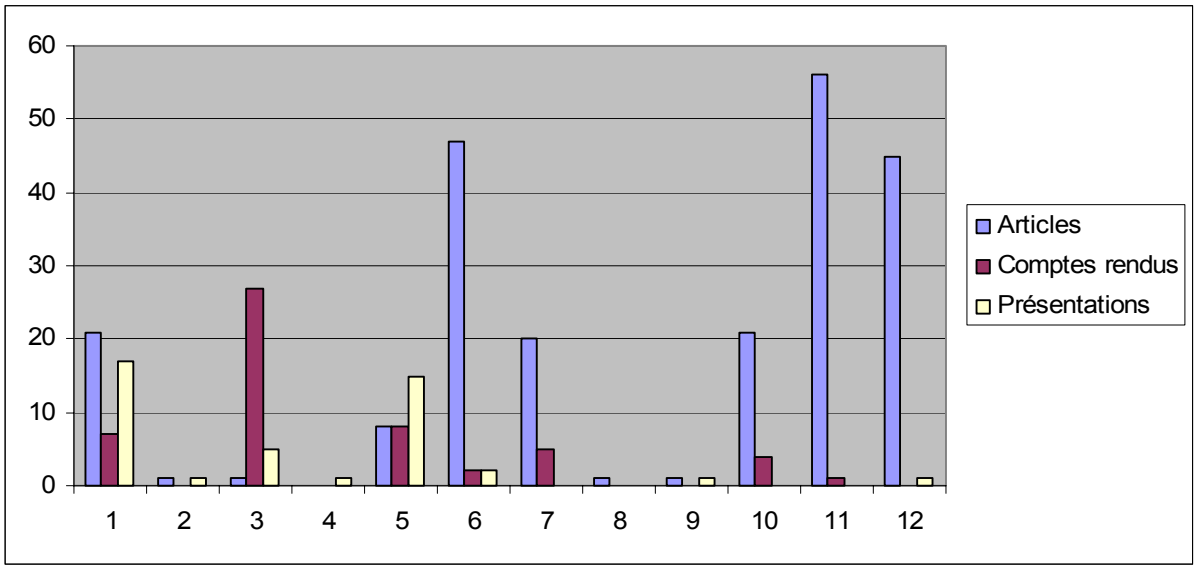
Graphique : Corpus de comparaison

[1] Dans lesquelles le genre de l'*article* s'opposerait par exemple au genre plus polémique de la *discussion,* ou encore au genre post-discussion de la *réponse* (*response*).

[2] Bien qu'il aurait été certainement plus approprié, si nous avions pu disposer de telles données, de le confronter à ses concurrents génériques (acte de colloque, genre oral de la communication, etc.).

8.1.1. Indépendance et proximités des trois genres

Nous avons d'abord cherché à apprécier l'indépendance et les éventuelles proximités des trois genres linguistiques observés. Dans cette perspective, nous avons mené une CAH en douze classes, afin d'observer si les trois catégories étaient correctement classifiées :



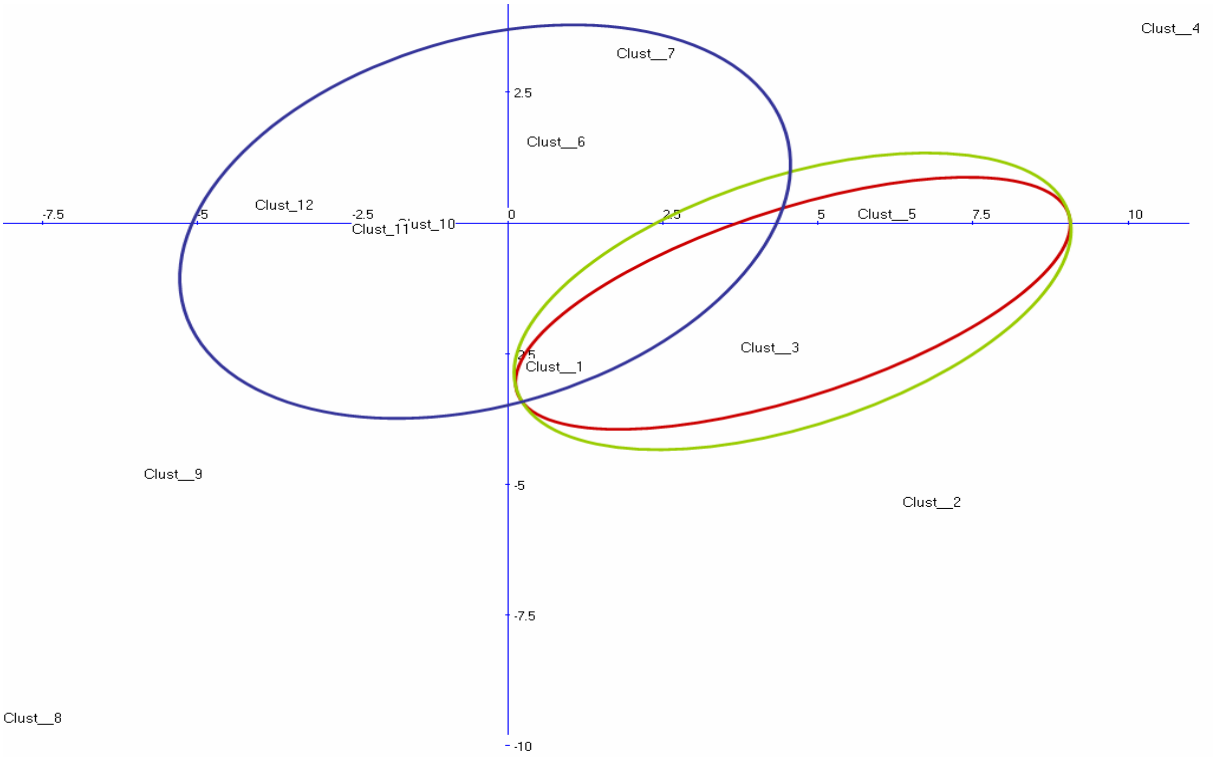
Graphique : CAH en 12 classes : répartition des trois genres

On remarque d'abord que nous n'obtenons aucun cluster strictement homogène en genre – si l'on isole les singletons et les classes de deux items. Rappelons à titre comparatif que certains styles d'auteurs étaient bien plus nettement isolés.

Le compte rendu et la présentation de revue semblent très proches : si les articles sont globalement classifiés dans des classes dans lesquelles ils sont très nettement majoritaires, les classes 1, 3 et 5 fédèrent l'ensemble des textes des deux autres genres.

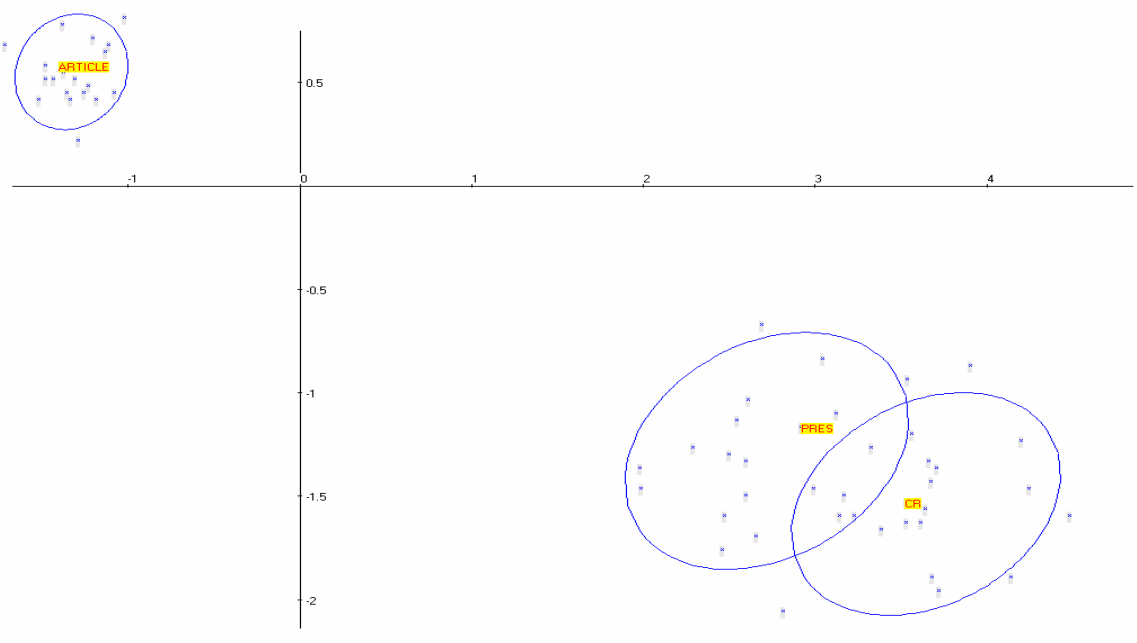
La présentation paraît de surcroît plus proche de l'article que le compte rendu : l'ensemble des comptes rendus est en effet fédéré au sein de la classe 3 qui est presque exempte d'articles, tandis que les présentations se répartissent dans les classes 1 et 5 qui en contiennent une proportion significative.

Ces observations sont confirmées par l'examen du positionnement des classes sur les deux premiers axes factoriels : le genre de l'article (classes cerclées de bleu) s'oppose aux présentations (en vert) et aux comptes rendus (en rouge), qui se recouvrent d'ailleurs largement :



Graphique : Positionnement des classes sur les deux premiers facteurs

L'examen des ellipses de confiance valide ces analyses : on observe bien une intersection des genres du compte rendu et de la présentation, tandis que le genre de l'article est isolé et significativement identifiable (ellipse significativement plus petite) :



Graphique : Ellipses de confiance autour des trois genres

Malgré la taille restreinte des corpus de comparaison et les biais potentiels qu'ils entraînent, le genre de l'article semble morphosyntaxiquement distinct des deux autres genres de la revue, qui seraient comparativement plus proches ; cette proximité est sans doute liée à la visée expositoire et récapitulative que partagent les deux genres.

Le champ générique de la revue se diviserait ainsi en genres centraux et autonomes (l'article) et périphériques – ou secondaires – et tributaires (genres optionnels comme la présentation et le compte rendu), qui partageraient des caractéristiques communes.

Afin d'approfondir et de préciser ces différences, examinons les axes d'organisation interne des trois genres.

8.1.2. Axes d'organisation internes des trois genres

Si l'on a pu observer que le jeu de descripteurs morphosyntaxique développé était peu adapté à la caractérisation des exemples du corpus (chapitre 5), on observe qu'il est globalement plus approprié aux deux autres genres linguistiques ; seuls huit descripteurs sont absents des comptes rendus (pronoms possessifs de première et de seconde personne, disjoints de seconde personne, modaux au passé simple et antislashs) tandis que quinze variables sont absentes du genre de la présentation (déterminants possessifs de seconde personne du singulier, pronoms possessifs, disjoints et clitiques de seconde personne, subjonctif imparfait, modaux au passé simple, accolades et antislashs). Ces variations sont liées au fait que les deux genres ne comportent généralement pas d'exemples, premières composantes écartées d'un travail de récapitulation.

En repositionnant les groupements de descripteurs associés aux pôles d'opposition du genre de l'article sur les nouveaux plans factoriels, on observe des différences d'organisation importantes qui dépassent largement les écarts observés sur le corpus stylistique (v. chapitre précédent). Ces divergences illustrent bien la forte stabilisation morphosyntaxique des genres et l'hétérogénéité intrinsèque du discours scientifique – appréhendé ici au sein d'un même domaine :

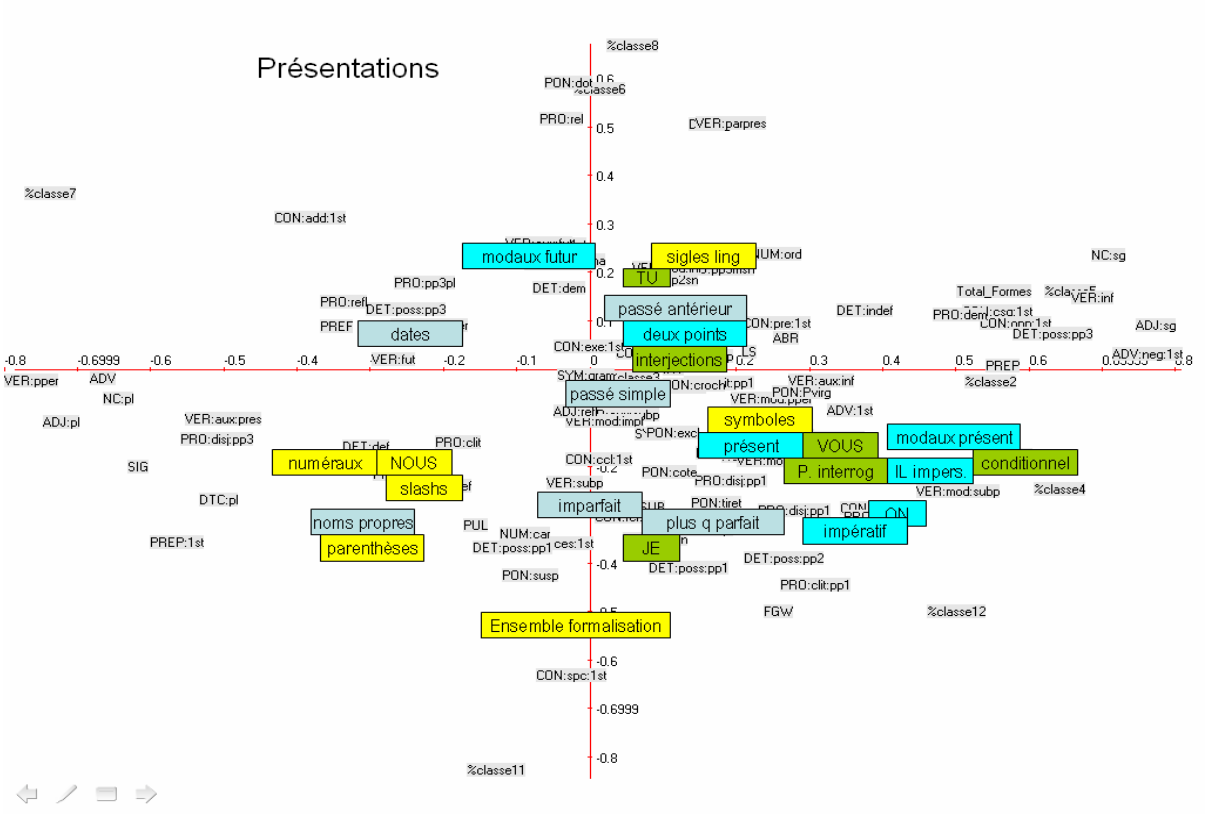


Figure : Positionnement des variables sur les deux premiers axes factoriels – corpus « Présentations »

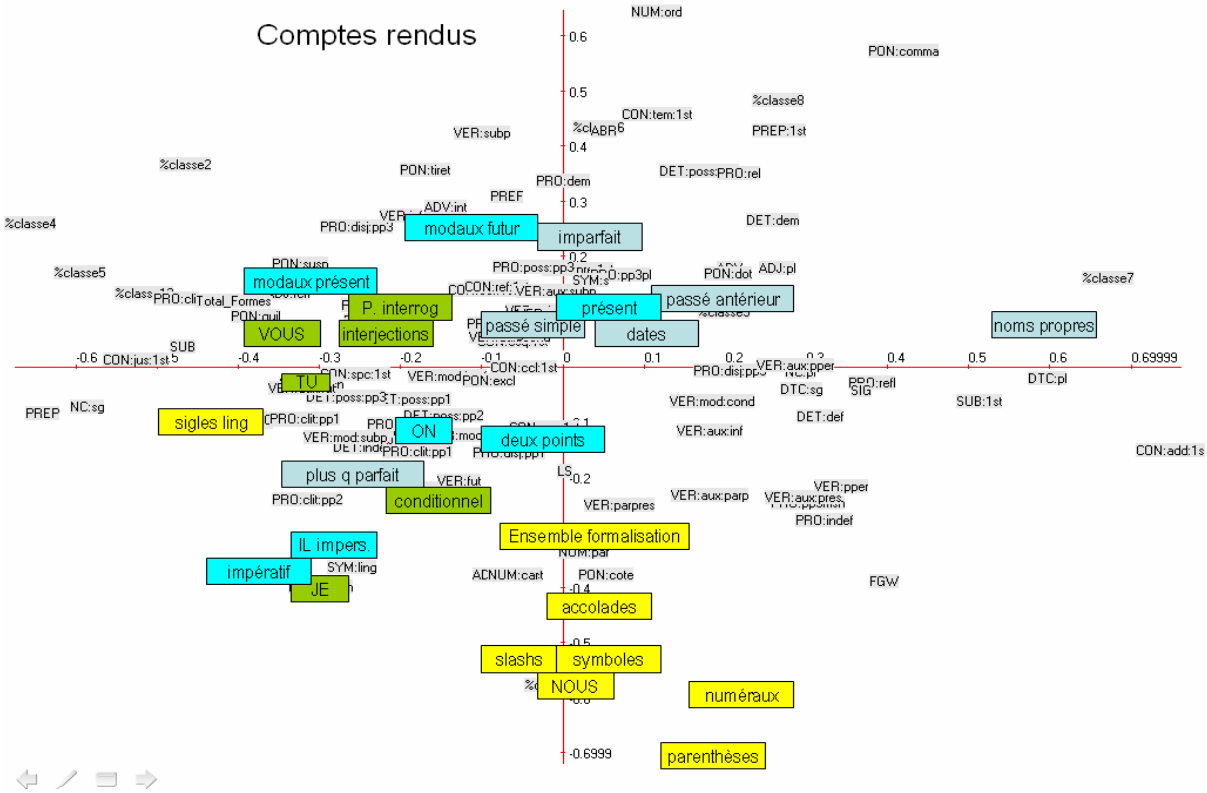


Figure : Positionnement des variables sur les deux premiers axes factoriels – corpus « Comptes rendus »

Bien que les présentations s'étaient avérées plus proches des articles dans la section précédente, il semble que l'organisation morphosyntaxique des comptes rendus et des articles soit finalement plus similaire, ou moins divergente ; le pôle *formalisation* est par exemple plus marqué.

On observe en revanche un pôle additionnel temporel/énumératif (groupement intercorrélé des numéros ordinaux, connecteurs temporels, virgules) des comptes rendus (haut du graphique), tandis que les présentations se caractériseraient par une tension *exposition / introduction vs. discussion / état de l'art* très visible sur l'axe 1, qui renvoie aux deux conceptions du genre que l'on peut observer dans les textes.

8.2. Genre et domaines

Si l'on a pu observer que les genres scientifiques étaient morphosyntaxiquement distincts au sein d'un même domaine, et a fortiori, d'un même champ générique, il convient maintenant d'apprécier la stabilité morphosyntaxique de l'article d'un domaine scientifique à l'autre. On admettra en effet qu'un *article* est généralement identifiable en tant que tel, quel que soit le domaine considéré[3].

Dans cette perspective, le corpus d'articles a été contrasté au corpus « Mécanique » de 49 textes développé par V. Clavier. Domaine appliqué rattaché aux sciences de la nature, la mécanique s'oppose à la linguistique sur différents plans : sciences de la culture *vs.* sciences de la nature, domaine théorique[4] / appliqué, etc. et ce sont précisément ces oppositions qui intéressent notre étude, dans la mesure où les éventuelles similarités qui rapprocheraient les deux disciplines nous semblent plus significatives et plus informatives quant à notre connaissance du genre que celles que partageraient deux disciplines des sciences humaines.

8.2.1. Caractérisation négative : des descripteurs absents des textes de mécanique

Les corpus comparés devant être décrits selon les mêmes procédures et les mêmes descripteurs, on a d'abord constaté que 30 variables, soit 1/5^{e} du jeu d'étiquettes développé, étaient absentes des textes de mécanique : outre les marqueurs spécifiques à la linguistique (*e.g.* symboles linguistiques et grammaticaux), on a pu noter que les marques de première personne du singulier et de seconde personne étaient globalement absentes du corpus. Malgré l'internationalisation de la science et l'usage recommandé du 'I', l'auteur de mécanique semble se conformer à la tradition rédactionnelle française et n'apparaître qu'à travers *nous*.

Certains temps conjugués s'avèrent inemployés (le subjonctif imparfait, de même que les modaux passé simple, imparfait et infinitif), tandis que certaines ponctuations sont absentes des textes : les antislashs, les points virgules et les points de suspension. La mécanique semble ainsi préférer les ponctuations simples, voire plus arrêtées et plus résolues.

Soulignons enfin que l'adjectif réflexif *même* (lui-même, elle-même, etc.) n'a pas été relevé, et qu'on n'a observé aucun connecteur de concession (*à la limite, à la rigueur, quoi qu'il en soit, certes,* etc.), ce qui est notable : il conviendrait ainsi de déterminer sur des

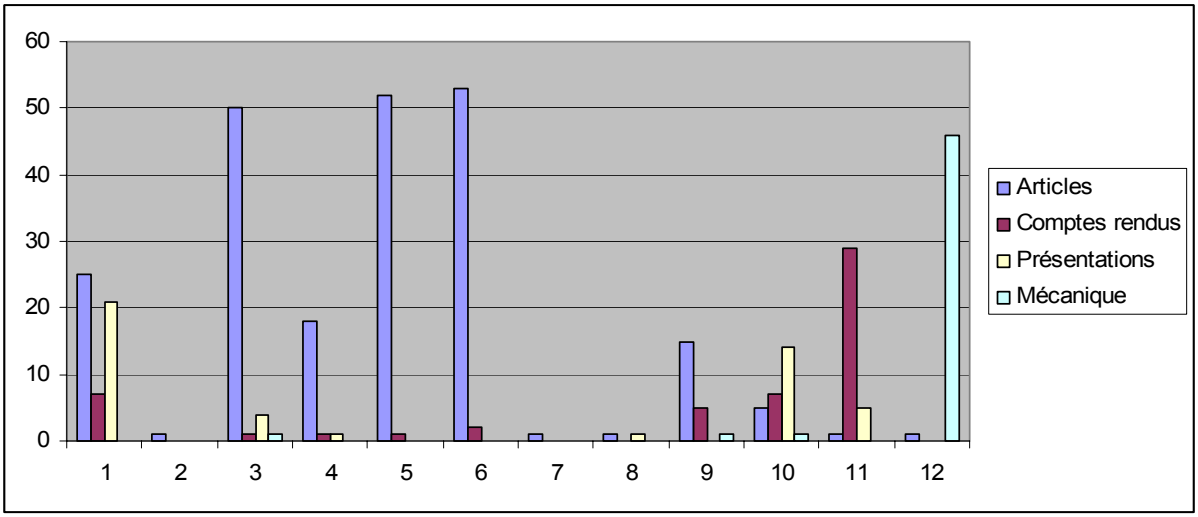
[3] Ou du moins le champ scientifique général considéré (sciences de l'homme et de la société *vs.* sciences de la nature) si l'on se positionne du point de vue du chercheur, les conventions différant sensiblement d'un champ à l'autre.

[4] Si l'on qualifie naturellement l'ensemble de la discipline mécanique comme *appliquée*, il en va différemment de la linguistique.

corpus plus étendus si l'expression de la *concession* est plus spécifique aux sciences humaines car affectée d'une valeur négative dans les sciences dites 'dures', peut-être plus inhibées face à l'interprétation.

8.2.2. Indépendance et proximités des deux domaines

Une CAH a de nouveau été effectuée après adjonction des textes de mécanique aux textes de linguistique (tous genres confondus), et on observe d'emblée que la mécanique est nettement isolée (classe 12) :



Graphique : CAH en 12 classes : genres et domaines

Les catégories domaniales semblent ainsi l'emporter sur les catégories génériques, ce qu'il conviendrait bien entendu de vérifier à plus grande échelle – d'abord en augmentant le corpus de mécanique de genres distincts, puis en ajoutant d'autres domaines scientifiques.

8.2.3. Axes d'organisation internes

Examinons maintenant les deux premiers axes factoriels de l'ACP effectuée sur le corpus de mécanique :

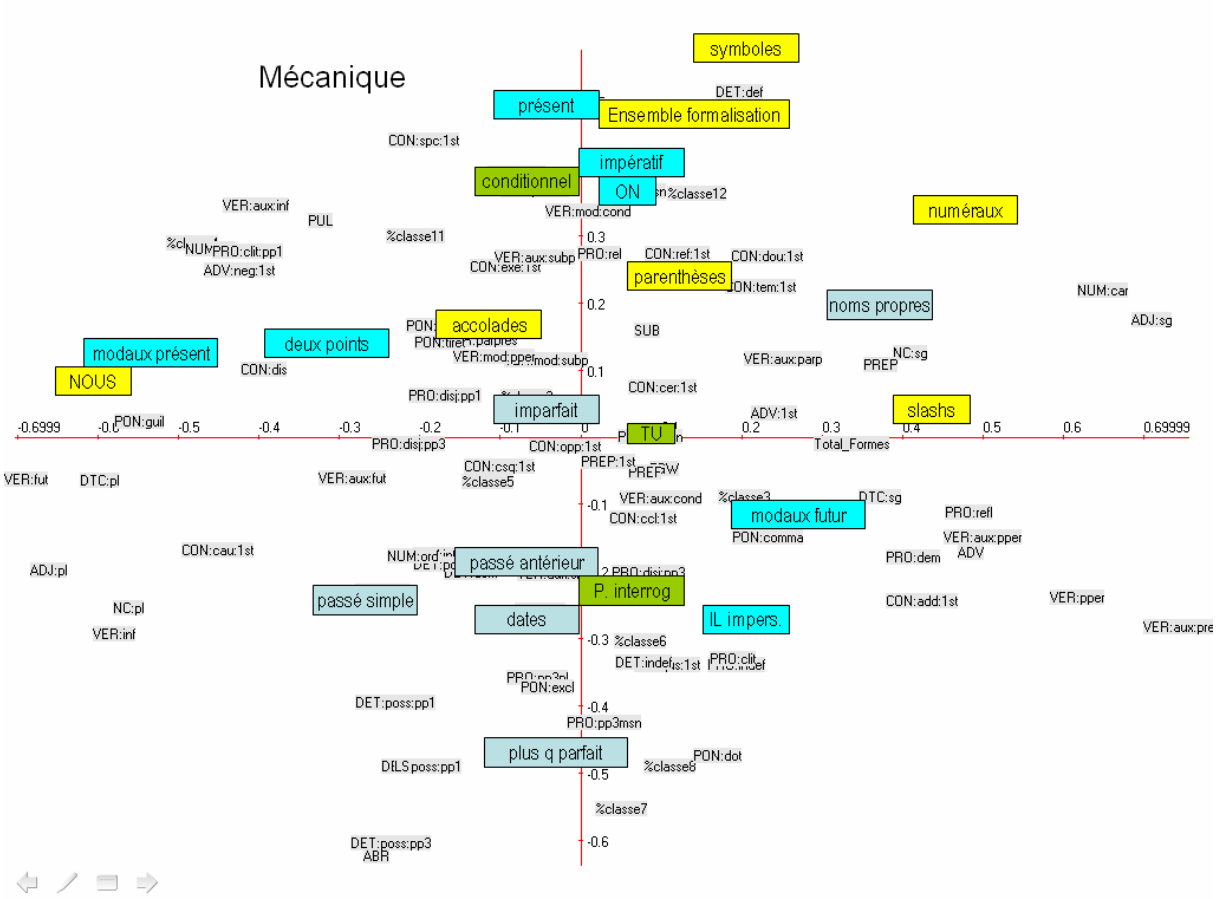


Figure : Positionnement des variables sur les deux premiers axes factoriels – corpus « Mécanique »

Si la dimension orale observée dans les articles de linguistique est bien entendu absente du domaine mécanique, l'organisation morphosyntaxique de ces textes se rapproche déjà plus de celle de l'article de linguistique : on retrouve ainsi l'opposition des pôles *formalisation / mode de narration historico-narratif* sur l'axe 2, et on notera d'ailleurs que la très grande majorité des individus se situe naturellement sur le versant positif de l'axe – rappelons que la majorité des textes de linguistique s'orientait également vers le même pôle :

Figure : Positionnement des individus sur les deux premiers axes factoriels – corpus « Mécanique »

Cette opposition semble ainsi fortement caractéristique du genre de l'article, et on observera que les marqueurs énonciatifs de la rhétorique scientifique (temps du présent, impératif, *on* auxquels on peut ajouter ici les modaux au conditionnel, les connecteurs de spatialité et de présupposition) s'opposent aux temps du passé, et sont davantage corrélés aux indices de formalisation. La rhétorique scientifique à l'œuvre en mécanique diffère d'ailleurs sensiblement de celle observée en linguistique : si *on* est corrélé aux marqueurs de formalisation en mécanique, c'est à *nous* qu'ils étaient associés en linguistique, comme si l'emploi d'un métalangage et de données plus objectives nécessitait une présence plus personnelle – et plus humaine ! – que l'indéfini.

Les rôles des pronoms *on* et *nous* sont globalement inversés : *nous* est ainsi associé aux modaux au présent et aux deux points, de même qu'au futur ; bien que l'impératif soit corrélé à *on,* c'est davantage à *nous* que semble dévolue la fonction de *guide*.

Aussi l'axe 1 opposerait finalement le personnel (sur le versant négatif de l'axe) à l'impersonnel (versant positif, pôle formalisation et données chiffrées), dimension moins marquée dans les textes de linguistique.

On remarque d'ailleurs que les pronoms *il impersonnel* et *on* sont en opposition : *on* s'inscrit dans le pôle scientifique formel, tandis que *il impersonnel* s'inscrit dans le pôle plus marginal *passé*, alors qu'ils étaient fortement corrélés en linguistique.

Ainsi et malgré des différences rhétoriques notables, le genre de l'article imposerait certains phénomènes morphosyntaxiques *a priori* transversaux aux disciplines.

8.3. Ouverture applicative : classification des domaines et des genres

Les sections précédentes ayant mis au jour d'importantes variations morphosyntaxiques des genres et des domaines considérés, nous avons cherché à approfondir ces résultats en évaluant l'impact classificatoire des descripteurs morphosyntaxiques développés dans une perspective plus appliquée de Recherche d'Information (RI).

Un travail collaboratif a ainsi été mené avec G. Cleuziou (LIFO, Equipe *Contraintes et Apprentissage,* Université d'Orléans) et V. Clavier (Gresec, Equipe *Connaissance, Recherche d'Informations, Interfaces et Systèmes de Traitement Automatique de la Langue,* Université de Grenoble), sur les données précédemment décrites. La dimension mathématique et informatique de l'étude a naturellement été effectuée par G. Cleuziou.

Soulignons que cette section est applicative plus que descriptive : les résultats proposés sont souvent triviaux sur le plan descriptif (*e.g.* caractérisation des textes de mécaniques avec le terme *écoulement*), bien que pertinents d'un point de vue classificatoire. La frontière qui sépare les études descriptives des études appliquées est ici particulièrement visible.

De manière générale, les classifications en domaines et en genres représentent un enjeu dans ce domaine applicatif, et nous avons pu constater que les deux catégories étaient rarement utilisées conjointement. Elles sont de fait le plus souvent associées à des variables ou traits appartenant à des niveaux linguistiques différents : les domaines se situeraient sur le plan lexical, tandis que les genres, ou les styles, seraient déterminés au niveau morphosyntaxique.

Ainsi, quand il s'agit de classification thématique ou domaniale, les textes sont souvent décrits en termes de relations lexicales, dans la mesure où ils sont supposés être le reflet de champs de connaissance particuliers. Ils se positionnent donc sur le plan du contenu, que différentes techniques de classification de documents ont tenté d'appréhender. Les mesures les plus fréquentes sont calculées sur les mots, les clusters de mots – inégalement appelés "thèmes", "sujets", "topics", etc. – ou encore les racines (ou word stems) (Porter, 1980), et se sont avérées plutôt efficaces dans diverses entreprises. De manière générale, on demeure au niveau du mot en raison de son faible coût de traitement. Les textes sont donc généralement réduits à l'état de "sacs de mots". Chaque document est alors décrit par le vocabulaire présent dans le corpus. Étant donné la taille de ce vocabulaire, une étape de réduction de l'espace de description est indispensable : sélection d'attributs par des mesures d'intérêt (mesure d'Information Mutuelle, Gain d'Information et mesure du chi2, etc.), reparamétrage de l'espace (LSI, pLSI) ou regroupement d'attributs. Ces formalismes d'indexation permettent d'obtenir des classifieurs performants, atteignant jusqu'à 90% de précision sur grands corpus (Hofmann, 1999, Dhillon et al., 2003).

La classification en genres est quant à elle généralement menée à partir d'un ensemble de parties du discours et/ou de catégories fonctionnelles. A la suite de Biber, c'est l'utilisation de variables morphosyntaxiques qui a été privilégiée pour valider des typologies textuelles et classifier les genres. Les classifications en genres à partir d'un jeu de variables morphosyntaxiques robuste sont d'ailleurs à même d'obtenir de très bons résultats en matière de validation de typologies textuelles (Karlgren et Cutting, 1994, Kessler et al., 1997, Malrieu et Rastier, 2001).

On considère généralement que les genres et les domaines sont des notions orthogonales. Il est souvent souligné qu'on peut retrouver des domaines distincts à l'intérieur de genres

différents, et inversement, ce qui laisserait en effet penser que les deux dimensions sont distinctes. Les deux niveaux de caractérisation des notions sont par conséquent rarement utilisés de manière conjointe ; certaines études (e.g. Lee, 2002) ont pourtant corrélé des variables lexicales aux genres et ont obtenu des résultats tout à fait encourageants. La classification des domaines à partir du niveau morphosyntaxique reste encore, à notre connaissance, en suspens. Pourtant, il semble qu'à l'instar des genres, les domaines sont susceptibles d'entraîner des régularités stylistiques. Prenons par exemple le cas du discours scientifique : la pratique sociale de la "communication scientifique" a donné lieu à la création d'un ensemble de genres tant écrits qu'oraux (articles, présentations de conférence, etc.), dans laquelle on retrouve des "domaines" correspondant aux différentes aires de l'activité scientifique (médecine, économie, recherche d'information, informatique, etc.). L'ensemble des productions de cette pratique communicative, qui partagent des propriétés linguistiques communes, forme ce que l'on appelle le "discours scientifique". Si les genres ont développé au sein de cette pratique une structure et un style propre qui permettent de les identifier par-delà les domaines – on reconnaîtra un article scientifique, qu'il porte sur le domaine médical, biologique ou informatique-, il ne paraît pas inconcevable d'émettre l'hypothèse que les domaines sont discriminables au moyen de variables morphosyntaxiques.

Soulignons a fortiori que la plupart des travaux recensés effectuent de la classification domaniale sur corpus génériquement homogènes (e.g. Reuters ou Newsgroup), et de la classification générique sur corpus discursivement[5] hétérogènes (e.g. Karlgren, Kessler, Malrieu et Rastier), ce qui augmente le pouvoir classificatoire des variables employées mais limite l'utilisation conjointe et l'évaluation de la portée des deux niveaux descriptifs. Bien que de nombreuses applications de Recherche d'Information[6] partent de données génériquement hétérogènes mais de même domaine, ce type de classification demeure problématique et est rarement mené étant donné la robustesse des jeux de variables utilisés.

Notre objectif étant d'évaluer l'intérêt des niveaux morphosyntaxiques et thématiques en matière de classifications en genre et en domaine, il nous a semblé primordial d'engager cette entreprise sur un corpus textuel discursivement homogène[7], quitte à étendre l'étude à un corpus plus large et plus hétérogène dans une étape ultérieure.

Après avoir exposé la méthodologie adoptée pour réaliser l'expérience, on présentera les expérimentations menées et les conclusions qu'elles appellent.

8.3.1. Méthodologie

8.3.1.1. Descripteurs et corpus

Parmi les variables lexicales envisageables, ce sont les substantifs les plus fréquents que nous avons sélectionnés. En effet, les noms sont des parties du discours non vides susceptibles de pointer sur des concepts scientifiques, contrairement aux adverbes, verbes ou adjectifs. Ils

[5] Discours littéraire, juridique, scientifique, journalistique, etc. Les types de discours sont reliés à des pratiques sociales distinctes et organisent en leur sein les typologies génériques et domaniales. Le discours juridique inclut ainsi les genres de l'arrêt, du décret, de la loi, etc.

[6] Entendue ici comme un ensemble de techniques et de technologies de traitement automatique, de gestion et de diffusion de l'information.

[7] Il semblerait en effet que les *discours* sont les premiers à émerger au niveau morphosyntaxique, bien avant les genres, les domaines ou les styles personnels [MAL 01]). Étant donné que nous nous intéressons aux notions de genres et de domaines dans la présente étude, il semble pertinent de négliger momentanément le problème des discours.

sont donc potentiellement plus discriminants et présentent l'avantage d'être facilement extractibles. Le poids des substantifs au singulier et au pluriel (dans la mesure où ils sont susceptibles de renvoyer à des concepts différents, e.g. "la langue" en linguistique ne renvoie pas à la même notion que "les langues") a également été pris en compte.

La classification morphosyntaxique s'est fondée sur le jeu de 136 descripteurs morphosyntaxiques dédiés à l'observation du discours scientifique linguistique.

Si nous nous sommes fondée sur les quatre corpus précédemment décrits, les spécificités des expérimentations présentées *infra* nous ont amené à effectuer différentes partitions du corpus correspondant à des tâches de classification distinctes :

ART-corpus renvoie aux *articles* du corpus (224 articles de linguistique + 49 articles de mécanique)

LING-corpus renvoie aux textes appartenant au domaine linguistique (224 + 45 + 53 textes)

Nous différencierons également des corpus *local* et *global* : *global* renvoie à l'intégralité du corpus, tandis que *local* renvoie à un sous-corpus homogène en genre ou en domaine.

8.3.1.3. Méthodes de classification

La classification (ou catégorisation) automatique de documents a donné lieu à de nombreux travaux recourant aux méthodes d'apprentissage automatique. Parmi les méthodes les plus utilisées dans ce domaine d'application, on mentionnera le classifieur naïf de Bayes (Lewis et Ringuette, 1994), les machines à support vectoriel (SVM) (Joachims, 1997) ou encore les arbres de décision (Cohen et Hirsch, 1998).

Les expérimentations proposées visent à (1) évaluer l'influence de chaque type de description sur la classification (précision du classifieur) et (2) observer l'articulation des deux ensembles d'attributs combinés dans un même classifieur. Dans cette perspective, nous utilisons deux méthodes très différentes bien que complémentaires de ce point de vue, à savoir la classification par SVM et par arbres de décision.

Les SVMs sont reconnus pour leurs performances inégalées dans l'application à la catégorisation de textes (Dumais *et al.*, 1998). De manière simplifiée, cette méthode consiste à apprendre un classifieur dans un nouvel espace d'attributs de dimension plus importante que l'espace initial. Ce nouvel espace peut-être obtenu par différents types de fonctions noyaux (*e.g.* linéaire, polynomial, RBF, etc. *v.* Vapnik, 1995 pour plus de précisions sur la technique d'apprentissage par SVM). Plusieurs études empiriques (*e.g.* Dumais, 1998) ayant montré que les meilleurs performances en classification textuelle sont obtenues avec des SVMs linéaires, c'est ce type de noyau que nous avons retenu dans nos expérimentations. La classification par SVM permettra alors d'appréhender quantitativement l'importance de chaque ensemble d'attributs : lexical, morphosyntaxique et combiné, notés respectivement **S, M** et {**M + S**}.

Les Arbres de décision (AD), contrairement aux SVMs, procèdent par apprentissage symbolique. Bien que moins performants sur cette application, les arbres générés par cette méthode permettent l'analyse et l'interprétation du rôle joué par chaque attribut. La présence et la position d'un attribut dans l'arbre indiquent son importance dans le processus de classification ainsi que la classe favorisée par ce dernier. De l'arbre peut être extrait un ensemble de règles explicatives « caractérisant » les classes ciblées. Dans nos expérimentations, nous utiliserons l'algorithme C4.5 (Quinlan, 1993).

8.3.1.4. Evaluation

Soient **D** un ensemble de textes scientifiques et **C** un ensemble de classes (genre ou domaine selon l'étude) tel qu'à chaque document $d_i \in \mathbf{D}$ est associée une classe $c_j \in \mathbf{C}$, on notera **S** l'ensemble des substantifs singuliers et pluriels apparaissant dans les documents de **D** et **M** un ensemble pertinent de variables morphosyntaxiques pour caractériser ces mêmes documents.

Afin d'évaluer le pouvoir classifiant de chaque type de description (lexicale et/ou morphosyntaxique), nous serons amenés à considérer 3 ensembles d'attributs (étant donné une taille nb pour l'ensemble) :

- $\mathbf{S}_{nb}$: constitué des nb premiers substantifs de **S**, ordonnés par Information Mutuelle (IM) décroissante,
- $\mathbf{M}_{nb}$: constitué des nb premières variables morphosyntaxiques ordonnées arbitrairement dans M,
- $\{\mathbf{M} + \mathbf{S}\}_{nb}$: constitué pour moitié de variables morphosyntaxiques et pour moitié des premiers substantifs de **S**.

L'ordonnancement des substantifs dans **S** est effectué sur la base de l'IM entre chaque substantif $\mathbf{m_i}$ et chaque classe c_j :

$$MI(s_i,c_j)=\log\frac{P(s_i,c_j)}{P(s_i)P(c_j)}$$

Dans cette définition, $\mathbf{P(s_i)}$ est donnée par le rapport du nombre d'occurrences de $\mathbf{s_i}$ dans les documents de **D** sur le nombre total d'occurrences de substantifs dans ces mêmes documents. $\mathbf{P(c_j)}$ représente la proportion de documents étiquetés $\mathbf{c_j}$ dans **D**. Enfin, $\mathbf{P(s_i, c_j)}$ correspond à la probabilité que, pour un document **d**, le mot si apparaissent dans d et que ce document appartienne à la classe $\mathbf{c_j}$. Par la règle de Bayes, on se ramène à la définition suivante :

$$P(s_i,c_j)=\frac{\sum_d n(s_i,d)}{\sum_s\sum_d n(s,d)}+\frac{\sum_{d\in c_j}\delta(s_i,d)}{\sum_d\delta(s_i,d)}$$

Le choix de cette mesure d'intérêt des mots est le résultat d'une étude comparative entre plusieurs mesures ou coefficients (Information Mutuelle, Rapport de chance, coefficient GSS)[8]. De même, cette étude préliminaire nous a conduit à privilégier la fonction globale d'intérêt suivante (plutôt qu'une fonction de maximum ou de moyenne non-pondérée) :

$$intérêt(s_i)=\sum_{c_j}P(c_j).IM(s_i,c_j)$$

Les expérimentations présentées par la suite correspondent à des résultats moyens obtenus sur cinq validations croisées à deux blocs (*2-fold cross validations*) : **D** est divisé en deux

[8] Voir (Sebastiani, 2002) pour plus de détails sur ces mesures.

sous-ensembles de tailles équivalentes, chaque sous-ensemble étant utilisé à son tour comme corpus d'entraînement et de test. Les valeurs reportées correspondent à des micro-précisions[9].

Concernant l'apprentissage par SVM, dans le cas de problèmes multiclasses, plusieurs SVMs sont appris (un par classe) puis combinés.

8.3.2. Expérimentations

Nous considérerons, dans ce qui suit, plusieurs sous-corpus correspondant chacun à une tâche différente de classification. En premier lieu, les expérimentations présentées porteront sur une classification en domaines et utiliseront deux corpus : un corpus « local » homogène en genre, rassemblant les 273 articles du corpus *vs.* le corpus « global ».

Dans le premier cas la tâche de classification consistera à distinguer les deux domaines "linguistique" et "mécanique" pour un ensemble de documents homogène en genre (uniquement des articles). Le corpus "global" permettra en revanche d'appréhender l'introduction d'un paramètre de variation générique (articles, présentations et compte-rendu).

De façon analogue, dans un second temps, la classification en genre sera expérimentée sur un corpus "local" homogène en domaine (322 documents de linguistique) puis sur le corpus "global" faisant intervenir une variation générique au sein des domaines.

8.3.2.1. Classification en domaines

Les résultats obtenus avec la méthode SVM (figures 180 et 181) montrent que les variables morphosyntaxiques sont plus discriminantes que les variables lexicales. De plus, on note qu'une utilisation conjointe des deux types de variables est globalement plus efficace que chacun des deux ensembles choisis séparément.

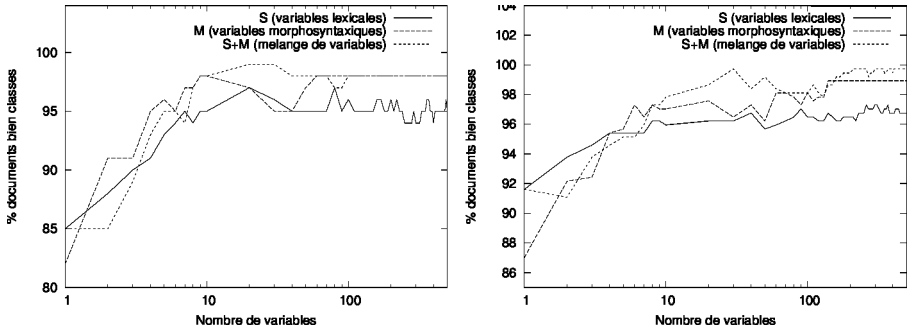


Figure : Classification en domaine par la méthode SVM sur corpus "global".

Figure : Classification en domaine par la méthode SVM sur corpus "local".

On obtient donc l'ordre de précédence suivant, avec ou sans variations génériques :

{Utilisation conjointe des deux niveaux} > {Niv. morphosyntaxique} >> {Niv. lexical}

[9] La *micro-précision* mesure la proportion de textes classés correctement, quelle que soit la classe. *A contrario,* la *macro-précision* mesure pour chaque classe séparément la proportion de textes bien classés avant d'effectuer la moyenne.

D'autres tests, effectués avec un classifieur de type "arbre de décision" indiquent les mêmes tendances, bien que les taux de précision obtenus par le second classifieur soient moins bons qu'avec la méthode SVM. L'indexation par le lexique semble également moins pertinente qu'une indexation morphosyntaxique ou mixte.

Il semble donc que les domaines scientifiques se distinguent davantage par des traits stylométriques que par des informations lexicales, constat surprenant si l'on considère que les deux domaines à discriminer (linguistique et mécanique) sont conceptuellement très éloignés.

8.3.2.2. Classification en genres

Les résultats obtenus avec le classifieur SVM (figures 182 et 183) confirmeraient l'hypothèse selon laquelle les genres sont effectivement corrélés au niveau morphosyntaxique : le taux de précision obtenu est plus élevé avec le jeu de variables morphosyntaxiques qu'avec les variables lexicales. Notons que les différences de domaines ne perturbent pas cet ordre.

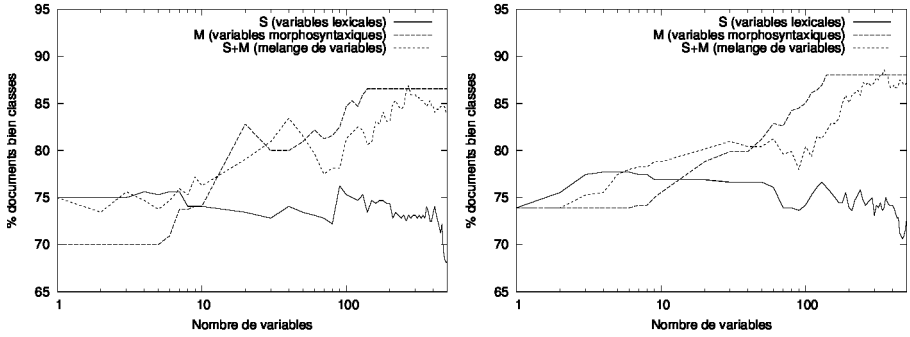


Figure : Classification en genre par la méthode SVM sur corpus "global".

Figure : Classification en genre par la méthode SVM sur corpus "local".

Nous présentons en figures 184 et 185 les résultats obtenus avec le classifieur de type "arbre de décision". On observe en premier lieu que les taux de précision obtenus avec cette méthode sont encore une fois sensiblement inférieurs aux résultats obtenus avec la méthode SVM : 84% au mieux avec C4.5 contre 88% avec SVM.

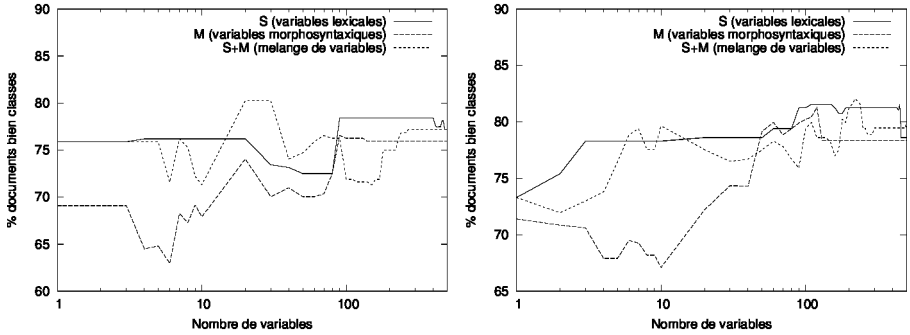


Figure : Classification en genre par l'algorithme C4.5 sur corpus "global".

Figure : Classification en genre par l'algorithme C4.5 sur corpus "local".

L'ordre de précédence établi précédemment *({Niv. morphosyntaxique} > {Utilisation conjointe des deux niveaux} >> {Niv. lexical}*) diffère avec cette nouvelle approche : les variables lexicales sont plus efficaces au niveau global, ce qui confirmerait l'existence d'une possible corrélation des genres avec le niveau lexical, hypothèse soutenue par Lee et Myaeng [LEE 02], qui associent des traits lexicaux au genre de la "homepage".

D'un point de vue plus technique, ces différences obtenues entre les deux classifieurs[10] peuvent en partie s'expliquer par les méthodes très différentes auxquelles ces deux classifieurs font appel. Notamment, l'approche SVM considère un nouvel espace de représentation des documents, à forte dimensionalité, et dont les dimensions sont définies par combinaisons - ici linéaires - des descripteurs initiaux. Cette méthode fait donc intervenir de façon plus ou moins marquée l'ensemble des descripteurs tandis que la construction d'un arbre de décision nécessite généralement très peu de descripteurs bien choisis.

8.3.2.3. Analyse complémentaire : micro- vs. macro-précision

Avant de fournir une explication plus précise des résultats précédents par l'étude des arbres de décision appris, nous proposons un résultat intermédiaire synthétisant l'ensemble des expérimentations présentées ci-dessus.

Pour un nombre fixé de descripteurs, nous étudions dans le tableau 186, les macro et micro-précisions induites par les arbres de décisions appris sur le corpus global. Cette étude a son importance compte tenu des grandes variations de tailles entre les classes, aussi bien pour la classification en domaines que pour la classification en genres.

Type de classification	**Type de précision**	**Nature et taille de l'ensemble de descripteurs**		
		M136	S500	$\{M + S\}_{500}$
Domaine	micro	92.2%	93.3%	94.1%
	macro	**80.3%**	**80.4%**	**84.8%**
Genre	micro	79.9%	80.1%	81.1%
	macro	59.3%	61.9%	61.4%

Tableau : Micro et macro-précisions sur corpus global par l'algorithme C4.5.

L'analyse en terme de macro-précision révèle certains phénomènes masqués par l'influence d'une classe fortement majoritaire (60% des documents du corpus global sont des articles linguistiques). Notamment pour la tâche de classification en domaine, on observe une nette accentuation de la pertinence de l'ensemble combiné de variables lexicales et morphosyntaxiques (+4.5%). En effet, on note beaucoup plus de documents du domaine de la mécanique classés en linguistique avec les niveaux de descriptions M ou S qu'avec une description combinée.

Cette remarque confirme une nouvelle fois la complémentarité entre les deux niveaux de description pour la classification en domaines.

[10] Cette observation confirme l'importance d'utiliser plusieurs méthodes de classification, utilisant des principes d'apprentissage différents, voire complémentaires.

8.3.3. Analyse des descripteurs discriminants

8.3.3.1. Les descripteurs de domaine

On reporte dans le tableau 187 les variables apparaissant dans au moins deux arbres de décision sur l'ensemble des 10 arbres obtenus sur 5 validations croisées.

Morphosyntaxiques	**Variables Lexicales**	**Mixtes**
Indices de renvois (e.g. "voir en **1.1")** Pronoms personnels Prépositions Symboles, sigles, abréviations Participes passés modaux Adverbes et connecteurs Pronoms clitiques	*équation* *écoulement* *vitesse* *coefficient* *déformation* *amélioration* *augmentation* *courbes* *essais* *laboratoire* *mécanique* *vitesses*	*équation* *vitesse* *écoulement* *vitesses* *laboratoire* Adjectifs réflexifs Locutions adverbiales Adverbes et connecteurs Connecteurs de concession Nombre de "JE" Prépositions Ponctuation (points)

Tableau : Descripteurs morphosyntaxiques, lexicaux et mixtes discriminants en matière de classification en domaines.

Les variables lexicales discriminantes sont toutes caractéristiques du domaine scientifique mécanique. Par exemple, on observe sur un échantillon que si le terme "écoulement" apparaît au moins deux fois, il permet de discriminer la moitié des textes de mécanique du corpus d'entraînement. Les textes de linguistique sont donc différenciés de manière négative : 90% du corpus linguistique est bien classé dans le même échantillon s'ils contiennent au plus une fois le terme "écoulement" et ne contiennent ni "mécanique", ni "vitesse" et ni "essais". Cette discrimination par des termes de mécanique nous semble liée à deux raisons : d'une part, la taille plus importante des textes de linguistique augmente le nombre et la diversité des descripteurs, et d'autre part, les textes de mécanique semblent plus homogènes au niveau lexical.

Les descripteurs morphosyntaxiques les plus discriminants semblent par contre plus caractéristiques du domaine linguistique : par exemple, on observe sur un échantillon que la variable "préposition", lorsqu'elle dépasse un certain seuil, permet de différencier jusqu'à 90% du corpus linguistique d'entraînement. De même, un nombre élevé de pronoms personnels et de marques de renvois discrimine les textes de linguistique des textes de mécanique.

En ce qui concerne les classifications mixtes, notons qu'elles recourent davantage aux variables morphosyntaxiques qu'aux variables lexicales malgré la prépondérance des traits lexicaux dans l'espace de description (364 L *vs*. 136 M). Pourtant, les variables lexicales interviennent toujours en premier dans l'arbre de classification, les traits morphosyntaxiques permettant de préciser les classes. Elles sont donc les plus discriminantes, mais ne suffisent pas à classer les documents de manière satisfaisante. Le rôle du niveau morphosyntaxique est donc loin d'être négligeable en matière de classification en domaines.

On notera que ces résultats contiennent des indices descriptifs susceptibles d'intéresser la caractérisation des domaines.

8.3.3.2. Les descripteurs de genre

On reporte dans le tableau 188 les variables apparaissant dans au moins trois arbres de décision sur l'ensemble des arbre appris.

On notera que les arbres de décision font intervenir plus de variables lexicales pour classifier les genres que pour la classification des domaines, ce qui ne semble pas surprenant. Les substantifs présentés dans le tableau 188 sont caractéristiques des comptes-rendus et des présentations de revue. Les articles sont donc classés relativement à l'absence de marqueurs caractéristiques des deux autres genres : ainsi, la quasi totalité des articles est correctement classée si les textes ne contiennent ni "contributions", ni "chapitres" et au plus une occurrence de "chapitre", les contributions étant aussi bien caractéristiques des comptes-rendus que des présentations de revue. "chapitres" permettrait par contre de discriminer les comptes-rendus. Certains indices lexicaux semblent donc caractéristiques du genre, conformément à ce que soutiennent (Lee et Myaeng, 2002). Toutefois, les éléments lexicaux ne sont pas aussi efficaces pour distinguer les genres que pour la classification des domaines, les genres n'étant pas discriminés de manière aussi claire que les domaines.

Les variables morphosyntaxiques semblent caractéristiques des articles scientifiques : ainsi, les indices de structuration textuelles sont particulièrement discriminants et interviennent d'ailleurs en premier dans la plupart des arbres de classification. En effet, les comptes rendus ne sont jamais structurés, à l'inverse des articles et des présentations de revue. Notons que si les articles sont caractérisés par un niveau élevé de structuration, il n'en va pas de même des présentations, qui peuvent être structurées sans que cela soit pour autant caractéristique du genre.

Morphosyntaxiques	**Variables Lexicales**	**Mixtes**
Indices de structuration textuelle Noms propres Passifs/passés composés Symboles Ponctuation (deux points) Ponctuation (points) Connecteurs de conséquence Éléments de langue étrangère Indices de renvois Pronom personnel "NOUS" clitique	*chapitres* *contributions* *articles* *presses* *chapitre* *bibliographie* *journées* *linguistique* *numéro* *politique*	LS *articles* *chapitres* *contributions* Passifs/passés composés Connecteurs de concession Connecteurs spatiaux Éléments de langue étrangère Indices de renvois Pronom personne "NOUS" clitique

Tableau : Descripteurs morphosyntaxiques, lexicaux et mixtes discriminants en matière de classification en genres

Enfin, en ce qui concerne la classification mixte, on note que seuls trois items lexicaux participent à la classification de manière significative : les substantifs "articles", "chapitres" et "contributions", qui sont d'ailleurs non caractéristiques des articles. De la même manière que pour la classification à partir du plan morphosyntaxique seul, les indices de structuration interviennent en premier dans l'arbre de classification.

8.3.4. Conclusion

Nous avons cherché à évaluer de manière expérimentale l'incidence des niveaux morphosyntaxiques et lexicaux sur la classification en domaines et en genres dans le cas particulier des textes scientifiques.

Dans cette perspective, un ensemble de descripteurs morphosyntaxiques adapté aux caractéristiques du discours scientifique a été développé. Nous avons parallèlement opté pour le choix des substantifs au singulier et au pluriel au niveau lexical, dans la mesure où ils pointent potentiellement sur des concepts.

Bien qu'ils aient été obtenus sur un corpus de taille restreinte, les résultats de l'expérience sont particulièrement encourageants parce qu'ils soulignent l'intérêt d'une complémentarité des deux niveaux pour la classification en domaines et confirment celui des variables morphosyntaxiques en matière de classification en genres. En effet, la discrimination des domaines est nettement plus précise si l'on utilise les deux jeux de variables conjointement avec les deux types de classifieurs employés, dans la mesure où les variables morphosyntaxiques permettent d'affiner considérablement les partitions obtenues avec le lexique. Enfin, nous avons pu apprécier l'intérêt de la méthode SVM par rapport à la méthode C4.5 en matière de classification générique morphosyntaxique. Nous n'écartons pas toutefois l'intérêt de l'utilisation de variables lexicales pour discriminer les genres, l'étude des descripteurs s'étant avérée encourageante.

Nous envisageons d'approfondir et de préciser les résultats obtenus sur d'autres types de domaines et de genres. La pertinence des descripteurs utilisés sera également évaluée plus exactement : le jeu de variables morphosyntaxiques employé sera ainsi comparé aux jeu d'étiquettes du Penn Treebank Corpus utilisé par des taggers comme Brill ou TreeTagger par exemple, et d'autres types de descripteurs lexicaux seront extraits afin de d'évaluer la pertinence de l'approche substantivale que nous avons adoptée.