Le genre comme point d'accès au document : analyse comparée de textes scientifiques en mécanique et linguistique

Viviane Clavier

GRESEC – Université Stendhal (Grenoble 3) – France

Abstract

We consider that, as an important feature of textuality, text genre can be described thanks to morphosyntactic cues. This characterization may improve information retrieval which is traditionally concerned with lexical content analysis of texts. We compare two similar text genres, research articles and conference proceedings, in both domains of linguistics and mechanics. The descriptive analysis shows that the morphosyntactic cues reveal common genre features. The document classification can therefore be proceeded from the annotated corpus to classify texts in genre and domain. Further investigations should be made on the interest of text genre for textual indexation.

Résumé

Nous nous intéressons au genre textuel comme plan d'organisation de la textualité et comme point d'entrée possible pour des applications en recherche d'information. En nous appuyant sur une caractérisation morphosyntaxique des genres, nous montrons que l'analyse statistique descriptive et multidimensionnelle de genres proches, des articles de revue et des actes de communication, dans des disciplines distinctes, la linguistique et la mécanique, présente des traits communs. Cette analyse valide la notion de genre au-delà des domaines. Nous montrons que l'utilisation conjointe de variables morphosyntaxiques et lexicales améliore la classification documentaire en genre et domaine. Le rôle du genre est abordée manière programmatique pour l'indexation.

Mots-clés : recherche d'information, analyse de corpus, genre et domaine, catégorisation morphosyntaxique.

1. Introduction

Il est largement admis que les textes écrits utilisent de façon normée le système graphique, la langue, la structure schématique, l'organisation linéaire des textes. Ces contraintes témoignent d'une codification socio-discursive et culturelle et, il est bien établi que les genres textuels sont révélateurs de ce système de normes (Malrieu et Rastier, 2001). Si la connaissance des genres est indissociable des activités d'écriture, elle conditionne aussi la réception et l'interprétation des textes. De ce point de vue, le genre présente un intérêt pour la recherche d'information.

La recherche d'information spécialisée constitue un terrain d'observation privilégiée des usages liés à la documentation scientifique (Blanquet, 2005). Les textes scientifiques sont actuellement omniprésents sur le web, qu'ils soient diffusés dans le cadre du web invisible, telles les bases de données scientifiques payantes, ou du web visible, les archives numériques en accès libre. Or, les modalités de recherche de l'information se limitent souvent à la recherche par thèmes selon des plans de classification et/ou par mots-clés sur le texte intégral. Les mêmes fonctionnalités de recherche s'appliquent à tous les documents quels que soient la discipline et le genre concernés. La dimension textuelle se trouve de fait évacuée de la caractérisation documentaire et le texte se voit réduit à un sac de mots appartenant ou non à

un langage contrôlé. Or, l'observation des pratiques des chercheurs montrent que ces derniers attachent tout autant d'importance à la pertinence thématique d'une information qu'au type de genre textuel auquel il se rattache. Par exemple, un chercheur ou enseignant-chercheur, va, dans le cadre de son activité, sélectionner des documents qui seront des articles de recherche, des résumés, des procédures, des analyses critiques d'articles, des rapports de recherche, des cours en lignes *etc*. Ces discours relèvent tous d'un genre particulier et l'usage de ces genres semble même relativement codifié.

Nous proposons d'explorer la complexité textuelle de deux collections de documents relevant de genres proches, articles et actes de communication, et de domaines distincts, la linguistique et la mécanique. Nous cherchons à valider l'hypothèse selon laquelle la dimension générique des textes traverse les disciplines. La comparaison des deux corpus annotés sur le plan morphosyntaxique permet, de fait, de faire émerger de grandes régularités. Selon Rastier (1989), la prise en compte des lieux de régulation du genre au niveau du texte, et non au niveau du lexique ou de la phraséologie s'impose afin de rendre compte « *des contraintes globales sur le local* ». Cette perspective nous conduit à considérer le texte comme le lieu de convergence de deux dimensions, lexicale et morphosyntaxique, l'une étant révélatrice du domaine, et l'autre du genre. Ces axes de description de la textualité sont exploités pour la classification automatique en genre et domaine. En guise de perspective, nous évoquerons le rôle du genre pour indexer des documents.

2. Caractérisation morphosyntaxique du genre de l'article : comparaison en mécanique et linguistique

2.1. Contexte de l'étude

L'étude envisagée s'appuie sur un travail mené en collaboration avec C. Poudat dans le cadre de ses travaux de thèse de doctorat sur la caractérisation du genre de l'article scientifique de revues en linguistique. Afin de contraster son corpus de linguistique avec un autre domaine disciplinaire, j'ai constitué un corpus de mécanique et procédé à une analyse morphosyntaxique qui s'appuie sur un jeu d'étiquettes propre au discours scientifique. Cette analyse poursuit un objectif de validation des hypothèses soutenues par Poudat (2006) selon lesquelles le genre de l'article est un niveau de description pertinent des textes scientifiques.

2.2. Corpus et méthodologie

Notre approche repose sur l'analyse d'un corpus de 49 textes extraits des actes du XV^{e} congrès français de mécanique de l'Association Française de Mécanique. Les actes datent de 2001. Ce corpus est contrasté aux 224 textes qui ont été recueillis et traités par C. Poudat. Ce sont des articles scientifiques français de revues linguistiques qui ont été publiés entre 1999 et 2002. La méthodologie adoptée pour le traitement des textes en mécanique est parallèle à celle qui a été définie par l'auteur. L'approche a consisté à annoter les textes au moyen d'un jeu de descripteurs morphosyntaxiques (15 catégories majeures et 145 variables) caractéristiques du genre de l'article. Une fois étiquetés, les textes ont été d'abord été soumis à une analyse statistique descriptive qui permet de faire émerger les constantes du genre de l'article ; ensuite, ils ont fait l'objet d'une analyse en composantes principales (ACP) pour observer les lieux de corrélations des variables du genre.

2.3. Principaux résultats

Parmi les résultats obtenus, nous observons que l'article se distingue nettement d'autres genres de discours (textes juridiques, essais, romans, etc.). Ces constats résultent d'une comparaison du corpus d'étude avec un corpus de référence étiqueté avec le logiciel Cordial®, développé par la société Synapse Développement.

En adoptant un système d'étiquetage propre au discours scientifique, on observe que les textes de linguistique et de mécanique annotés au moyen de TnT treetagger présentent également des points communs : ainsi, parmi les variables sur-représentées dans les deux corpus, on recense les catégories des connecteurs, des symboles et des abréviations, des adjectifs, des numéraux. Parmi les variables sous-représentées, les catégories de marqueurs qui indiquent une relation d'appartenance exclusive avec l'énonciateur sont quasiment proscrites dans ce type de textes : pas de marques de 1ère, 2ème personne aussi bien au niveau des déterminants possessifs, des clitiques. Les temps verbaux absents ou quasi-absents des textes sont sensiblement les mêmes : passé simple, subjonctif imparfait, conditionnel passé, *etc*.

Les deux premiers axes factoriels de l'ACP réalisée sur le corpus de mécanique[i] montre que l'organisation morphosyntaxique des textes se rapproche de celle de l'article de linguistique. Ainsi, on observe une opposition entre deux pôles, celui de la formalisation (présence de numéraux, de symboles, de ponctuations formelles comme le slash ou les parenthèses) et celui d'un mode plus *historico-narratif* (passé simple, passé antérieur, plus-que-parfait, dates). Par ailleurs, on observe que les marqueurs de la rhétorique scientifique (temps du présent, impératif, modaux au conditionnel) sont corrélés aux marqueurs énonciatifs de la personne : *on* est corrélé aux marques de formalisation et *nous* aux marques de modalisation, le *nous* ayant pour fonction de guider le lecteur[ii]. Inversement, le pronom *il* impersonnel est corrélé aux marques de formalisation et données chiffrées, alors que c'était le *nous* en linguistique.

2.4. Analyse

Il convient d'être prudent quant à l'interprétation des résultats, nos observations ne reposant que sur un nombre restreint de domaines. En effet, rien ne permet d'affirmer que le genre de l'article présente des caractéristiques communes dans toutes les disciplines. Cependant, la description des corpus de mécanique et de linguistique permet de confirmer, et dans une certaine mesure, de valider l'existence d'un genre relativement stabilisé dans ces deux domaines qui, pourtant, s'opposent fortement.

En ce qui concerne la pertinence de la morphosyntaxe pour atteindre le palier du genre, nous pouvons, en revanche, être beaucoup plus affirmative. La catégorisation morphosyntaxique permet d'atteindre les dimensions discursives, voire stylistiques des textes, dimensions qui sont constitutives du genre de l'article. C'est en tout cas ce que confirment les travaux de Rinck (2006), qui met en perspective les composantes macro-textuelles du texte avec des aspects micro-linguistiques révélés par la morphosyntaxe (temps verbaux, marques de personnes, ponctuations, etc.) pour analyser la figure de l'auteur et l'identité disciplinaire du genre de l'article.

Le jeu de catégories morphosyntaxiques développé par Poudat (2006) devrait, de notre point de vue, être enrichi de variables morphologiques qui permettraient de rendre compte des modes de construction du lexique de spécialité. Lors de l'analyse, il nous est apparu que la

[i] La carte factorielle est consultable sur (Poudat, 2006 : 310)
[ii] Remarque : le rôle des pronoms est inversé en linguistique.

plupart des noms et adjectifs étaient des mots dérivés et/ou composés extrêmement réguliers[iii], ce qui permettrait d'envisager un profilage purement morphologique des textes.

3. Applications du genre à la recherche d'information

3.1. Genre et classification de textes

La classification automatique de textes est l'une des applications les plus courantes du genre pour la recherche d'information. Pour classer des textes, on peut en effet s'appuyer sur des approches statistiques et des méthodes de TAL pour recueillir des corrélats de marqueurs issus de la textualité qui se combinent aux marqueurs documentaires et éditoriaux. Différents types de marqueurs caractéristiques du genre ont été relevés dans la littérature. En général, il y a une combinaison de traits structuraux : catégories grammaticales, ponctuation, longueur des phrases, complexité syntaxique, *etc*. (Kessler et *al*., 1997), (Wolters et Kirsten, 1999) et (Lee et Myaeng, 2002). La disponibilité de documents numériques sur internet a récemment contribué à remettre le genre sur le devant de la scène (Prime-Claverie et *al.,* 2002). C'est ici la portée typologisante du genre qui est « utilisée » à des fins classificatoires, d'autres travaux plus anciens en avaient déjà posé les principes (Kalrgren et Cutting*,* 1994).

Le travail de description menée en corpus a été le point de départ de différents tests pour classer des textes automatiquement en genre et domaine (Poudat et *al.* 2006). Le choix de la méthode et la mise en œuvre effective de la classification sont le fruit du travail de G. Cleuziou, qui, dans le cadre de sa thèse de doctorat s'est intéressé au classement de données textuelles pour la recherche d'information (Cleuziou, 2004). Parmi les résultats intéressants, il apparaît que la prise en compte conjointe de descripteurs lexicaux et morphosyntaxiques[iv] améliore le classement des textes en genre, et plus surprenant, le classement thématique.

Ces résultats sont encourageants même s'ils ne portent que sur deux domaines (linguistique et mécanique) et trois genres scientifiques (articles, présentations de revues, comptes rendus). Ils laissent imaginer que la classification au sein du discours de la recherche, c'est-à-dire avec des textes relevant de genres relativement proches, est possible. Il faut remarquer que lorsque le corpus comporte des genres très hétérogènes et peu stabilisés, les performances des classifieurs se dégradent ainsi que le font remarquer Wolters et Kirsten (2002) sur le corpus LIMAS, l'équivalent du corpus Brown pour l'allemand.

3.2. Genre et indexation

L'indexation a pour but de représenter le contenu d'un document à l'aide d'un vocabulaire contrôlé ou non. Ce vocabulaire peut consister en l'extraction de la terminologie d'un domaine. La prise en compte de la textualité peut permettre d'éclairer le sens d'un terme en fonction de sa position dans le texte. Ainsi, C. Poudat a montré que certains termes, lorsqu'ils apparaissent en début de texte, sont problématisés, alors qu'à la fin, ils sont définis. Les mêmes constats peuvent être faits sur le corpus de mécanique. Il est donc important de considérer que le genre a aussi un impact sur la caractérisation du lexique de spécialité.

En revanche, ce qui semble plus problématique, et qui constitue l'un des point de divergence les plus importants entre les deux corpus, est l'interprétation linguistique (ou rhétorique) que

[iii] *atromatique, densitométrie, poroélastique, piézorésistance/piézoréistif, thermoacoustique vibroacoustique,etc.*
[iv] 136 variables morphosyntaxiques caractéristiques du discours scientifique

l'on peut attribuer à cette structure. Si le « fameux » plan IMRaD[v] s'applique à quasiment tous les articles de mécanique, il ne s'applique pas à la linguistique. Il semble qu'il y ait là une distance certaine affichée par les linguistes vis-à-vis des standards imposés par l'écriture scientifique. Doit-on de ce point de vue, distinguer les genres selon leur appartenance discursive, c'est-à-dire, pour dire vite, les communautés de discours littéraires des sciences de l'homme *versus* les communautés de discours des sciences de la nature ? Si oui, existe-t-il des régularités de quelque ordre que ce soit et, sont-elles automatisables ?

Références

BLANQUET M.-F. (2005). « Quelle recherche d'information pour une discipline donnée ? », *in Babel - edit,* Rencontres Formist. ENSSIB - juin 2005. [en ligne] http://babel.enssib.fr/

CLEUZIOU G., (2004). *Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information.* Thèse de doctorat en informatique soutenue en décembre 2004 à l'Université d'Orléans.

KARLGREN J. and CUTTING D. (1994). « Recognizing text genres with simple metrics using discriminant analysis », *in* Proceedings of COLING 94, Kyoto.

KESSLER B., NUNBERG G. and SCHÜTZE H. (1997). « Automatic Detection of Genre » *in* Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'97*),* p. 32-38.

LEE Y.-B. and MYAENG S.-H. (2002). « Text genre classification with genre-revealing and subject-revealing features » *in* SIGIR'02, Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, pp. 145-150.

MALRIEU D. et RASTIER F. (2001) « Genres et variations morphosyntaxiques » *in* TAL, Vol. 42, n°2, pp. 547-577.

MOUNIER E. et PAGANELLI C. (2003). « La segmentation du texte en paragraphes : une application à la recherche d'information dans les documents techniques volumineux ». *in* Modèles linguistiques. Tome XXIV, Fascicule 2, pp. 85-97.

POUDAT C. (2006). *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*. Thèse de doctorat en linguistique soutenue en juin 2006 à l'Université d'Orléans. [en ligne] http://www.revue-texto.net/Corpus/Corpus.html

POUDAT C., CLEUZIOU G. et CLAVIER V. (2006). « Catégorisation de textes en domaines et genres : complémentarité des indexations lexicale et morphosyntaxique » *in* Document numérique vol. 9, n°1, Paris : Hermès, Editions Lavoisier, pp. 25-42.

PRIME-CLAVERIE C., BEIGBEDER M. and LAFOUGE T. (2002). « Clusterisation du web en vue d'extraction de corpus homogènes », INFORSID 2002, 20e congrès informatique des organisations et des systèmes d'information et de décision, Nantes, 4-7 juin 2002.

RASTIER F. (1989). *Sens et textualité*. Paris : Hachette.

RINCK F., (2006). *L'article de recherche en Sciences du langage et en Lettres : figure de l'auteur et identité disciplinaire du genre.* Thèse de doctorat en linguistique soutenue en novembre 2006 à l'Université Stendhal, Grenoble III.

WOLTERS M. and KIRSTEN M. (1999). « Exploring the Use of Linguistic Features in Domain and Genre Classification » *in* Ninth Conference of the European Chapter of the Association for Computational Linguistics, pp. 142-149.

[v] Introduction, Methods, Results and Discussion.