Introduction et plan de la thèse

La présente thèse est née d'une interrogation sur le texte, et plus spécifiquement, sur ce qui permet, à ce niveau de constitution, de le différencier et de l'identifier formellement sur le plan linguistique. Deux types d'entreprises ont à l'origine orienté et nourri notre réflexion : les travaux de Denise Malrieu et François Rastier sur les variations morphosyntaxiques des genres textuels (Malrieu & Rastier 2001) et les travaux de Benoît Habert sur le *profilage de textes* (Habert 2000a, 2000b), tous deux intéressés ou influencés par les travaux pionniers de D. Biber (1988).

Si les deux derniers travaux cités s'intéressent peu ou prou à la notion de *genre*, c'est bien cet objet qui nous a finalement interpellée, tant sur le plan théorique que méthodologique :

- *théorique* car la linguistique est de plus en plus confrontée à la question de l'attestation de ses corpus d'étude, et du bien-fondé de ses descriptions, aux exigences de représentativité et de congruence : objet empirique, le texte constitue un palier de description pertinent qui ne peut être défini que relativement aux contraintes globales qui régulent les phénomènes locaux qui l'affectent ; parmi ces contraintes, le genre apparaît comme crucial si l'on admet avec Rastier qu'il représente un palier de normalisation linguistique permettant de relier les textes aux discours, entités plus larges et donc plus complexes à caractériser ;
- *méthodologique*, car une telle entreprise relève un défi réel : le genre n'est pas immédiatement observable, les textes qu'il affilie n'en étant que des réalisations exemplaires.

En partant du principe que le global détermine le local et que tout texte est par conséquent linguistiquement normé par son genre, et donc son corpus d'appartenance (Rastier, 2001), nous avons tenté de saisir ces régulations en exploitant les moyens dont dispose aujourd'hui la linguistique.

Sans être dénuée d'hypothèses sous-jacentes, notre étude est fondamentalement empirique : c'est en corpus que nos observations sont systématiquement objectivées, et que se sont élaborés nos choix méthodologiques et nos descriptions.

La perspective adoptée est ainsi empirique et utilitariste, dans la mesure où nous avons globalement cherché à dépasser l'opposition *linguistique théorique / appliquée* en nous orientant vers une linguistique applicable et objectivée, qui s'opposerait à une linguistique d'opinion. Notre projet vise d'abord à construire des observables indépendamment de l'intuition, qui pourra ainsi être validée ou invalidée, mais toujours dans un second temps de l'analyse. La présente thèse s'inscrit ainsi dans une certaine mesure dans un cadre poppérien.

Les genres étant innombrables et en constante évolution, c'est finalement sur l'article scientifique que s'est porté notre choix. En effet et comme l'expose le chapitre 1, les genres scientifiques/académiques et professionnels sont peu décrits en français et demeurent encore appréhendés de manière très locale et incomplète par l'ensemble du courant rhétorico-fonctionnel appelé *English for Specific Purpose*. Dans la mesure où il représente le genre le plus accrédité, le plus répandu et le plus observé du discours scientifique si l'on s'intéresse à l'écrit, l'article de revue s'est naturellement imposé.

Le genre de l'article semblant varier considérablement d'un domaine, ou d'une discipline de spécialisation à l'autre[1], c'est au sein du domaine scientifique linguistique qu'il a été décrit : la décision peut certes paraître singulière, voire réflexive, d'autant que la linguistique a déjà tendance à fonder ses descriptions sur ses propres productions, mais l'adoption d'un domaine inconnu de l'analyste entraîne des problèmes d'expertise et de (mé)connaissance des standards régissant le genre, qui entravent considérablement l'interprétation des données. De surcroît, l'observation des productions scientifiques du champ linguistique est intéressante d'un point de vue auto-réflexif, et pourrait donner lieu à diverses applications didactiques : les pratiques de rédaction scientifique en français sont encore peu répandues dans les cursus.

Outre son intérêt applicatif et descriptif, la présente thèse propose une réflexion méthodologique sur l'élaboration et la mise en œuvre d'un observatoire de genre : après avoir collecté et construit un corpus de 224 articles de revues linguistiques parus autour de 2000 – puisqu'un genre s'observe d'abord en synchronie -, nous avons mis en place une chaîne de traitement exploitant les outils et les méthodes du Traitement Automatique des Langues (TAL), des statistiques textuelles et de la linguistique de corpus en général.

La démarche est globalement la suivante : après avoir décrit le corpus génériquement homogène (ou du moins faiblement hétérogène), on créée des contrastes en explorant la hiérarchie des variables typologiques possibles (*e.g.* auteur, domaine, genre). On opère donc par cycles de validation, en s'écartant du corpus initial par le biais d'hypothèses contrastives pour finalement y revenir, et le saisir de manière plus pertinente. En ce sens, l'approche adoptée est bien distincte du profilage, qui s'intéresse aux affinités plus qu'aux contrastes, même s'il constate des différences.

L'une des originalités de l'entreprise réside probablement dans l'adaptation des descripteurs aux caractéristiques des textes scientifiques, afin de permettre, voire de garantir l'interprétation des données : étant donné son caractère discriminant en matière de typologies textuelles (*e.g.* Karlgren 1994, Karlgren et Cutting 1994, Kessler et al. 1997, Rayson et Garside 2000, Habert 2000, Malrieu et Rastier 2002, etc.) et son apport considérable à la description des langues, c'est le niveau morphosyntaxique que nous avons d'emblée privilégié, d'autant que de nombreux outils d'étiquetage sont disponibles.

Après avoir recouru un temps à Cordial et rencontré maintes difficultés d'interprétation et de fiabilité des données et des descripteurs, nous avons exploré d'autres voies, et mené un travail de recension et d'exploitation des étiqueteurs morphosyntaxiques à AKSIS (Bergen, Norvège) dans le cadre d'une bourse Marie Curie. Suite à ce projet, nous avons pris la décision d'entraîner plusieurs annotateurs morphosyntaxiques à l'étiquetage des textes scientifiques, à partir d'un système de descripteurs original adapté aux caractéristiques de l'article scientifique de revue linguistique. Malgré son coût élevé, cette solution s'est avérée extrêmement profitable à l'étude, d'une part parce qu'elle nous a permis de saisir des phénomènes linguistiques que nous n'aurions pas pu appréhender par d'autres biais, et d'autre part parce que la pertinence de la description du genre observé s'en est trouvée considérablement accrue.

[1] Comme l'a montré par exemple le projet KIAP (Cultural Identity in Academic Prose) http://kiap.aksis.uib.no/index-e.htm.

Les constituants de l'observatoire de genre mis en place (corpus d'étude et systèmes de descripteurs) sont présentés dans le second chapitre de ce travail.

Puisque toute description nécessite un point d'entrée, c'est ce niveau morphosyntaxique, qui a fondé nos investigations. Le chapitre 3 dresse ainsi un profil morphosyntaxique du genre de l'article, les descripteurs étant d'abord appréhendés plus localement par sous-systèmes de description (personnes, temps verbaux, ponctuations, connecteurs, etc.), avant de servir d'entrée à une analyse multidimensionnelle, visant à explorer la structure générique de l'article.

Etant donné le nombre de dimensions prises en compte, on imaginera facilement à quel point nous avons été submergée de données – et cette observation vaut pour l'ensemble de notre entreprise et des travaux quantitatifs, nécessairement confrontés à la richesse empirique ; puisqu'il est exclu de tout restituer et que le passage du quantitatif au qualitatif doit être raisonné, nous avons souvent dû restreindre les résultats obtenus en différenciant, et en privilégiant ce que nous étions en mesure d'interpréter.

Les données sont peut-être inégalement interprétées, et les spécialistes de phénomènes plus locaux pourraient à juste titre invalider certaines analyses. Cependant, notre projet aura eu le mérite d'affronter la richesse et la complexité de l'objet : les éléments peu ou mal interprétés pourront être revisités par les scoliastes futurs. Dans cette recherche, ce sont les relations établies entre les données plus que les données en elles-mêmes qui importent.

Afin de saisir le genre dans sa complexité textuelle, différents axes d'organisation ont été ensuite explorés : les textes scientifiques étant particulièrement structurés, on s'est ainsi intéressé à la structuration et à l'organisation du genre de l'article, non soumis à une structure IMRAD[2] en français (chapitre 4).

L'article étant constitué de composantes optionnelles, ou de configurations optatives nombreuses et linguistiquement distinctes de son corps, c'est l'exemple de linguistique que nous avons particulièrement approfondi (chapitre 5). Outre les variations morphosyntaxiques importantes du genre selon son degré d'exemplification, l'exemple s'est avéré représenter en moyenne plus de 10% de l'article, ce qui est important.

Le plan de l'expression ayant été jusqu'ici privilégié sur le plan du contenu, c'est la thématique du genre qui a ensuite été explorée, dans le cadre d'un chapitre visant à décrire et à discriminer les concepts linguistiques et leur textualisation à l'aune de différents critères (chapitre 6).

Si les chapitres 3, 4, 5 et 6 explorent les lieux de stabilité et de variation de l'article en corpus génériquement homogène, les chapitres 7, 8 et 9 posent différentes hypothèses de variation externes qui dépassent le cadre du corpus initial et requièrent sa confrontation à d'autres collections : incidence du style d'auteur sur le genre (chapitre 7), confrontation du genre de l'article à d'autres genres scientifiques linguistiques et variations du genre d'un domaine à l'autre (chapitre 8) et d'une langue à l'autre (chapitre 9).

[2] *Introduction, Materials and methods, Results, Analysis, Discussion*.

Le programme paraît d'emblée ambitieux, mais on soulignera qu'il est essentiellement exploratoire, en grande partie parce que s'est posé le problème de la disponibilité des corpus de référence et de comparaison. Peu d'entreprises poursuivent en effet des objectifs similaires aux nôtres, avec des données comparables, *i.e.* génériquement homogènes et étiquetées. Par conséquent, la mise en contraste de notre corpus avec d'autres collections, afin de déterminer ce qui précisément le caractérise, s'est avérée problématique, et a affecté l'ensemble de la thèse : par exemple, les propriétés morphosyntaxiques *spécifiques* au genre de l'article (chapitre 3) n'ont pu être déterminées que grossièrement, en les comparant à des discours éloignés.

Les textes du chapitre 7, qui résulte dans son ensemble d'un travail collaboratif avec F. Rinck (Lidilem, Grenoble), ont été l'objet d'une nouvelle récolte, mise en place dans cet objectif ; les deux nouveaux corpus homogènes en genre (présentations de revue et comptes rendus) sont constitués de textes tirés des revues linguistiques que l'on a bien voulu nous confier, tandis que le corpus « Mécanique » appartient à V. Clavier (Gresec, Grenoble) avec qui nous avons travaillé. Le corpus anglais a quant à lui été constitué parallèlement au corpus français.

L'ensemble des corpus mobilisés a été l'objet d'un lourd traitement d'étiquetage (et de vérification) des textes avec le système d'annotation spécifique présenté chapitre 2 : la procédure étant coûteuse, l'examen des variations genres / domaines n'aurait par exemple pu être étendue davantage. Le choix de la « Mécanique », qui peut paraître pour le moins singulier, s'est d'ailleurs imposé de fait, et non de droit, en fonction de la disponibilité des textes.

Le bilan et les orientations et analyses futures de notre travail sont présentés au sein du chapitre 10.

On notera que nous n'avons pas regroupé l'ensemble des méthodes et des outils employés au sein d'un chapitre méthodologique, choix qui nous semble justifié du fait que les chapitres du présent travail ne recourent pas systématiquement aux mêmes procédés ; les méthodes mobilisées par les objectifs internes de chaque chapitre sont donc présentées en amont.