Caractériser les discours académiques et de vulgarisation : quelles propriétés ?

Amalia Todirascu[1], Beatriz Sánchez-Cárdenas[2]
(1) LiLPa, Université de Strasbourg, 22, Rue René Descartes, BP 80010, 67084 STRASBOURG cedex, France
(2) LexiCon, Universidad de Grananda, Calle Buensuceso, 11 18002 Granada, Espagne
todiras@unistra.fr, bsc@ugr.es

Résumé. L'article présente une étude des propriétés linguistiques (lexicales, morpho-syntaxiques, syntaxiques) permettant la classification automatique de documents selon leur genre (articles scientifiques et articles de vulgarisation), dans deux domaines différentes (médecine et informatique). Notre analyse, effectuée sur des corpus comparables en genre et en thèmes disponibles en français, permet de valider certaines propriétés identifiées dans la littérature comme caractéristiques des discours académiques ou de vulgarisation scientifique. Les premières expériences de classification évaluent l'influence de ces propriétés pour l'identification automatique du genre pour le cas spécifique des textes scientifiques ou de vulgarisation.

Abstract.

Characterizing scientific genre for academia and general public: which properties should be taken into account?
The article focuses on the study of a set of morpho-syntactic properties for audience-based classification. The linguistic analysis of academic discourse and of popular science discourse reveals that both discourse types are characterized by specific linguistic and textual properties. This research used two French comparable corpora in regards to genre and subject matter. The corpora was composed of scientific articles and popular science texts in the domains of medicine and computer science. The experiments performed as part of our study evaluated the influence of discourse-specific morpho-syntactic properties on genre-based classification, for scientific and popular science texts.

Mots-clés : analyse linguistique, discours scientifique et de vulgarisation, corpus comparables, classification selon le genre.
Keywords: linguistic analysis, academic and popular science discourse, comparable corpora, genre-based classification.

1 Introduction

L'identification automatique du genre du document est une tâche utile pour l'extraction automatique de terminologie ou de néologismes, pour la génération automatique de contenu destiné à un public cible ou pour la simplification automatique. En effet, plusieurs applications peuvent intégrer cette étape de classification par genre. Pour l'extraction de termes ou néologismes, il faut utiliser des textes riches en termes ou productifs en néonymie, tels que les articles scientifiques. Dans le domaine de la didactique des langues, il est nécessaire de proposer aux apprenants des textes adaptés à leur niveau de connaissances, c'est-à-dire des textes simplifiés en terme de lexique, de structures syntaxiques ou discursives. Pour cette raison, nous privilégions la sélection des textes de vulgarisation pour des apprenants ayant des compétences en langues. Etant donné que le genre est caractérisé par un ensemble complexe de paramètres linguistiques et extra-linguistiques, l'identification automatique des genres, en particulier des discours académiques et scientifiques n'est pas une tâche aisée.

Lors de l'identification automatique du genre, la plupart des systèmes proposent une identification basée sur des n-grammes (séquences de mots, de caractères ou de catégories lexicales) (Sebastiani, 2005) (Lee et Myaeng, 2004) (Kessler *et al*, 1997) exploitant le lien entre le thème et le genre traité (Bechet *et al*, 2008). Or, ces approches sont généralement liées au vocabulaire du domaine et, si l'on obtient des performances notables dans un domaine, l'adaptation des systèmes pour d'autres domaines émergents demande une phase de réapprentissage sur un corpus conséquent. Pourtant, quelques travaux se distinguent par l'exploitation des catégories lexicales (Karlgren et Cutting, 1994) ou des propriétés syntaxiques (Santini, 2007), (Goeuriot *et al*, 2005), (Stamatatos *et al*, 2000) (D'hondt *et al*, 2013) pour classer les documents selon le genre ou l'auteur des textes. Certaines approches proposent des propriétés

indépendantes de la langue (Petrenz et Webber, 2011), (Sun *et al*, 2012) qui s'avèrent efficaces pour la classification de genres journalistiques. Enfin, plusieurs travaux combinent propriétés stylométriques (longueur de phrases, signes de ponctuation), catégories lexicales et mots-clés pour la classification des genres littéraires (D'hondt *et al*, 2013), (Lecluze et Lejeune, 2014).

Des nombreux travaux en linguistique proposent des caractérisations des types et des genres textuels par le biais des propriétés morpho-syntaxiques (Biber et Conrad, 2009) (Poudat *et al*, 2006), (Malrieu et Rastier, 2001). Les résultats de ces études ont été exploités par des systèmes de classification automatique par genre (Charnois *et al*, 2008, Poudat, 2008, D'hondt *et al*, 2013). Ces systèmes utilisent des propriétés linguistiques (p.e. la fréquence ou la distribution de noms propres ou de déterminants, la fréquence de certaines catégories de verbes de modalité ou d'opinion, la fréquence des groupes nominaux complexes etc.), mais pas d'annotations linguistiques de plus haut niveau (syntaxe, sémantique discursive).

Nous adoptons une approche de classification automatique exploitant les études des genres textuels disponibles dans la littérature. Dans cet article, nous étudions plusieurs propriétés lexicales (termes du domaine), traits morpho-syntaxiques et structures syntaxiques utiles pour l'identification automatique du genre textuel selon le public visé : communauté d'experts scientifiques ou grand public. Contrairement à d'autres approches qui sélectionnent les propriétés automatiquement, nous partons des travaux en linguistique textuelle (Biber et Conrad, 2009) et des études des genres scientifiques (Hyland, 2009 ; Swales, 2004) et nous comparons les propriétés textuelles des articles scientifiques et de textes de vulgarisation scientifique sur deux corpus comparables[1] en termes de genres et du thème, disponible en français, un corpus médical et un corpus informatique. Les corpus sont comparables en genre : chaque corpus est constitué en parties égales d'articles scientifiques et d'articles de vulgarisation. D'autre part, à l'intérieur du domaine, nous avons sélectionné des textes scientifiques et de vulgarisation traitant des mêmes thèmes. Pour vérifier l'influence des propriétés sélectionnées suite à l'étape de l'analyse de corpus, nous utilisons des techniques de classification automatique et nous comparons avec une approche de classification basée exclusivement sur des termes du domaine ou les mots pleins contenus dans les documents.

2 Discours scientifiques et discours de vulgarisation scientifique

Pour analyser les différences entre discours scientifiques et de vulgarisation, nous adoptons la définition proposée par (Biber et Conrad, 2009), considérant que le genre est défini par un faisceau de propriétés linguistiques et extra-linguistiques (paramètres de production du texte et de réception du texte). Certaines structures et formules sont unanimement reconnues par les utilisateurs comme étant des caractéristiques propres à un genre donné. À ce titre, (Hyland, 2009) propose une classification des genres académiques, du point de vue des pratiques de rédaction et de lecture que la communauté scientifique adopte. Le domaine définit les caractéristiques rhétoriques et stylistiques du discours académique. (Swales, 2004) considère que les genres participent activement à la construction des connaissances d'un champ disciplinaire. Les pratiques disciplinaires se traduisent par la préférence pour certains procédés linguistiques, destinés à faciliter le partage des connaissances. Pour notre étude, nous nous focalisons sur les propriétés linguistiques et extra-linguistiques permettant de distinguer entre les discours académiques et de vulgarisation scientifique.

Des nombreux travaux existent sur la caractérisation des articles scientifiques (Hyland, 2009), (Swales, 2004). Ces recherches ont mis en évidence des procédés spécifiques, reconnus et appliqués par la communauté scientifique, pour exprimer le positionnement de l'auteur (Tutin, 2010) ou pour argumenter des choix méthodologiques (Rinck, 2006). Parmi les procédés linguistiques mentionnés par (Hyland, 2009), plusieurs sont typiques du discours académique : des expressions qui expriment la possibilité d'une hypothèse (verbes de modalité), l'autocitation (pronoms personnels de 1ère personne, articles possessifs), l'expression d'un point de vue (verbes de croyance). (Swales, 2004) propose une analyse détaillée de plusieurs catégories de genres scientifiques (thèse, article, présentation orale) en identifiant les fonctions rhétoriques et linguistiques qui expriment l'argumentation, la citation d'autres travaux ou les connaissances implicites.

Si certains travaux proposent l'identification d'un vocabulaire commun aux textes scientifiques, un meta-langage (Drouin, 2007), d'autres mettent en exergue l'apparition de phénomènes liés à la syntaxe : présence d'adjectifs relationnels (Daille, 1999), préférence pour le passif, utilisation de tournures impersonnelles, utilisation du pronom de 1ère personne. (Kocourek, 1991) souligne comme caractéristique du langage de spécialité les termes du domaine. Aussi, il identifie des phénomènes telles que les propositions participiales ou infinitives, les subordonnées relatives.

[1] Si le terme « corpus comparable » est généralement utilisé pour des corpus multilingue qui partagent les mêmes critères, dans cet article, nous utilisons le terme de corpus comparable pour des corpus qui ont des sous-parties comparables en terme de genre et thème.

D'autre part, le langage de vulgarisation scientifique a également fait l'objet de nombreuses études. (Hyland, 2009) identifie une volonté d'expliciter les notions au niveau de connaissances du public non averti : utilisation des explications, des dispositifs cohésifs tels que la répétition, la synonymie ou les articles démonstratifs pour renforcer la cohésion du texte. Le public auquel l'auteur s'adresse est inclus dans le discours, par l'utilisation du pronom personnel *vous* et par des questions rhétoriques accompagnées de réponses. (Jacobi et Schiele, 1988) proposent plusieurs paramètres linguistiques spécifiques aux textes de vulgarisation telle la reformulation (l'explication d'un terme pivot par d'autres expressions plus simples) pour renforcer les connaissances du public. La reformulation s'exprime par des marqueurs explicites (*« c'est-à-dire », « autrement dit »*) ou par des énoncés définitoires (*X est défini comme Y, X est nommé Y, X est un Y*).

Ces propriétés mises en évidence par des études détaillées des discours académiques ou de vulgarisation scientifiques seront utilisées pour classifier automatiquement les deux catégories de discours. Nous évaluons l'influence des ces propriétés décrites dans la littérature dans le cadre d'un système de classification automatique de genres, par une étude de corpus et par quelques expériences de classification.

3 Méthodologie

Nous partons des ces études linguistiques des discours académique et de vulgarisation scientifique pour identifier des propriétés exploitables pour la classification automatique. Pour atteindre cet objectif, nous avons suivi la méthodologie explicitée ci-dessous :

1. Identification des propriétés morpho-syntaxiques, syntaxiques et stylistiques dans la littérature pour l'analyse des discours académiques et de vulgarisation (Hyland, 2009; Swales, 2004; Jacobi et Schiele, 1988) ;
2. Constitution de corpus comparables pour deux domaines différents, composés d'articles scientifiques et de textes de vulgarisation. Les domaines choisis sont le domaine médical et l'informatique, deux domaines dans lesquelles les pratiques d'écriture scientifique sont différentes ;
3. Prétraitement : étiquetage, lemmatisation et analyse syntaxique (Bohnet, 2009) ;
4. Analyse des corpus, à l'aide du concordancier Antconc (Anthony, 2009), appliqué aux corpus annotés, pour l'identification des propriétés utiles pour la classification ;
5. Développement d'un extracteur de propriétés, appliqué sur le corpus annoté ;
6. Expériences de classification avec les propriétés choisies avec la plateforme Weka (Hall *et al*, 2009). Nous avons comparé les résultats obtenus à l'aide des propriétés sélectionnées avec un système utilisant des mots pleins identifiés dans les corpus.

Dans les prochaines sous-sections, nous présentons les étapes de création de corpus et le prétraitement appliqué sur le corpus, l'analyse des propriétés présentées dans la section 2 et les expériences de classification.

3.1 Création de corpus et prétraitement

Notre objectif est d'identifier automatiquement les articles scientifiques et les textes de vulgarisation dans les domaines de la médecine et de l'informatique. Compte tenu du fait que peu de corpus sont disponibles en français (Tutin 2010) nous avons constitué des corpus comparables dans les domaines mentionnés. Les corpus sont comparables par rapport aux genres sélectionnés (textes de vulgarisation, articles scientifiques) et par rapport aux thèmes traités dans les documents. Nous avons cherché les documents à l'aide des mêmes mots-clés (à l'intérieur du domaine) pour sélectionner des documents de genres différents traitant du même thème.

Pour le domaine médical, nous disposons de deux corpus de taille comparable pour le français : 302 textes/genres (environ 1 500 000 mots pour chaque genre). Les articles scientifiques proviennent du corpus Scientext (Tutin, 2010), des sites à destination de spécialistes, gérés par les réseaux de santé et les organismes publics, de quelques revues médicales (la revue « médecine/sciences », Revue française de Rhumatologie). Le corpus du discours de vulgarisation dans le domaine médical été constitué à partir de sites d'information à destination du grand public, créés pour la prévention sur certaines maladies[2]. Le corpus de textes scientifiques du domaine informatique est composé d'articles scientifiques disponibles sur le portail HAL. Le corpus de textes de vulgarisation en informatique est composé des textes disponibles sur des portails ou des magazines en ligne et des tutoriels destinés à l'apprentissage d'un langage de

[2] Nous remercions Guillaume Bertrand qui a contribué à la constitution du corpus médical dans le cadre de son travail de mémoire du master Linguistique, Informatique, Traduction (Université de Strasbourg, 2011).

programmation pour débutants. Certains textes de vulgarisation ont été retirés du corpus car jugés trop éloignés des textes de vulgarisation (par exemple des brèves annonçant le lancement d'un nouveau logiciel ou produit). Nous disposons au total de 301 textes/genre dans le corpus informatique.

Tout d'abord, nous avons procédé à un nettoyage des corpus recueillis. Pour le corpus informatique il s'agit de supprimer manuellement les images, les tableaux et les formules, ainsi que les parties contenant du code. Les corpus médicaux provenant des sites Web ont été extraits à l'aide d'un outil d'extraction du contenu textuel à partir des pages Web et des fichiers PDF (Todirascu *et al*, 2012). Les corpus ont été étiquetés, lemmatisés et annotés avec l'analyseur statistique en dépendances de (Bohnet, 2009) disponible pour le français. Nous avons utilisé cet analyseur dans le but d'identifier plusieurs catégories de propriétés syntaxiques des discours scientifiques (Biber et Conrad, 2009; Hyland, 2009). Par ce biais, nous cherchons à identifier des propriétés généralisables pour la classification de textes scientifiques et de vulgarisation entre plusieurs domaines.

3.2 Analyse de corpus

Nous avons réparti les propriétés identifiées dans la littérature dans plusieurs classes :

- propriétés statistiques : la longueur moyenne des phrases, le nombre total d'unités lexicales, la longueur moyenne des mots, la fréquence des mots longs ou courts, les signes de ponctuation (!,?). Il est possible de calculer ces propriétés sans faire appel aux annotations linguistiques ;
- propriétés lexicales : certaines classes de verbes (verbes de cognition, verbes de communication, verbes de modalité) ou d'adjectifs (relationnels). De plus, nous avons utilisé une liste de 100 termes monolexicaux et polylexicaux, extraits à l'aide de Termostat (Drouin, 2007) pour chaque corpus ;
- propriétés morpho-syntaxiques : les catégories lexicales spécifiques (nom, nom propre, verbe, adjectif, adverbe, pronoms personnels de 1re et 2e personne) ;
- propriétés syntaxiques et sémantiques : les séquences définitoires ou des explications, constructions passives, les tournures impersonnelles, le type de sujet ou d'objet (pronoms, groupes nominaux ou phrase).

Afin d'évaluer le lien entre ces propriétés et leur genre, nous avons étudié les corpus, étiquetés, lemmatisés et analysés en dépendances. Compte tenu des tailles variables des documents composant nos corpus, nous avons calculé la fréquence relative de chaque propriété (FT – le nombre d'occurrences comptées dans le corpus, Nb – le nombre total de mots et de signes de ponctuation du corpus) :

$$Freqrel = (\frac{FT}{Nb}) * 1000000$$

Pour compter la fréquence des propriétés, nous avons défini des patrons lexico-syntaxiques spécifiques, utilisant le langage d'interrogation de corpus CorpusQueryProcessing (CQP) (Christ, 1994). Pour identifier les sujets et les objets complexes, nous avons exploité les liens de dépendances entre le verbe et le sujet. Nous avons défini des règles heuristiques pour plusieurs phénomènes (tournures impersonnelles, énoncés définitoires). Certaines règles estiment la fréquence relative du phénomène (propositions relatives). Plusieurs patrons identifient des définitions (11 patrons) ou des emplois impersonnels du pronom 'il' (45 patrons) :

[lemma= "'être"] [lemma="définir|appeler|nommer"][word="comme"][pos="NOM"]

[lemma="il"][lemma="être"][word="nécessaire"][word="de"][pos="vinf"]

lemma – impose une contrainte sur le lemme ; word – sur la forme ; pos – sur la categorie lexicale (vinf – verbe à l'infinitif, NOM – nom).

Nos analyses (tableau 1) montrent que le discours académique est marqué par plusieurs propriétés caractéristiques : les pronoms personnels *(nous, je)*, les constructions passives, les pronoms impersonnels (*il, on*) et la préférence pour des sujets complexes (phrases subordonnées, groupes nominaux modifiés par plusieurs compléments de noms), pour des mots longs (plus de 9 caractères) ou courts. La fréquence relative de ces propriétés est plus importante pour les discours académiques que dans les corpus de vulgarisation scientifique. D'autre part, le discours de vulgarisation est caractérisé par une préférence marquée pour le pronom personnel de la 2e personne, une forte présence des questions, une préférence pour les marqueurs de reformulation, ainsi que pour les définitions qui éclaircissent les termes complexes au grand public. Nous avons retenu ces propriétés ayant un comportement similaire dans les deux domaines.

	Je	Nous	passif	définitions	Sujet complexe	Pronoms impersonnel	Reform.	Pronom 2e pers
DS MED	22	2316	3069	814	41536	326	214	0
DV MED	934	274	2546	2212	10451	14	132	315
DS INFO	188	2049	657	1704	15789	187	113	10
DV INFO	0	20	167	11070	4324	20	530	641

TABLE 1 : Quelques propriétés lexicales, morpho-syntaxiques ou syntaxiques et leur fréquence relative (valeurs en partie pour million) (DS – discours académique; DV -discours de vulgarisation)

Certaines propriétés n'ont pas été retenues pour les expériences de classification en raison de leur comportement similaire dans les deux genres (par exemple, la fréquence de noms propres, de noms ou d'adverbes qui ont des fréquences similaires dans les deux genres). D'autres propriétés ont des comportements contradictoires selon le domaine : le pronom personnel *je* est plus fréquent dans le discours de vulgarisation scientifique en médecine, alors qu'en informatique il est spécifique au discours scientifique. Les marqueurs de reformulation semblent plus fréquents dans le discours scientifique médical tandis ce qu'en informatique, ils sont plus fréquents dans les discours de vulgarisation.

Finalement, nous avons comparé le vocabulaire des textes scientifiques et celui de textes de vulgarisation. Nous avons remarqué la présence des termes du domaine dans le corpus de textes scientifiques et le corpus de textes de vulgarisation. Cependant, il existe des différences entre les deux types de discours. D'une part, le langage scientifique se caractérise par la préférence marquée pour des noms abstraits (*analyse, approche, processus, entité*) qui font partie du méta-lexique du domaine informatique et par une préférence marquée pour les noms d'événement liés aux processus d'examen médical et de prise en charge du patient (*examen, étude, traitement*), aux processus des maladies (*infection, apparition, augmentation*). D'autre part, le langage de vulgarisation se manifeste par la fréquence de mots du domaine de la biologie *(cellule, protéine, neurone)* et des parties du corps *(abdomen, foie, ganglion)* pour le corpus médical et par une fréquence des entités propres à l'univers de l'ordinateur (*souris, fenêtre*) ou des programmes (*code, fonction*). Ces classes de noms abstraits ont été rajoutées aux propriétés utilisées pour les expériences de classification.

3.3 Expériences de classification

Après avoir choisi les propriétés à l'issue de l'analyse de corpus, nous avons appliqué ces propriétés pour la classification automatique, afin d'évaluer leur utilité pour l'identification de genres. Les propriétés sélectionnées sont regroupées manuellement et sont utilisées pour représenter les documents :

- TERMES : un ensemble de 100 termes monolexicaux et polylexicaux ont été choisis pour chaque domaine et nous avons appliqué l'union des deux listes pour les expériences de classification entre domaines. Nous utilisons TfxIFD calculé pour chaque terme dans le vecteur représentant chaque document.
- STAT : les propriétés statistiques (longueur des mots courts et longs, longueur de la phrase, fréquence des signes d'interrogation ou d'exclamation, fréquence des parenthèses indiquant des explications) ;
- SYN : les propriétés syntaxiques et plus généralement des annotations de haut niveau (la fréquence relative des certains noms abstraits identifiés dans la section précédente, les pronoms *nous* et *vous*, les verbe de cognition ou de modalité, la fréquence des adjectifs relationnels, des constructions passives, des tournures impersonnelles, des sujets et des objets complexes, les définitions et des marqueurs de reformulation).
- ALL : toutes les propriétés statistiques et syntaxiques, à l'exception des termes;
- Unigrams : l'union des mots pleins apparaissant dans le corpus médical ou dans le corpus informatique. Il s'agit du système de base, ces propriétés peuvent être extraites sans recours aux annotations linguistiques de haut niveau.

Pour calculer la fréquence relative de ces propriétés pour chaque document nous avons appliqué la formule et les patrons (présentés en section 3.2) pour identifier certains phénomènes complexes tels que les tournures impersonnelles, les constructions passives, les définitions ou les explications.

Une fois les propriétés extraites, nous avons effectué des tests à l'aide de Weka, une plateforme de classification automatique (Hall *et al*, 2009). Nous avons classé manuellement les documents comme étant des textes scientifiques ou de vulgarisation. Notre corpus est constitué de 302 textes/genre dans le domaine médical et 301 textes/genre dans le domaine informatique. Les résultats présentés sont obtenus avec l'algorithme SMO, l'implémentation du classifieur SVM disponible sur Weka. Nous avons effectué plusieurs expériences présentées dans le tableau 2 :

Classes	Modèle	Corpus de test	ALL	STAT	SYN	TERMES	Unigrams
2 classes DS, DV	médecine	médecine	96.97%	93,15%	92,51%	87,42 %	99,08%
	informatique	informatique	93,12%	87,30%	83,63%	95,76 %	99,07%
	médecine	informatique	76,16%	74,18%	71,08%	55,43%	65,32%
	informatique	médecine	91,24%	87,89%	83,63%	47,55%	54,25%
4 classes DSMED, DVMED, DSINFO, DVINFO			85,18%	71,36%	77,54%	89,42 %	98,22%

TABLE 2 : L'exactitude (E) obtenue pour les 4 systèmes et le système de base sur un corpus de test du domaine et hors domaine

1) **Expériences de classification à l'intérieur du domaine, et entre les domaines considérant deux classes (discours de vulgarisation et discours scientifiques).** Notre objectif est de vérifier si certaines propriétés choisies peuvent être généralisées entre les domaines pour caractériser les discours scientifiques et de vulgarisation. Pour ces tests, nous avons appliquée la validation croisée (k=10). Nous avons construit un modèle pour chaque groupe de propriétés pour le corpus médical et informatique. Pour construire le modèle des termes du domaine sélectionnés manuellement (TERMES), nous avons utilisé l'union des deux ensembles de termes extraits à partir des corpus médical et informatique. Pour la classification à l'intérieur du domaine (corpus de test et d'entrainement du même domaine), on constate que les termes sont plus efficaces pour le corpus informatique (E=95,76%) mais le système ALL obtient 93,12% des instances correctement classées. Pour le corpus médical, le système ALL (E=96,97%) est meilleur que TERMES (E=87,42%) Pour la classification entre domaines, nous constatons que les listes de termes sélectionnés manuellement sont peu efficaces, aussi bien que le système de base Unigram. Pour la classification inter-domaine, le modèle ALL construit sur le corpus informatique est le meilleur (exactitude de 91,24 %), alors que dans l'autre direction, ALL reste toujours le plus performant (76.16%). Les systèmes utilisant exclusivement des annotations de haut niveau (SYN) ont obtenu des résultats plus faibles que les systèmes utilisant des propriétés statistiques (STAT). Le système Unigram reste très efficace pour la classification à l'intérieur du domaine.

2) **Expériences de classification, réalisées en considérant 4 classes (DSMED, DVMED, DSINFO, DVINFO)**. Dans ce deuxième cas, nous avons appliqué la méthode de validation croisée (avec k=10) sur le corpus d'entraînement. Parmi les quatre configurations de propriétés proposées dans notre approche, il s'avère que les termes restent toujours les meilleurs alors que ALL est en seconde position. Le meilleur score est obtenu par le système Unigram (98,22%).

Les résultats obtenus montrent que la combinaison de propriétés statistiques et syntaxiques ALL semblent généralisables plus facilement entre domaines. Toutefois, la taille du corpus est limitée et les résultats doivent être confirmés sur un corpus de taille plus importante. Le calcul de certaines propriétés s'appuie sur l'existence d'annotations syntaxiques automatiques, dont le résultat n'a pas été corrigé. Certains propriétés (par exemple la fréquence des pronoms relatifs ou les énoncés définitoires) sont évaluées à l'aide d'une catégorie lexicale ou d'une séquence d'étiquettes. Les erreurs provenant de l'analyse automatique peuvent influencer l'extraction de propriétés et donc les résultats de la classification.

4 Conclusion et perspectives

Les expériences de classification à l'intérieur du domaine montrent que l'ensemble de propriétés lexicales, morpho-syntaxiques et syntaxiques sont aussi performantes que les termes. En ce qui concerne la classification entre les domaines, l'ajout des propriétés morpho-syntaxiques se révèle plus effective que la simple sélection manuelle des termes du domaine. Les expériences de classification ont pris en compte des propriétés des genres étudiées dans la littérature par la suite vérifiées à l'aide d'une analyse de corpus détaillée. D'autres comparaisons avec un système de classifications utilisant des n-grams de caractères seront effectuées. Néanmoins, il est évident que ces résultats doivent être validés sur des corpus de taille plus importante, ainsi que l'évaluation des erreurs dues à l'annotation automatique et aux règles d'extraction de propriétés (par exemple des règles qui identifient les définitions ou les tournures impersonnelles), qui s'appuient sur cette annotation. La méthode peut être appliquée à d'autres domaines et elle peut s'avérer utile pour identifier des sous-genres académiques (thèse, article scientifique).

Références

BIBER D., CONRAD S. (2009). *Register, Genre and Style*, Cambridge University Press.

BOHNET B. (2009) Efficient Parsing of Syntactic and Semantic Dependency Structures. *Proceedings of Conference on Natural Language Learning (CoNLL)*, Boulder, 67-72.

CHARNOIS T., DOUCET A., MATHET Y. (2008). Trois approches du GREYC pour la classification de textes. *Actes TALN'08*, Avignon

CHRIST,O. (1994) A Modular and Flexible Architecture for an Integrated Corpus Query System Proceedings of COMPLEX'94, Budapest, Hungary, July 7-10, 1994, PP- 23-32

DAILLE B. (1999). « Identification des adjectifs relationnels en corpus », Conférence TALN1999, Cargèse

D'HONDT E., VERBERNE S., KOSTER C., BOVES L. (2013). Text Representations for Patent Classification. Computational Linguistics, 39(3), 755–775.

DROUIN, P. (2007) « Identification automatique du lexique scientifique transdisciplinaire », Revue française de linguistique appliquée 2/2007 (Vol. XII) , p. 45-64

HYLAND,K (2009). Academic Discourse, London: Continuum.

HALL M., ET AL (2009). « The WEKA Data Mining Software: An Update », SIGKDD Explorations, Vol. 11/1.

GOEURIOT L., MORIN E., ET AL. (2009). « Reconnaissance du type de discours dans des corpus comparables spécialisés », CORIA Conférence en Recherche d'Information et Applications.

JACOBI D., SCHIELE B. (ÉD.) (1988). *Vulgariser la science*, Seyssel, Éditions Champ Vallon (Milieux).

KARLGREN J., CUTTING D. (1994). Recognizing text genres with simple metrics using discriminant analysis. *Actes de COLING 94*, 1071-1075.

KESSLER B., NUNBERG G., SCHÜLTZE H. (1997). Automatic detection of text genre. Actes de EACL'97, 32-38.

KOCOUREK R. (1991). *La langue française de la technique et de la science*, Oscar Brandstetten Verlag.

LECLUZE, C, LEJEUNE, G. (2014). DEFT 2014, analyse automatique de textes littéraires et scientifiques en langue française (stylometrie et quelques catégories lexicales) , DEFT 2014, Marseille, 1er juillet 2014

LEE Y.-B., MYAENG S. H. (2004) Automatic identification of text genres and their roles in subject-based categorization, *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*

MALRIEU D., RASTIER F. (2001). Genres et variations morphosyntaxiques. *Traitement Automatique des langues*, vol. 42, n°2, pp. 548-577.

PETRENZ P., WEBBER, B. (2011). Stable Classification of Text Genres. *Computational Linguistics* 37:2, 385-393

POUDAT C., CLEUZIOU G., CLAVIER V. (2006). Catégorisation de textes en domaines et genres : complémentarité des indexations lexicale et morphosyntaxique. *Document numérique*, vol.9, n°1/2006, pp. 61-76.

RINCK F. (2007). Styles d'auteur et singularité des textes. Approche stylométrique du genre de l'article en linguistique, Pratiques, 135/136, 119-136.

SANTINI M. (2007). Automatic Identification of Genre in Web Pages, Ph.D.Thesis, University of Brighton

SEBASTIANI F. (2005). « Text categorization » In Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press, Southampton, UK, 2005, pp. 109-129.

SUN J., YANG Z., LIU S., WANG P. (2012). Applying stylometric analysis techniques to counter anonymity in cyberspace. *Journal of Networks*, 7(2)

STAMATATOS, E., FAKOTAKIS, N. ET KOKKINAKIS, G. (2000). Automatic Text Categorization in Terms of Genre and Author. In *ComputationalLinguistics*, Vol.26, No. 4, pages 471–497.

SWALES, J. (2004) Research Genres: Explorations and Applications, Cambridge Applied Linguistics.

TODIRASCU, A., PADO, S., KISSELEW, K., KRISCH, J., HEID, U. (2012) French and German corpora for audience-based text type classification, Proceedings of LREC 2012

TUTIN A. (2010). Evaluative adjectives in academic writing in the humanities and social sciences. In Lores-Sanz, R., Mur-Duenas, P., Lafuente-Millan, E. Constructing Interpersonality: Multiple Perspectives on Written Academic Genres. Cambridge: Cambridge Scholars Publishing..