# Projet *TextCOOP*

## Typologies textuelles pour le traitement automatique des documents électroniques : problèmes et enjeux

[Rapport d'analyse]

Maria ZIMINA-POIROT

(Version du janvier 2007)

LIPN - Laboratoire d'Informatique de l'université Paris-Nord (UMR7030)

# Sommaire

*« Obtenir de bons résultats à partir du traitement automatique de la langue sur corpus présuppose une différentiation théorique claire entre la représentation des connaissances et la construction du sens en discours. La première dérive de l'observation de l'état des choses, la deuxième rend compte de la diversité dénominative inhérente à une communauté linguistique spécialisée. »*

Rute COSTA [2006, p. 86]

## 1. La problématique des typologies textuelles

La problématique des typologies textuelles constitue un objet d'études assez complexe et délicat étudié depuis fort longtemps.[1] On peut affirmer qu'il existe tellement de critères pouvant contribuer à la démarche typologique dans l'analyse textuelle qu'il est difficile de dresser une liste exhaustive d'approches de la construction des typologies. Parmi les tentatives de systématisation dans ce vaste champ d'études, on pourrait citer, par exemple, les travaux de A. Petitjean [1989], J-M. Adam [1992], J.-P. Bronckart, [1996], S. Branca-Rosoff [1999].

### 1.1. Objet 'texte' : de l'hétérogénéité textuelle à l'unité sémantique globale

Les conclusions des chercheurs ayant abordé les problèmes de construction des typologies textuelles sont aussi très divergentes. On trouve dans la littérature récente consacrée à la problématique d'analyse automatique des textes des comptes-rendus d'expériences qui témoignent d'un progrès significatif dans l'exploration typologique [Argamon *et al.*, 2005], ainsi que des constats beaucoup plus réservées venant des chercheurs en sciences du langage qui vont même jusqu'à l'affirmation que la notion même de typologies de textes a un caractère prématuré [Roulet, 1991].

Formellement, la création d'une typologie nécessite une stabilisation d'un ensemble de traits. Quels sont donc les critères jugés pertinents pour la création des typologies de textes ?

Cette question nous mène directement à la principale difficulté de la démarche typologique dans l'analyse textuelle : le phénomène d'*hétérogénéité textuelle*. Cette hétérogénéité textuelle se manifeste à plusieurs niveaux.

Un texte peut contenir plusieurs séquences (ou micro-textes) de nature différente (*argumentatif*, *narratif*, etc.) La notion de *polyphonie textuelle*[2] est aussi convoquée dans ce contexte pour contribuer à l'identification des réseaux textuels et, par conséquent, à l'étude de la cohérence textuelle à l'aide des repérages énonciatifs [Fløttum et Holm, 1999]. La cohérence textuelle est alors assurée par divers moyens linguistiques tels que la progression thématique, les relations anaphoriques, les connecteurs, etc. [Charolles, 1988] ; [Pery-Woodley, 2001].[3]

Nombre de travaux sont consacrés à la notion de *paratexte* [Genette, 1987]. De façon extrêmement sommaire on pourrait résumer cette notion en signalant que tout texte est généralement entouré par l'ensemble des discours de commentaires, de présentations ou d'accompagnements qui viennent le compléter, l'étayer, l'argumenter. Pour un document papier, il y a des règles bien connues du monde de l'édition pour l'insertion de tels éléments (encadrés, images, figures, etc.). Leur rôle consiste à maintenir la linéarité du texte et à offrir au lecteur plusieurs parcours de lecture. Dans le monde de l'édition électronique l'intégration des liens hypertextes participe activement à la consolidation du document. La pose des liens ainsi que l'étude des règles sémantiques et sémiotiques qui les régisse constitue d'ailleurs une autre facette d'étude des typologies textuelles sur le Web.

Un texte ne peut s'écrire indépendamment de ce qui a déjà été écrit et porte les traces et la mémoire d'un héritage culturel et d'une tradition d'écriture. Le terme *intertexte* désigne alors l'ensemble des ressources textuelles antérieures au texte donné que ce dernier utilise sous forme de citations, de paraphrases ou d'allusions. Les possibilités offertes par la publication électronique ont beaucoup contribué à la formalisation de ces relations d'intertextualité en précisant les rapports spatiaux et temporels entre les textes. Dans les travaux actuels l'*hypertexte* est d'ailleurs perçu comme une sorte d'intertextualité technologique :

*« La textualité de l'hypertexte fait le choix du spectare plutôt que celui du scribere. Rupture. Et dans son immense majorité, elle continue de donner principalement à voir ce qui est écrit. Continuité. Un texte à un début et une fin. Un hypertexte n'a ni début ni fin. Il dispose d'un ou plusieurs points d'amorçage. A partir de l'activation de l'un de ses points il se met en mouvement, jusqu'à l'épuisement de celui qui l'a créé, de celui qui le parcourt, ou de ses propres ressources. »* [Ertzscheid, 2002].

### 1.2. Enjeux économiques : gestion d'accès aux contenus textuels pertinents

Un objet textuel peut se matérialiser sous formes différentes et comporte des contenus informationnels variés. L'environnement, le support, les modalités d'inscription du texte sont importants dans les relations qui unissent la connaissance que les hommes ont des textes et les connaissances que les textes donnent aux hommes [*ibid*].

La question du support est d'ailleurs privilégiée par le courant de l'analyse structurale des textes[4] qui considère l'organisation du support comme élément important pour l'établissement d'une typologie textuelle. L'étude de la structure du support offre des éléments précieux pour l'exploration d'ordre « archivistique » souvent privilégiée pour la gestion de l'information.

Sur le plan communicationnel, un document textuel est généralement produit pour faire passer un message. Sur le plan informationnel, il est caractérisé par une unité sémantique et logique globale. En même temps, il est constitué d'un ensemble de parties reliées entre elles. Chaque partie (unité) du document joue un rôle précis. Ce découpage est d'ailleurs utilisé dans les systèmes de navigation textuelle (recherche documentaire, questions-réponses) qui se servent de l'éclatement du document en parties (unités d'informations) pour présenter à l'utilisateur une information plus facile à assimiler [Ben Romdhane, 2001] ; [Benamara *et al*., 2005].

Malgré une certaine simplicité apparente, le découpage du texte en unités d'information et la modélisation des structures textuelles adéquates relèvent d'une tâche particulièrement complexe dont la formalisation est actuellement en cours de l'étude [Pery-Woodley, 2000] ; [Hernandez, 2004]. En effet, la structure des documents n'est pas uniforme, elle change d'un type de document à un autre, d'une discipline à une autre, etc. Certains textes (articles scientifiques, documents techniques) suivent une structuration normalisée et globalement plus stable que les autres qui permettent une plus grande liberté lors de leur création (textes littéraires, essais, etc.).

Beaucoup de travaux récents partent de ce constat pour explorer les possibilités de caractérisation automatique des propriétés des documents textuels. Dans ces travaux, les propriétés de description plus au moins classiques (titre, auteur, etc.) sont utilisées conjointement avec des propriétés se rapportant à l'environnement de production du document (champ disciplinaire, profession, et communauté de l'auteur) et au support de diffusion (type de l'environnement éditorial) [Marshman, 2003].

Ces « variables » utilisée pour l'attribution des propriétés aux documents sont intimement liées au profil de l'utilisateur (lecteur) potentiel. Dans ce contexte, on avance un modèle d'un

module d'aiguillage qui devrait aider à établir un lien entre, d'une part la requête et le profil de l'utilisateur, et d'autre part les contenus des documents textuels disponibles sur le thème de la question :

*« Pour répondre à l'exigence de pertinence de la recherche documentaire comme l'extraction d'information, il faut mettre en œuvre des moyens d'analyse du contenu reposant sur des ressources lexicales, syntaxiques et sémantiques spécifiques au domaine étudié. Des représentations normalisées et enrichies des documents peuvent alors être automatiquement construites sur lesquelles peuvent s'appliquer des moyens de sélection de documents et d'extraction d'information. Outre une meilleure précision, ces moyens d'accès à l'information permettent alors de répondre à d'autres besoins que l'exploration thématique, par exemple, des besoins de fiabilité, d'originalité, et de synthèse. »* [Nazarenko, 2003, p. 6].

On comprendra aisément les enjeux potentiels d'un tel système de régularisation des processus d'accès à l'information à l'époque où des volumes croissants des données textuelles sont générés quotidiennement sur Internet [Ceheux, 2002] ; [Meyer Zu Eissen et Stein, 2004]. Ces enjeux économiques expliquent le regain d'intérêt vers la problématique des typologies textuelles au sein de la communauté de la recherche d'information.

Le constat apparent est simple : si l'on parvient déjà à localiser l'ensemble des documents répondant au « thème » de la requête formulée par l'utilisateur, pourrait-on identifier plus précisément les différents profils des documents textuels « ramenés dans les filets » ? [Santini, 2003]. On remarque que ce profil de document est souvent véhiculé par la notion de *genre textuel*.

## 2. La notion de genre textuel pour une exploration typologique

De façon générale, les travaux consacrés à la notion de genre textuel se multiplient et se diversifient. Ils sont issus des recherches en sociolinguistique, analyse conversationnelle [Moirand, 2003], analyse de discours [Beacco, 1992], linguistique et sémantique textuelle [Adam, 1999] ; [Rastier, 1989], apprentissage automatique et traitement automatique des langues [Karlgren et Cutting, 1994] ; [Finn et Kushmerick, 2003], etc.

Une réflexion sur les modes de traitement linguistique de la généricité textuelle amène certains chercheurs à adopter plusieurs points de vue sur le genre. Un panorama de ces points de vue est exposé, par exemple, dans les articles de Petitjean [1989, 2005], Branca-Rosoff [1999], Moirand [2003].

Selon le point de vue sociologique, le genre fonctionne comme une sorte de régulateur des auteurs, des textes et des lecteurs.

Du point de vue des théories de la réception, le genre est perçu plutôt comme un ensemble de règles, de régularités, de conventions, en fonction desquelles le lecteur s'émerge dans le texte.

Les recherches en poétique sont menées autour d'une réflexion sur le statut universel de la généricité textuelle est la diversité culturelle des genres. Dans cette perspective, le genre sert d'étalon qui permet d'identifier le statut du texte (*genre premier* ou *genre second*)[5], son degré de complexité (unicité ou mixité générique), son originalité, etc.

Du point de vue historique, l'éloignement temporel ou culturel stimule la volonté d'homogénéiser les actualisations génériques. Les processus de transformation, radicalisation, hybridation, contamination et multiplication des genres sont pourtant des phénomènes de la vie courante qui font partie de la création textuelle [Petitjean, 2005].

La notion de domaine a une valeur en reconnaissance des genres. Selon F. Rastier, par exemple, chaque type de pratique sociale correspond à un domaine sémantique et à un discours qui l'articule. Le discours juridique, par exemple, comporte des textes de jurisprudence, des rapports, des décisions de justice, des comptes-rendus d'audiences, etc. [Malrieu et Rastier, 2001] ; [Rastier, 2006].

On pourrait reprendre la définition de S. Moirand pour tenter de résumer ces multiples facettes du *genre* :

*« une représentation socio-cognitive intériorisée que l'on a de la composition et du déroulement d'une classe d'unités discursives, auxquelles on a été « exposé » dans la vie quotidienne, la vie professionnelle et les différents mondes que l'on a traversés, une sorte de patron permettant à chacun de construire, de planifier et d'interpréter les activités verbales ou non verbales à l'intérieur d'une situation de communication, d'un lieu, d'une communauté langagière, d'un monde social, d'une société…»* [Moirand, 2003, p. 20]

***

A nos jours, de plus en plus de travaux qui visent l'automatisation de l'analyse textuelle considère la notion de genre comme une sorte de meta-catégorie d'aide susceptible d'améliorer ou optimiser les traitements spécialisés. Pour l'interprétation de textes spécialisés, par exemple, on souligne que la classification en domaines n'est plus suffisante et d'autres critères d'ordre sociologique et sociolinguistique devraient s'y ajouter, comme l'identification des communautés scientifiques productrices et réceptrices de ces textes, la diversité des situations de communication et l'encadrement spatio-temporel de leur diffusion [Silva, 2005] ; [Aussenac-Gilles *et al*, 2002].

Dans ce contexte, on convoque les typologies en genres pour tenter d'améliorer les performances d'outils informatiques existants [Sekines, 1997] ; [Illouz *et al*., 2000a]. On envisage, par exemple, d'entraîner des outils d'analyse syntaxique pour chaque genre de texte. Il est aussi question d'utiliser la notion de genre pour enrichir d'importants volumes de données textuelles en recherche d'information [Ouerfelli et Lallich-Boidin, 2000].

La reconnaissance automatique des genres intéresse également les acteurs du monde de recherche documentaire qui espèrent ainsi améliorer l'efficacité des applications qui rencontrent le problème de l'ambiguïté sémantique.

En terminologie, un terme donné n'entrera pas dans les mêmes réseaux d'association de termes et les éléments sémiques ne sont pas mobilisés de la même façon selon le genre [Slodzian, 2000] ; [Costa, 2006].

En linguistique de corpus, les chercheurs sont de plus en plus confrontés à des corpus volumineux et hétérogènes [Illouz *et al*., 1999b]. Les questions de profilage se posent de façon aigue et la reconnaissance automatique des genres pourrait aussi être très pertinente [Illouz *et al*., 2000b]. L'hypothèse avancée consiste à utiliser une description fine des informations concernant les profils des documents traités pour arriver à des analyses nuancées

des phénomènes et à des prédictions plus précises sur leur fonctionnement pour certains types de traitements automatisés [Pery-Woodley, 1995] ; [Lee, 2001].

La notion genre est également convoquée en traduction assistée par ordinateur. On utilise une base de données importante de traductions préalablement effectuées (mémoires de traduction) comprenant à chaque fois les données source et cible, le programme cherchant alors dans ce volumineux ensemble d'informations la version la plus adaptée. Pour cette tâche, chaque type de texte devrait être mis en rapport avec des documents appartenant à la même catégorie : par exemple, les écrits techniques doivent être comparés entre eux, et pas avec des extraits de roman, etc. [Ratliff, 2006].

Au sein de ces applications variées, le genre est perçu comme une forme d'aide à la compréhension et/ou interprétation de textes mais la définition formelle de la notion de genre pour sa reconnaissance automatique reste problématique.

Peut-on avancer une définition du genre qui pourrait aider à caractériser les textes dans l'objectif typologique ?

Rappelons que depuis plusieurs décennies, en sciences humaines est particulièrement en sciences du langage, le genre est perçu comme un outil sémiotique complexe, c'est-à-dire une forme langagière prescriptive permettant à la fois la production et la compréhension des textes. Les travaux théoriques sur les genres discursifs remontent à M. Bakhtine qui a noté déjà dans les années trente que chaque sphère d'utilisation de la langue correspond à des types relativement stables d'énoncés. D'où la possibilité d'étudier les corrélations des caractéristiques extralinguistiques avec des caractéristiques linguistiques *régulièrement attestées* dans les textes. Dans ce contexte, les typologies textuelles sont perçues comme des outils d'analyse permettant une maîtrise plus consciente des genres [Mangenot, 1996] ; [Schneuwly, 1994]. La difficulté principale vient du fait que les combinaisons potentielles d'ensembles de caractéristiques extralinguistiques sont pratiquement inépuisables comme la variété des pratiques et des activités humaines liées aux productions/utilisations des ressources textuelles.

Face à ce constat méthodologique important, il faut souligner l'intérêt des *différentes* typologies textuelles qui reposent sur des critères variés [Sinclair et Ball, 1996]. Autrement dit, une typologie textuelle seule n'a pas d'intérêt en *soi*. En revanche, on s'attend à ce qu'elle apporte des éléments de réponse à des *besoins applicatifs précis* :

*« En résumé, les conditions générales de l'établissement d'une typologie des genres exigent que soit précisé le type d'*objet-texte *auquel s'applique la typologie, que soit déterminé le* lieu de pertinence *dans lequel agit la typologie et que soient définis des axes de typologisation selon les critères homogènes d'organisation discursive (que le matériau soit verbal ou visuel). »*

[Charaudeau, 1997, p. 5].

## 3. Problèmes de reconnaissance automatique des genres textuels

### 3.1 Multiples axes de typologisation

L'automatisation de la reconnaissance des genres textuels est intimement liée aux difficultés propres au traitement automatique de la langue naturelle. Plusieurs projets en cours s'intéressent aux méthodes permettent l'indentification des genres. Un panorama de ces recherches est présenté, par exemple, dans [Luštrek, 2006] ; [Santini, 2004] ; [Finn et Kushmerick, 2003]. D'un point de vue formel, cette démarche s'appuie sur une consolidation d'un ensemble de traits. Or, cette tâche est particulièrement complexe dû à l'ensemble des phénomènes de l'hétérogénéité textuelle.

Existent-il vraiment des variables typiques de genre ? Peut-on capter par un ensemble déterminé de traits des phénomènes liés à la parenté, mutation, rupture de genres ? Sur quelles bases formelles peut-on envisager la mise en place des procédures de comparaison, de rapprochement ou de différenciation entre les genres ? Comment capter les phénomènes de filiation entre les genres et les sous-genres ? Ces questions se posent de façon aigue lorsque l'on s'intéresse à la problématique des typologies du point de vue de l'automatisation des traitements textuels.

Une piste de réflexion qui permettrait d'apporter des éléments de réponse à cette interrogation consiste à considérer le texte comme une réalité sémantique et sémiotique complexe. Dans cette optique, il est possible de comparer plusieurs textes selon plusieurs axes. Cette nécessité de comparaisons multiples est palpable si l'on considère la variété des paramètres suggérées pour l'analyse textuelle, tels que *l'ancrage institutionnel*, *l'intention communicationnelle*, *le mode énonciatif* et *la situation de production*, *l'organisation formelle*, *les contenus thématiques*, *les indices para et péritextuels*, etc. [Petitjean, 2005] ; [Bommier-Pincemin, 1999].

Les combinaisons de comparaisons potentielles pourraient être multipliées pratiquement à l'infini. D'où la nécessité d'une réflexion sur le statut des critères de comparaisons, les modalités de leurs interactions, leur degré de généralité et leur lien avec les pratiques sociales concernées [Malrieu et Rastier, 2001].

Les choix sont multiples mais une catégorisation doit s'appuyer sur une constellation de propriétés relativement stabilisée. Dans les travaux actuels, cet ensemble de propriétés est évoqué sous les noms différents : matrice discursive, critère définitoire, conditions

d'énonciation, support et modes de diffusion, mode d'organisation, sémantique, etc. [Petitjean, 2005] ; [Charaudeau 1997] ; [Charaudeau et Maingueneau, 2002]. On évoque des catégories qui touchent aux différentes niveaux d'analyse : sémantique, énonciatif, longueur, pragmatique, compositionnel, stylistique, etc. [Adam, 1992].

Quel axe de typologisation faut-il privilégier ? Quel modèle sera le plus efficace ? Quel système de traits faut-il choisir ?

Lorsque l'on essaye d'intégrer le plus grand nombre de variables possible, on gagne en compréhension mais on perd en lisibilité : une typologie trop complexe devient rapidement inefficace. Si l'on diminue le nombre des variables, on gagne en lisibilité mais on perd en compréhension : la typologie devient alors trop réductrice. Une des solutions consisterait à faire appel au principe de *hiérarchisation* : on construit une première typologie approximative, puis en faisant intervenir d'autres variables à l'intérieur des axes de base, on construit des typologies successives qui s'enchâssent dans le modèle de base [Charaudeau, 1997].

La dimension applicative reste essentielle pour la construction des fécaux d'indices. Mangenot [1996] souligne qu'il est important d'utiliser les traits qui reposent sur les marques de surface linguistiques et permettent d'établir un rapport clair à des situations d'énonciation typées ; il note aussi qu'il convient de privilégier les activités que les classifications obtenues aident à mettre en place plutôt que les typologies elles-mêmes. Le dernier point semble particulièrement pertinent pour l'utilisation des typologies textuelles dans les différents domaines du TAL.

### 3.2. Sélection de traits pertinents

Dans les travaux existants consacrés aux questions de classification automatique en genres, la constitution des fécaux d'indices a pour objectif de capter les régularités en fonction desquelles on pourrait amorcer le travail d'identification générique. Dans la communauté du TAL, la sélection de traits linguistiques pertinents est perçu comme un enjeu important : un jeu de traits «équilibré » permettrait alors à optimiser les traitements sur les données pour lesquelles il a été conçu [Karlgren et Cutting, 1994] ; [Copeck *et al.*, 2000] ; [Santini, 2005].

Nombres de choix sont possibles, partant de traits surfaciques (comme les caractères) jusqu'à des traits linguistiques très variés (la fréquence des verbes à l'impérative, la fréquence d'un certain type de subordonnée, etc.) [Malrieu, 2004]. Le coût de l'interaction est souvent perçu

comme un paramètre très important, ainsi que la possibilité de corrélation entre le jeu de traits et le traitement que l'on vise à améliorer [Kessler *et al.*, 1997] ; [Meyer Zu Eissen et Stein, 2004].

Avant de faire appel à des mécanismes de calcul des jeux de traits complexes, il est utile de réfléchir à des niveaux plus accessibles, comme, par exemple, le niveau de caractères. Cette information peut sembler très pauvre du point de vue de l'analyse linguistique fine mais elle fournit néanmoins des indices sur la forme du texte [Dewdney *et al.*, 2001] ; [Rauber et Muller-Kogler, 2001]. On note, par exemple, que la fréquence du point est corrélée à la combinaison de la longueur des phrases, de la présence d'acronymes et de nombres. L'espace est un indicateur de la longueur des mots, la proportion de lettres capitales fournit des renseignements concernant le nombre des noms propres, etc. [Illouz, 1999a]. Les autres indices calculés au niveau des caractères comprennent, par exemple, les sigles commerciaux ou monétaires ($, £, ©, ®, ™ …), les dates (avec prise en compte des formats), les acronymes (ANAES, POS, …), la ponctuation (.,:;!?« »" …), la longueur moyenne des mots et des phrases, la proportion de mots longs (plus de *n* caractères, par exemple), etc.

La plupart des travaux fixant comme objectif l'analyse au niveau du genre affiche une volonté de minimiser le recours au lexique. Cependant, on note que le bilan quantitatif est souvent réalisé en s'appuyant sur les résultats de segmentation automatique en formes graphiques. Ces informations permettent de mesurer un certain nombre de caractéristiques stylistiques de chaque texte et sont généralement faciles et rapide à calculer [Stamatatos *et al.*, 2000].

Certains décomptes qui relèvent du niveau lexical font appel à des ressources de dictionnaires existants pour identifier, par exemple, les mots «communs », les termes appartenant à des « langues de spécialités », les fautes d'orthographes, les mots « inconnus » des dictionnaires, etc. [Meyer Zu Eissen et Stein, 2004].

Au sens large, le bilan quantitatif fondé sur des indices linguistiques peut être réalisé à plusieurs niveaux d'analyse : niveau caractères / formes graphiques, niveau lexical, niveau morpho-syntaxique (POS), niveau syntaxique (analyse de dépendances syntaxiques), niveau sémantique, niveau typo-dispositionnel (on compte, par exemple, le nombre de figures, tables, titres), caractéristiques générales du document (longueur en phrases/paragraphes, type/token ratio, etc.) C'est précisément une utilisation conjointe de ces différents niveaux d'analyse qui est visée dans les classifications multi-facettes en genres utilisées dans les recherches en cours [Crowston et Kwasnik, 2004] ; [Argamon *et al.*, 2005] ; [Crowston et Kwasnik, 2004] ; [Sugar Boese, 2005], [Santini, 2006] ; [Clavier, 2006].

En revanche, la complexité de traitements sous-jacents à certains niveaux d'analyse linguistique fine (comme par exemple l'analyse syntaxique) est telle que ces traitements ne sont envisagés que sur des données textuelles relativement limitées lorsque le cadre applicatif permet d'accepter le temps de calcul plutôt élevé.[6] Ainsi, le recours à l'analyse linguistique fine n'est pas vraiment à l'ordre du jour pour typer les documents du Web en temps réel. Dans ce contexte, la classification en genres réalisée *après* l'indentification des documents pertinents par un moteur de recherche, doit être particulièrement rapide pour tenir compte des délais d'attente de l'utilisateur.

La prise en compte des niveaux relevant de l'analyse linguistique fine pose aussi des problèmes de disponibilité de ressources et d'outils linguistiques adéquats. Rappelons que c'est précisément pour améliorer la qualité de ces derniers que l'on s'intéresse à l'identification automatique du genre. Cette dualité n'est pas résolue dans la plupart des travaux actuels qui proposent le plus souvent des jeux de traits linguistiques *hétérogènes* dont la composition est guidée par l'intuition des chercheurs, la nature des données d'observation, la disponibilités d'outils permettant d'automatiser le calcul à certains niveaux d'analyse et le résultat final visé.

Un souci méthodologique important resurgit dans plusieurs travaux en reconnaissance automatique des genres : les des données textuelles sont représentées par un vecteur de traits (*feature vector*) relevant des niveaux d'analyse sensiblement différents. Comme le note [Liao *et. al*., 2003, p. 1] n'importe quel indice relevé dans un texte pourrait être représenté par un trait : *« A feature can be as simple as a single token, or a linguistic phrase, or a much more complicated syntax template. A feature can be a characteristic quantity at different linguistic levels. »* Au final, c'est précisément le bilan quantitatif fondé sur les occurrences de ces traits que l'on cherche à utiliser à des fins de catégorisation. Pour évaluer le poids quantitatif d'un trait dans un document donné on porte une estimation sur son niveau de couverture : *« A feature weight should show the degree of information represented by local feature occurrences in a document, at a minimum. »* [*ibid.*]

Selon cette approche, plus il y a de marques qui relèvent des fécaux d'indices élaborés pour l'exploration typologique (lexicaux, morpho-syntaxiques, typo-dispositionnels, etc.), plus il y a de chances de reconnaître correctement le profil visé. Pourtant, ce modèle ne reflète que très peu la réalité discursive. L'utilisation des taux de concentration de certaines marques linguistiques pour amorcer le travail d'identification générique soulève beaucoup de questions. Comme le remarque Charaudeau [1997] au sujet des genres audiovisuels, on

confond les dominantes et les procédés d'organisation du discours (descriptif, narratif, argumentatif), puis, pour pallier à ce défaut, on fait appel à une opération de pondération, en constatant que certains procédés sont plus dominants que d'autres dans tel ou tel texte : *«… il y a parfois des coïncidences entre dominantes et type de texte : par exemple les articles de dictionnaire ne sont que descriptifs, les panneaux de la circulation ne sont qu'injonctifs. Mais quelques coïncidences font-elles un principe de typologisation des genres ? »* [*ibid*, p. 7]

Par conséquent, la constitution des « jeux de traits » pour une typologie des genres se heurte à un problème méthodologique important lié à la nature même du *genre* :

*« Le genre est du côté de la configuration textuelle comme résultat global de ce qui a présidé à sa construction (résultats global et en même temps composite, ce qui rend si difficile et si discutable toute tentative de classement en genre) ; le procédé est du côté de l'outillage sémio-discursif dont chacun des éléments ne signifie en soi que partiellement et contribue au processus de configuration textuelle : une argumentation, une image de montage ou une feintise sont autant de procédés qui, certes, peuvent intervenir comme trait définitoire d'un genre, mais ne peuvent être confondus avec celui-ci. »*

## 4. Profilages, typologies, catégorisations : axes de recherches

L'aperçu de la littérature existante montre qu'il existe plusieurs axes de recherches autour des questions de profilages/typologies/catégorisations textuels. On remarque au moins trois pistes de réflexion méthodologique : la démarche inductive, les catégorisations supervisées et la mise en place des procédés d'analyse de la distance intertextuelle. Comme on peut l'imaginer facilement, des applications particulières mélangeant ces trois approches sont également possibles.

### 4.1. Typologies induites des textes

Les travaux de Biber [1988 ; 1995 ; 2004] ont fourni une base de réflexion solide pour la démarche inductive dans l'exploration typologique. Dans le contexte francophone, ils ont été repris, enrichis et diversifiés grâce au travail de plusieurs chercheurs, tels que [Habert *et al.*, 2000] ; [Beaudouin *et al.*, 2001] ; [Malrieu, 2001 ; 2004] ; [Beauvisage, 2004].

L'approche inductive consiste à s'intéresser aux ensembles de traits divers (marqueurs de temps et d'aspect, questions, passifs, modaux, etc.), d'étudier lesquels d'entre eux sont associés ou dissociés du point de vue de leur fonctionnement linguistique et de faire émerger au travers de ces corrélations des regroupements textuels cohérents (*types de textes*). Ce sont des méthodes statistiques multidimensionnelles (*analyse des correspondances*, *classification automatique*, *analyse discriminante*) qui sont mises à contribution pour découvrir simultanément les regroupements pertinents à la fois dans l'espace des traits et dans l'espace des textes [Lebart et Salem, 1994] ; [Habert et Salem, 1995] ; [Habert *et al.*, 1997].

La souplesse des méthodes statistiques s'adapte particulièrement bien aux comparaisons textuelles multiaxiales qui sont incontournables lorsque l'on cherche à découvrir des ensembles de traits associés ou dissociés en terme de leurs distributions. Ces attractions/répulsions entre les agrégats de traits divers sont identifiées automatiquement en corpus. La statistique multidimensionnelle permet d'obtenir des pôles multiples, positifs et négatifs, correspondant à des constellations de traits linguistiquement corrélés.

Généralement, l'architecture d'un tel traitement comprend les étapes suivantes [Habert *et al.*, 2000] :

1. Etiquetage morpho-syntaxique / analyse syntaxique des données textuelles.

2. Analyse typologique des données textuelles étiquetées/annotées : regroupement, dégroupement, transformations, enrichissement, omissions de certains traits.

3. Etablissement de catégories correspondant aux traits linguistiques retenus comme pertinents.

4. Analyse statistique de la matrice des fréquences de chaque trait dans chaque texte : l'individus statistique donnant lieu à des comptages pour chaque case du tableau sera l'occurrence d'un trait retenu par l'analyse typologique.

La partie centrale de la démarche réside dans la préparation des données pour l'exploration typologique. Les opérations de regroupement, contraste, etc. sont intimement liées à l'étiquetage et le marquage. Cette partie est pourtant souvent difficile à formaliser de façon univoque : la définition des traits est une tâche particulièrement complexe.[7] Comme le note S. Branca-Rosoff [1999, p. 16] au sujet des travaux de D. Biber, le jugement subjectif du chercheur ainsi que la lignée de travaux dans laquelle s'inscrivent ces recherches sont essentiels dans l'interprétation des résultats de l'exploration typologique qui repose sur le marquage linguistique :

*« Le choix des paramètres est cependant posé comme naturel; c'est-à-dire qu'il n'est pas explicité par l'analyste. Or, les traits retenus n'ont rien "d'objectif". Leur sélection résulte des hypothèses du chercheur et de la tradition, plus précisément des bases typologiques qu'il juge intéressantes. »* [*ibid*]

### 4.2. Catégorisations supervisées

Dans le cadre des catégorisations supervisées, on isole d'abord des catégories de textes pour définir ensuite les traits spécifiques de chacune d'entre elles. Les spécificités des catégories pré-établies (genres, domaines, contextes de production textuelle) sont étudiées en observant des données langagières rassemblées à des fins de l'étude. Ce sont des catégories intuitives établies par les locuteurs qui sont utilisées pour amorcer l'exploration typologique.[8]

Les « variables pertinentes » récoltées à l'issue de l'analyse sur corpus sont attribuées aux documents catégorisés d'un *corpus d'apprentissage*. Inutile de rappeler que la principale fragilité de ces variables réside dans le fait qu'elles ne sont que très peu généralisables à d'autres ensembles des données textuelles et ne sont d'actualité qu'au moment de la constitution du corpus (dû au développement mouvementé des pratiques sociales, langagières, etc.) [Habert *et al*., 1998, p. 40-43].

L'émergence des machines à vecteur de support (*SVM*) a largement contribué au développement des catégorisations supervisées sur les données textuelles [Joachims, 1998].[9]

L'objectif de ces méthodes consiste à approximer la fonction de catégorisation exacte : on associe à chaque couple (document textuel/catégorie) une valeur (vraie ou fausse), en fonction de l'appartenance ou non de ce document à une classe donnée.

Très sommairement, on pourrait décomposé ce processus en trois phases :

1. La phase d'apprentissage qui a pour objectif d'induire une *fonction de catégorisation* à partir d'un *jeu d'entraînement* (un jeu de documents dont la catégorie est connue *a priori*).

2. La phase de validation qui a pour objectif d'ajuster automatiquement les paramètres requis par des algorithmes d'apprentissage.[10]

3. La phase de test qui permet d'évaluer le processus de catégorisation grâce au *jeu de test* (un ensemble de couples document textuel/catégorie dont l'information associée à la catégorie n'est utilisée que pour l'évaluation finale) [Jaillet, 2004].

Après une catégorisation manuelle d'un jeu de test, les affectations fournies par le système sont comparées aux affectations préalablement définies par un expert.[11] L'objectif du processus de catégorisation consiste à maximiser une fonction de score basée sur le *rappel* et la *précision* [Sebastiani, 2002, p. 41-47].

Dans ce type d'approche, c'est le *formalisme vectoriel* qui est utilisé pour représenter les textes. Chaque dimension de l'espace vectoriel correspond à un *descripteur* (appelé aussi *terme*), extrait du jeu d'apprentissage. Pour réduire le nombre de dimensions potentiellement élevé, on passe par un prétraitement linguistique souvent problématique du point de vue de la cohésion discursive. Les descripteurs linguistiques sont décorrélés et mis à plat dans les décomptes. L'éparpillement d'indices au niveau local rend particulièrement délicat l'interprétation globale. Au final, les vecteurs statistiques obtenus sont composés par les occurrences de traits relevés dans les documents analysés : on dispose d'un tableau *descripteurs x individus* pour chaque texte (la valeur des descripteurs et la catégorie sont connus pour chaque texte du corpus d'apprentissage).

Dans les travaux en cours, les dimensions de l'espace vectoriel sont parfois associées à des concepts prédéfins recensés dans les ressources dictionnairiques, thésaurus, ontologies, etc. Par exemple, le vecteur conceptuel du verbe « activer » sera défini par les concepts *X, Y, Z,…* répertoriés dans un thésaurus. Dans ce type d'approche, on cherche à établir un profil sémantique global d'un document textuel en projetant ce dernier sur un ensemble de concepts.

Cette projection vise à déduire les « champs sémantiques » du texte en question [Jaillet *et al*., 2003]. Les recherches en cours montrent que ce type de représentation doit être manipulé avec précaution :

*« Dès lors qu'on parle de « vrais » textes, de « vrais » corpus, et pas simplement de phrases d'exemples artificiellement construites en dehors d'un contexte linguistique et pragmatique, il convient de se rendre compte que la dimension interprétative personnelle fait qu'il n'y a pas de consensus évident sur ce qu'est ou n'est pas le sens d'un texte. ... Le sens n'est pas de nature symbolique ; c'est un processus sémiotique au centre de l'activité de l'interprétant qui est complexe, notamment parce qu'il est réflexif. »* [Beust et Roy, 2006, p. 58].

***

De façon générale, le point faible des catégorisations appliquées aux objets textuels réside dans le fait qu'il est particulièrement difficile de démontrer si une classification obtenue est justifiée, si les descripteurs identifiés contribuent effectivement à discriminer les ensembles d'objets textuels visés :

*« La classification repose sur des représentations des données (objets) très classiques dont les limites de traitement sont connues et souvent très consommateurs de ressources en temps de calcul. La sémantique des objets doit être très clairement définie, non ambiguë. Cette même sémantique doit être typique pour pouvoir trouver les classes d'objet c'est-à-dire proche de relations de type présence/absence. Un biais de traitement se produit dans le cas d'objets ayant un contexte sémantique complexe comme des objets du langage naturel. Les problèmes émanent de la fréquence multicontextuelle des termes et de l'analyse inverse du processus de classification souvent impossible. »* [Turenne, 2000, p. 16]

### 4.3. Analyse de la distance intertextuelle

L'existence de la littérature très riche consacrée aux problèmes de comparaison des ensembles textuels témoigne de l'importance et de la complexité méthodologique de ces questions.[12]

Il existe différentes manières de calculer des proximités/oppositions entre textes. Dans le domaine de la statistique textuelle ou *textométrie*, les approches quantitatives ont permis de développer plusieurs méthodes de calcul de la distance intertextuelle qui pourrait être mesurées, par exemple, à partir de la comparaison globale de stocks lexicaux [Lebart et Salem, 1994] ; [Salem, 2006] ; [Brunet, 1998 ; 2003 ; 2004].

Les typologies appuyées sur différentes méthodes de calcul (la distance du *chi-deux*, l'*indice de Jaccard*) partagent des caractéristiques communes malgré les différences notables dans les méthodes de calcul [Salem, 2006].

Les expériences sur corpus montrent que les différents types d'unités textuelles mobilisées (*formes, lemmes, POS*, etc.) ont peu d'incidence sur les rapprochements/oppositions textuels observés [Lebart et Salem, 1994] ; [Brunet, 2000]. Une typologie des éléments textuels rassemblés ne reflète que les dimensions de variation les plus importantes. Ainsi, le changement de l'unité de décompte ou de la méthode de calcul n'influence que très peu les résultats obtenus [Kastberg Sjöblom, 2006].

En revanche, le vrai souci méthodologique réside dans le fait qu'une typologie ne résume pas la diversité inhérente propre à chaque collection de textes. Les différences qui existent nécessairement entre les textes sur le plan lexical, syntaxique, sémantique, pragmatique, etc. sont souvent complexes. L'objectif qui parait réaliste du point de vue de l'analyse automatique consisterait à convoquer des méthodes permettant d'organiser les parties d'un corpus en sous-ensembles qui relèvent *a posteriori* une certaine cohérence et qui rapprochent des textes qui relèvent d'un même genre, période, auteur, etc. :

*« À supposer qu'on puisse avec sûreté répartir les textes dans l'espace, comme on distribue les villes sur une carte, il resterait à décrire et à expliquer les oppositions et les rapprochements. »* [Brunet, 2003].

Face aux incertitudes des typologies textuelles basées sur des relevés d'indices, des descripteurs isolés, des dénombrements éparpillés d'unités textuelles de toute sorte, une voie de recherche complémentaire consisterait à s'intéresser aux procédures permettant de considérer le texte dans son intégrité, comme une structure ordonnée dont les régularités se manifeste au plan syntagmatique et paradigmatique.

Les comparaisons intertextuelles de structures récurrentes dans les corpus qui autorisent des comparaisons du point de vue de leurs contextes de production (*corpus parallèles* ou *comparables*, *séries textuelles chronologiques*, etc.) attire l'attention sur la circulation d'unités textuelles étendues qui reflètent l'existence de concepts partagés par la communauté produisant ces textes [Salem, 1987]. Ces formations discursives communes pourraient être repérées à travers l'analyse des répétitions segmentales partagées par un groupe de textes [Salem, 2006] ou encore par l'analyse de structures lexicales complexes qui se réalisent, par exemple, dans les réseaux de concurrences chaînées [Martinez et Zimina, 2002] ; [Veronis,

2003] ; [Leblanc et Martinez, 2005]. Plusieurs travaux ont abordé ces questions de repérages/mise en relation de structures textuelles spécifiques repérables au fil des textes [Serant et Thoiron, 1992] ; [Zweigenbaum et Habert, 2006] ; [Pierard et Bestgen, 2006]. Appuyés par de nouveaux développements logiciels, ces recherches sont en train de s'étendre à l'étude des phénomènes de *résonances textuelles* et intertextualité [Lamalle et Salem, 2002] ; [Salem, 2004 ; 2006] ; [Zimina, 2005] ; [Brunet, 2006].

## 5. Perspectives

Quels seront les axes de recherches consacrées aux questions des typologies textuelles dans les années à venir ? A l'issue de ce survol bibliographique des recherches en cours, nous constatons que la volonté de typer les textes, de tracer des parcours prédéfinis pour certains groupes d'utilisateurs commence à montrer ses limites lorsqu'elle est confrontée à la dynamique d'exploration textuelle interprétative, centrale dans la compréhension humaine du sens d'un texte. Les recherches en typologies textuelles se rapprochent alors de plus en plus des questions de topologies/topographies textuelles. Les questions de visualisation, de *cartographie*, de *navigation interactive* dans les vastes ensembles textuels sont à l'ordre du jour et les développements logiciels connexes se multiplient.[13]

## Références :


ADAM, J.-M. (1992). *Les textes : types et prototypes - Récit, description, argumentation, explication et dialogue*. Paris : Nathan.

ADAM, J.-M. (1999). *Linguistique textuelle : Des genres de discours aux textes*. Paris : Nathan.

ARISTOTE (1932). *Rhétorique I*, texte établi et traduit par M. Dufour. Paris : Les Belles Lettres.

ARGAMON, SH., KARLGREN, J., SHANAHAN, J. G. (Eds.) (2005). « Stylistic Analysis Of Text For Information Access. » *Papers from the workshop held in conjunction with the 28th Annual International ACM Conference on Research and Development in Information Retrieval*, Salvador, Bahia, Brazil, August 13-19, 2005. Disponible sur : http://www.sics.se/jussi/Artiklar/2005_SIGIR_Salvador/Style2005/style2005.pdf

AUSSENAC-GILLES, N., CONDAMINES, A., SZULMAN, S. (2002). « Prise en compte de l'application dans la constitution de produits terminologiques. » In J. Le Maitre (Eds.), *Information, Interaction, Intelligence : Actes des 2e Assises Nationales du GDR I3*, Cépaduès Editions, p. 289-303. Disponible sur : http://www.irit.fr/GDR-I3/fichiers/assises2002/papers/17-ConstitutionDeProduitsTerminologiques.pdf

BAKHTINE, M. (1929/1977). *Marxisme et philosophie du langage*. Paris : Minuit.

BAKHTINE, M.(1984). *Esthétique de la création verbale*. Paris : Gallimard.

BEACCO, J-C. (1992) « Les genres textuels dans l'analyse du discours. » *Langages* n. 105, « Ethnolinguistique de l'écrit ».

BEAUDOUIN, V., FLEURY, S., HABERT, B., ILLOUZ, G., LICOPPE, C., PASQUIER, M. (2001). « TyPWeb : décrire la Toile pour mieux comprendre les parcours. » In Actes du *Colloque International sur les Usages et les Services des Télécommunications, e-Usages CIUST'01*, Paris, 12-14 juin 2001.
Disponible sur : http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/Articles/typweb.pdf

BEAUVISAGE, Th. (2004). « Sémantique des parcours des utilisateurs sur le web. » *Thèse de Doctorat en Sciences du langage*. Université de Paris X – Nanterre. Disponible sur : http://www.revue-texto.net/Inedits/Beauvisage/Beauvisage_Parcours.html

BENAMARA, F., MOENS, M-F., SAINT-DIZIER, P. (Eds.) (2005). *Proceedings of IJCAI05 Workshop « Knowledge and Reasoning for Answering Questions. »* Edinburgh, July 30, 2005. Disponible sur : http://www.irit.fr/recherches/ILPL/kraq05V1.pdf

BEN ROMDHANE, M. (2001). « Navigation dans un espace textuel, accès à l'information scientifique ». *Thèse en Sciences de l'Information et de la Communication*. Université Jean Moulin Lyon 3. Disponible sur : http://www.recodoc.univ-lyon1.fr/these_MBR.pdf

BEUST, P., ROY, Th. (2006). « Prendre en compte la dimension globale d'un corpus dans la contextualisation du sens : expérimentations en informatique linguistique. » *GLOTTOPOL* n. 8 « Traitements automatisés des corpus spécialisés : contextes et sens. » Juillet 2006. Disponible sur : http://www.univ-rouen.fr/dyalang/glottopol/telecharger/numero_8/gpl8_05beust_roy.pdf

BIBER, D. (1988). *Variations across speech and writing*. Cambridge : Cambridge University Press.

BIBER, D. (1995). *Dimensions of register variation*. Cambridge : Cambridge University Press.

BIBER, D. (2004). *Conversation text types: A multi-dimensional analysis*. In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles JADT'04*, Louvain La Neuve, 10-12 mars, 2004, p. 15-34.
Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_000.pdf

BOMMIER-PINCEMIN B. (1999). «Diffusion ciblée automatique d'informations : conception et mise en oeuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents. » *Thèse de Doctorat en Sciences du Langage*. Université Paris 4.
Disponible sur : http://www.revue-texto.net/Inedits/Pincemin/Pincemin_these.html

BRANCA-ROSOFF, S. (1999). « Types, modes et genres entre langue et discours. » *Langage et société* n. 87, p. 5–24. Disponible sur : http://www.cavi.univ-paris3.fr/Ilpga/ed/dr/drsb/sb-pdf/intro-Branca-LS87.pdf

BRONCKART, J.-P. (1996). « Genres de textes, types de discours et opérations psycholinguistiques. » *ENJEUX* n. 37/38, p. 31-47.

BRUNET, É. (1988). « Une mesure de la distance intertextuelle : la connexion lexicale. » *Revue Informatique et Statistique dans les Sciences humaines RISSH* n. 24(1-4), p. 81-116.

BRUNET, É. (2000). « Qui lemmatise, dilemme attise. » Revue électronique *Lexicometrica* 2. Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/article/numero2/brunet2000.PDF

BRUNET, É. (2003). « Peut-on mesurer la distance entre deux textes ? » *CORPUS* n. 2 « La distance intertextuelle. » Décembre 2003. Disponible sur : http://corpus.revues.org/document30.html

BRUNET, É. (2004). « Où l'on mesure la distance entre les distances. » *Texto ! Dits et Inédits*. Mars 2004. Disponible sur : http://www.revue-texto.net/Inedits/Brunet/Brunet_Distance.html

BRUNET, E. (2006). «Navigation dans les rafales. » In *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles JADT'06*, Besançon, 19-21 avril 2006. Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_EB.pdf

CEHEUX, G. R. (nom collectif) (2002). « Stratégie pour l'interprétation de documents. » In *Actes des deuxièmes assises nationales du GdR I3 (Information-Interaction-Intelligence)*, p. 275-288. Disponible sur : http://www.irit.fr/GDR-I3/fichiers/assises2002/papers/16-InterpretationDeDocuments.pdf

CHARAUDEAU, P. (1997). « Les conditions d'une typologie des genres télévisuels d'information. » *Réseaux* n. 81, CNET. Disponible sur : http://www.enssib.fr/autres-sites/reseaux-cnet/81/05-chara.pdf

CHARAUDEAU, P., MAINGUENEAU, D. (2002). *Dictionnaire d'analyse du discours*. Paris : Seuil.

CHAROLLES, M. (1988). « Les plans d'organisation textuelle, périodes, chaînes, portées et séquences. » *Pratiques* n. 57, Metz, p. 3-14.

CLAVIER, V. (2006). «Le genre comme point d'accès au document : analyse comparée de textes scientifiques en mécanique et linguistique. » In *Contributions à la Journée ATALA « Typologie de textes pour le traitement automatique »*, Paris, 9 décembre, 2006. Disponible sur : http://www.atala.org/article.php3?id_article=312

COPECK, T., BARKER, K., DELISLE, S., SZPAKOWICZ, S. (2000). « Automating the Measurement of Linguistic Features to Help Classify Texts as Technical. » In *Actes de la 7ème conférence annuelle sur le Traitement Automatique des Langues Naturelles TALN'2000*, Lausanne, 16-18 octobre, 2000.
Disponible sur : http://www.cs.utexas.edu/~kbarker/papers/taln00-tt.pdf

COSTA, R. (2005). « Texte, terme, contexte. » In *Actes des 7es Journées scientifiques du Réseau de chercheurs 'Lexicologie, Terminologie, Traduction'*, Bruxelles, 8-10 septembre 2005, p. 79-87. Disponible sur : http://perso.univ-lyon2.fr/~thoiron/JS%20LTT%202005/pdf/Costa.pdf

CROWSTON K., KWASNIK B. (2004). « A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality. » In *Proceedings of the 37th Hawaii International Conference on System Sciences – 2004*. Disponible sur : http://csdl2.computer.org/comp/proceedings/hicss/2004/2056/04/205640100a.pdf

DEWDNEY, N., VANESS-DYKEMA, C., MACMILLAN, R. (2001). « The form is the substance: Classification of genres in text. » In *Proceedings of the ACL Workshop on Human Lanugage Technology and Knowledge Management*, Toulouse, France, 6-7 Juillet 2001.

ERTZSCHEID, O. (2002). « L'hypertexte : haut lieu de l'intertexte. » *Revue des ressources* [revue électronique publié par *Points de Fuite multimédia*]. Octobre 2002. Disponible sur : http://www.larevuedesressources.org/article.php3?id_article=27

FINN, A., KUSHMERICK, N. (2003). « Learning to classify documents according to genre. » In *Proceedings of IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, August 10, 2003, Acapulco, Mexico.
Disponible sur : http://www.aidanf.net/publications/finn03learninggenre.pdf

FLØTTUM, K., HOLM, H-V. (Eds.) (1999). « Polyfoni. » *TRIBUNE* n.9, Skriftserie for Romansk Institutt, Universitetet i Bergen. Disponible sur : http://www.hum.au.dk/romansk/polyfoni/

GENETTE, G. (1987). *Seuils*. Paris : Éditions du Seuil.

GREIMAS, A.-J. (1966). *Sémantique structurale*. Paris : Larousse.

HABERT, B., ILLOUZ, G., LAFON, P., FLEURY, S., FOLCH, H., HEIDEN, S., PREVOST, S. (2000). « Profilage de textes : cadre de travail et expérience. » In *Actes des Journées internationales d'Analyse des Données Textuelles JADT'07*, Lausanne, 9-11 mars 2000. Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/56/56.pdf

HABERT, B., FABRE, C., ISSAC, F. (1998). *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques.* Paris : Masson.

HABERT, B., NAZARENKO, A., SALEM, A. (1997). *Les linguistiques de corpus.* Paris : Armand Colin / Masson.

HABERT, B., SALEM, A. (1995). « L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles. » *TAL* n. 36(1–2), p. 249–276.

HERNANDEZ, N. (2004). « Description et Détection Automatique de Structures de Textes. » *Thèse de Doctorat en Informatique.* Université Paris-Sud.

HOLZEM M., DIONISI, D., LABICHE, J., TRUPIN, E. (2005). « Le document dans son agir organisationnel : le modèle de l'organisation dans l'interaction usager système. » In *Actes du huitième colloque international sur le document électronique CIDE'05*, Beyrouth, Liban, 25-28 mai 2005, p. 133-151.
Disponible sur : http://archivesic.ccsd.cnrs.fr/docs/00/06/25/90/PDF/sic_00001381.pdf

ILLOUZ, G. (1999a). « Méta-Étiqueteur Adaptatif : Vers une utilisation pragmatique des ressources linguistiques. » In *Actes de l'Atelier thématique TALN'99 : Corpus et Traitement Automatique des Langues : pour une réflexion méthodologique*, Cargèse, 12-17 juillet 1999.
Disponible sur : http://www.limsi.fr/Individu/gabrieli/CV/Publis/Articles/taln99.pdf

ILLOUZ, G., HABERT, B., FLEURY, S., FOLCH, H., HEIDEN, S., LAFON, P. (1999b). « Maîtriser les déluges de données hétérogènes. » In *Actes de l'Atelier thématique TALN'99 : Corpus et Traitement Automatique des Langues : pour une réflexion méthodologique* Cargèse, 12-17 juillet 1999. Disponible sur :
http://www.limsi.fr/Individu/gabrieli/CV/Publis/Articles/taln99-typweb.pdf

ILLOUZ, G. (2000a). « Vers un apprentissage en TALN dépendant du type de Texte. » In *Actes de la 7ème conférence annuelle sur le Traitement Automatique des Langues Naturelles TALN'2000*, Lausanne, 16-18 octobre 2000. Disponible sur :
http://www.limsi.fr/Individu/gabrieli/CV/Publis/Articles/FinaleTALN2000illouz.pdf

ILLOUZ, G., HABERT, B., FLEURY, S., FOLCH, H., HEIDEN, S., LAFON, P., PREVOST, S. (2000b). « Typtex : Generic Features for Text Profiler. » In *Proceedings of 6th RIAO Conference on "Content-Based Multimedia Information Access"*, Paris, 12-14 avril 2000. Disponible sur : www.limsi.fr/Individu/gabrieli/CV/Publis/Articles/RIAO2000-TyPTex.ps.gz

JAILLET, S. (2004). « Catégorisation automatique de documents. » In *Actes de la douzième session des journées des doctorants DOCTISS'04*, ED « Information, Structures et Systèmes », Université Montpellier 2. Disponible sur : http://www.lirmm.fr/doctiss04/art/I02.pdf

JAILLET, S., TEISSEIRE, M., CHAUCHE, J., PRINCE, V. (2003). « Classification automatique de documents. Le coefficient des deux écarts. » In *Actes du XXIème Congrès INFORSID'2003*, Nancy, 3-6 juin 2003. p. 87-102. Disponible sur : http://134.214.81.35/articles/a447c1bguEFeLWdKY.pdf

JALAM, R. (2003). « Apprentissage automatique et catégorisation de textes multilingues. » *Thèse de Doctorat en Informatique.* Université Lumière Lyon 2. Disponible sur : http://www.agrocampus-rennes.fr/math/jalam/these/these_radwan.pdf

JOACHIMS, T. (1998). «Text categorization with support vector machines: learning with many relevant features. » In *Proceedings of 10th European Conference on Machine Learning ECML-98*, p. 137–142.
Disponible sur : http://www.cs.cornell.edu/People/tj/publications/joachims_98a.ps.gz

KARLGREN, J., CUTTING, D. (1994). « Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. » In *Proceedings of the 15th International Conference on Computational Linguistics COLING'94*, p. 1071-1075.
Disponible sur : http://eprints.sics.se/56/01/cmplglixcol.pdf

KASTBERG SJÖBLOM, M. (2006). « La variation typologique : analyse systématique d'un corpus québécois. » In *Contributions à la Journée ATALA « Typologie de textes pour le traitement automatique »*, Paris, 9 décembre 2006.
Disponible sur : http://www.atala.org/article.php3?id_article=312

KESSLER, B., NUNBERG, G., SCHÜTZE, H. (1997). « Automatic Detection of Text Genre. » In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics.* Disponible sur : http://acl.ldc.upenn.edu/P/P97/P97-1005.pdf

LAMALLE, C., SALEM, A. (2002). «Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. » In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles JADT'2002*, Saint-Malo, 13-15 mars 2002. Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/lamalle_salem.pdf

LEBART, L., SALEM, A. (1994). *Statistique textuelle*. Paris : Dunod.

LEBLANC J-M., MARTINEZ, W. (2005). « Positionnements énonciatifs dans les vœux présidentiels sous la cinquième République ». *Corpus n. 4* « Les corpus politiques : objet, méthode et contenu. » Décembre 2005. Disponible sur : http://corpus.revues.org/document347.html

LEE, D. (2001). « Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. » *Language Learning & Technology* n. 3, vol. 5, p. 37-72. Disponible sur : http://llt.msu.edu/vol5num3/lee/

LIAO, C., ALPHA, Sh., DIXON, P. (2003). « Feature Preparation in Text Categorization. » In *Proceedings of the Second Australasian Data Mining Conference AusDM03*, Canberra, Australia (in conjunction with the 2003 *Congress on Evolutionary Computation CEC 2003*), 8 December 2003. Disponible sur : http://www.oracle.com/technology/products/text/pdf/feature_preparation.pdf

LUNDQUIST, L., MINEL, J-L., COUTO, J. (2006). « NaviLire, Teaching French by Navigating in Texts. » In *Proceeding of Eleventh International Conference Information Processing and Management of Uncertainty in Knowledge-based Systems IMPU'2006*, Paris, July 2-7, 2006. p. 1093-1099. Disponible sur : http://halshs.archives-ouvertes.fr/docs/00/09/78/49/PDF/Lundquist_Minel_Couto_IPMU06.PDF

LUONG, X. (dir.) (2003). *CORPUS* n. 2 « La distance intertextuelle. » Décembre 2003. Disponible sur : http://corpus.revues.org/sommaire52.html

LUŠTREK, M. (2006). *Overview of automatic Genre Identification.* Technical Report. Jozef Stefan Institute, Slovenia. Disponible sur : http://dis.ijs.si/mitjal/documents/Overview_of_Automatic_Genre_Identification-TR-06.pdf

MALRIEU, D. (2001). « Stylistique et Statistique textuelle: À partir de l'article de C. Muller sur les "pronoms de dialogue" .» *Texto ! Dits et Inédits.* [Revue électronique en ligne]. Disponible sur : http://www.revue-texto.net/Inedits/Malrieu_Stylistique.pdf

MALRIEU D. (2004). « Linguistique de corpus, genres textuels, temps et personnes. *Langages* n. 153, p. 73-85. Disponible sur : http://infolang.u-paris10.fr/modyco/textes/malrieu/DM_Genres_temps_personnes_03.pdf

MALRIEU, D., RASTIER, F. (2001). « Genres et variations morphosyntaxiques. » *TAL* n. 2, vol. 42, p. 548-577. Disponible sur : http://www.revue-texto.net/Inedits/Malrieu_Rastier/Malrieu-Rastier_Genres1.html

MANGENOT, F. (1996). « Outils textuels pour l'apprentissage de l'écriture en L1 et en L2. » In *Actes du Xe colloque international FOCAL 'Pratiques discursives et acquisition des langues étrangères'*, Besançon, 19-21 septembre 1996, p. 515-525. Disponible sur : http://w3.u-grenoble3.fr/espace_pedagogique/focal.htm

MARANINCHI, L. (2001). « Etude de marqueurs linguistiques pour une interprétation sémantique du référent du pronom sujet 'nous' ». *Mémoire de maîtrise en Sciences du Langage*, mention Industrie de la Langue. Université de Paris III, Sorbonne Nouvelle. Disponible sur : http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/sitespp/maitrise-2001/laetitia/memoireLM2001.pdf

MARSHMAN, E. (2003). « Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie. » Étude réalisée pour le groupe *ÉCLECTIK*. L'Observatoire de linguistique Sens-Texte, Université de Montréal. Disponible sur : http://www.olst.umontreal.ca/pdf/terminotique/corpusnormes.pdf

MARTINEZ, W., ZIMINA, M. (2002). « Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues. » In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles JADT'2002*, Saint-Malo, 13-15 mars 2002, p. 495-506. Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/martinez_zimina.pdf

MEYER ZU EISSEN, S., STEIN, B. (2004). « Genre Classification of Web Pages: User Study and Feasibility Analysis. » In Biundo S., Fruhwirth T., Palm G. (Eds.), *Advances in Artificial Intelligence*. Berlin : Springer, p. 256-269. Disponible sur : http://www-ai.upb.de/aisearch/ki04-frame.pdf

MOIRAND, S. (2003). « Quelles catégories descriptives pour la mise au jour des genres du discours ? » In *Contributions à la Journée d'études : les genres de l'oral*, le 18 avril 2003, Université Lumière – Lyon 2. Disponible sur : http://icar.univ-lyon2.fr/Equipe1/actes/Journee_Genre/Moirand_cat_genres.rtf

NAZARENKO, A. (2003). *Projet ExtraPloDocs : Rapport d'analyse de l'existant et des besoins.* Consortium LIPN, MIG, ISOFT. Rapport 2.1 et 2.2. Disponible sur : http://www-lipn.univ-paris13.fr/~poibeau/Extra/D21.pdf

OUERFELLI, T., LALLICH-BOIDIN, G. (2000). « Pratiques d'indexation dans les Bases Textuelles Structurées : Application aux Textes Techniques sous Format HTML. » In *Travaux du 28e congrès annuel de l'Association canadienne des sciences de l'information*. Disponible sur : http://www.slis.ualberta.ca/cais2000/ouerfelli.htm

PERY-WOODLEY, M-P. (1995). « Quels corpus pour quels traitements automatiques ? » *TAL* n. 36(1-2), p. 213-232. Disponible sur : http://w3.univ-tlse2.fr/erss/membres/pery/articles/TALcorpus.pdf

PERY-WOODLEY, M-P. (2000). « Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle. » Mémoire d'HDR, In *Carnets de grammaire* n. 8, juillet 2000, Université de Toulouse-LeMirail : ERSS. Disponible sur : http://w3.univ-tlse2.fr/erss/membres/pery/articles/habilmppw.pdf

PERY-WOODLEY, M-P. (2001). « Modes d'organisation et de signalisation dans des textes procéduraux. » *Langages* n. 141, p. 28-46. Disponible sur : http://w3.univ-tlse2.fr/erss/membres/pery/articles/lang00.pdf

PETITJEAN, A. (1989). « Les typologies textuelles. » *Pratiques* n. 62, p. 86-125.

PETITJEAN, A. (2005). « Pour une problématisation linguistique de la notion de genre textuel. » In *Actes du VIe Congrès des Romanistes Scandinaves*, Copenhague, 24-27 août 2005. Disponible sur : www.ruc.dk/isok/skriftserier/XVI-SRK-Pub/MOL/MOL07-Petitjean/

PIERARD, S, BESTGEN, Y. (2006). «A la pêche aux marqueurs linguistiques de la structure du discours. » In *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles JADT'06*, Besançon, 19-21 avril 2006. Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/PDF/II-067.pdf

POIBEAU, T., NAZARENKO, A. (1999). « L'extraction d'information, une nouvelle conception de la compréhension de texte ? » *TAL* n. 40(2).

POUDAT C. (2006). «Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres. » *Thèse de Doctorat en Sciences du langage*. Université d'Orléans. Disponible sur : http://www.revue-texto.net/Corpus/Publications/Poudat/

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

RASTIER, F. (2006) « De l'origine du langage à l'émergence du milieu sémiotique. » *Marges linguistiques* n. 11. Disponible sur : http://marg.lng1.free.fr/documents/13_ml112006_rastier_f/13_ml112006_rastier_f.pdf

RATLIFF, E. (2006). « Me Translate Pretty One Day. » *WIRED* n. 14.12, December 2006.

Disponible sur : http://www.wired.com/wired/archive/14.12/translate.html?pg=1&topic=translate&topic_set=

RAUBER A., MULLER-KOGLER, A. (2001). « Integrating automatic genre analysis into digital libraries. » In *Proceedings of the First ACM-IEEE Joint Conference on Digital Libraries*. ACM Press, Roanoke, Virginia. Disponible sur : http://www.ifs.tuwien.ac.at/ifs/research/pub_pdf/rau_jcdl01.pdf

ROULET, E. (1991). « Une approche discursive de l'hétérogénéité discursive. » *Etudes de linguistique appliquée* n. 83, p. 117-130.

SALEM, A. (1987). *Pratique des segments répétés : essai de statistique textuelle*. Paris : Klincksieck.

SALEM, A. (2004). « Introduction à la résonance textuelle. » In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles JADT'04*, Louvain La Neuve, 10-12 mars 2006. Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_096.pdf

SALEM, A. (2006). «Proximités segmentales. » In *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles JADT'06*, Besançon, 19-21 avril 2006. Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/PDF/II-075.pdf

SANTINI, M. (2003). « Identifying Genres on the Web: PhD Thesis Outline. » *ITRI report series*: ITRI-03-06, Brighton University, UK. Disponible sur : ftp://ftp.itri.bton.ac.uk/reports/ITRI-03-06.pdf

SANTINI, M. (2004). « State-of-the-art on Automatic Genre Identification. » *ITRI report series*: ITRI-04-03, Brighton University, UK. Disponible sur : ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-03.pdf

SANTINI, M. (2005). « Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically-Motivated Features. » *ITRI report series*: ITRI-05-02, Brighton University, UK. Disponible sur : ftp://ftp.itri.bton.ac.uk/reports/ITRI-05-02.pdf

SANTINI, M. (2006). « Some Issues in Automatic Genre Classification of Web Pages. » In *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles JADT'06*, Besançon, 19-21 avril 2006, p. 865-876. Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/PDF/II-077.pdf

SCHNEUWLY, B. (1994). « Genres et types de discours: considérations psychologiques et ontogénétiques. » In Y. Reuter (Ed.), *Les interactions lecture-écriture.* Bern : Lang, p. 155-174.

SEBASTIANI, F. (2002). «Machine learning in automated text categorization. » *ACM Computing Surveys* n. 34(1), p. 1-47. Disponible sur : http://br.endernet.org/~akrowne/elaine/dlib/papers/sebastiani/sebastiani_classification_survey.pdf

SEKINES, S. (1997). « The domain dependence of parsing. » In *Proceedings of the fifth Conference on Applied Natural Language Processing*, Washington, p. 96–102. Disponible sur : http://acl.ldc.upenn.edu/A/A97/A97-1015.pdf

SERANT, D., THOIRON, P. (1992). Topographie des formes répétées. *Lingua* n.3, vol. 87, p. 333-343.

SILVA, R. (2005). « Morphologie de spécialité : regard(s) sur le(s) contexte(s). » In *Actes des 7es Journées scientifiques du Réseau de chercheurs Lexicologie, Terminologie, Traduction*, ISTI, Bruxelles, 8-10 septembre 2005, p. 421-427. Disponible sur : http://perso.univ-lyon2.fr/~thoiron/JS%20LTT%202005/pdf/Silva.pdf

SINCLAIR, J., BALL, J. (1996). *EAGLES Preliminary Recommendations on Text Typology*. Birmingham University. EAG---TCWG---TTYP/P. June 1996. Disponible sur : http://www.ilc.cnr.it/EAGLES/texttyp/texttyp.html

SLODZIAN, M. (2000). « L'émergence d'une terminologie textuelle et le retour du sens. », In Bejoint H., Ph. Thoiron (dir.), *Le sens en terminologie*. Lyon : Presses Universitaires de Lyon, coll. « Travaux du C.R.T.T. », p. 61-85.

STAMATATOS, E., FAKOTAKIS, N., KOKKINAKIS, G. (2000). « Text genre detection using common word frequencies. » In *Proceedings of the 18th International Conference on Computational Linguistics*, Association for Computational Linguistics, volume 2, Luxembourg, p. 808–814.

SUGAR BOESE E. (2005). « Stereotyping the Web: Genre Classification of Web documents. » *Master's Thesis*. Computer Science Department, Colorado State University. Disponible sur : http://www.cs.colostate.edu/~boese/Research/masters.pdf

TURENNE, N. (2000). «Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles. » *Thèse de Doctorat en Informatique.* Université Louis-Pasteur Strasbourg/ENSAIS. Disponible sur : http://genome.jouy.inra.fr/~turenne/com_files/TheseTurenne.pdf

VAN DIJK, T.A. (1980). *Macrostructures*. Hilsdale, N.J.: Lawrence Erlbaum Associates, Publ.

VERONIS, J. (2003). « HyperLex : Cartographie lexicale pour la recherche d'informations. » *Rapport interne.* Equipe DELIC, Université de Provence. Disponible sur : http://www.up.univ-mrs.fr/veronis/pdf/2003-hyperlex-rapport.pdf

WERLICH, E. (1976). *A Text Grammar of English.* Heidelberg: Quelle & Meyer.

ZWEIGENBAUM, P., HABERT, B. (2006). « Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue. » GLOTTOPOL n. 8 « Traitements automatisés des corpus spécialisés : contextes et sens. » Juillet 2006. Disponible sur : http://www.univ-rouen.fr/dyalang/glottopol/telecharger/numero_8/gpl8_03zweigenbaum_habert.pdf

ZIMINA, M. (2005). « Topographie bi-textuelle et approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles », » In *Actes des 7es Journées scientifiques du Réseau de chercheurs 'Lexicologie, Terminologie, Traduction'*, Bruxelles, 8-10 septembre 2005, p. 175-186.
Disponible sur : http://perso.univ-lyon2.fr/~thoiron/JS%20LTT%202005/pdf/Zimina.pdf

**NOTES :**

[1] C'est la théorie littéraire qui a repris le thème de genres de la rhétorique antique. Dans son analyse de l'art rhétorique. Aristote distingue des *types de preuve* destinés à convaincre l'auditoire [Aristote, 1932]. Au cours des années, plusieurs approches de la problématique des typologies textuelles ont été considérées au sein des courants linguistiques différents. Par exemple, dans les année soixante-dix, Egon Werlich a élaborée une typologie basée sur des *phrases types* [Werlich, 1976]. On a vu ensuite se développer une réflexion sur la notion de *superstructure textuelle* [Van Dijk, 1980]. Dans les années quatre-vingt, Jean-Michel Adam a commencé à travailler sur la notion de *séquence élémentaire* (argumentative, narrative, descriptive et dialogale) [Adam, 1992]. Actuellement, le modèle typologique d'Adam est jugé opérationnel par beaucoup de linguistes.

[2] La notion de polyphonie se présente pour la première fois dans les travaux de Mikhaïl Bakhtine [1929]. L'idée de Bakhtine était d'envisager le texte comme une sorte d'espace de discussion ou d'affrontement de plusieurs voix (celle du locuteur, du discours social, de la sagesse populaire, etc.) d'où le principe de *dialogisme* qui contribue à la compréhension du texte et à son analyse.

[3] Citons aussi le travail de Maraninchi [2001] sur l'étude de marqueurs linguistiques : http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/sitespp/maitrise2001/laetitia/memoireLM2001.pdf

[4] L'analyse structurale des textes plonge ses racines au sein de la théorie sémantique développée, par exemple, dans les travaux de Greimas [1966].

[5] Selon M. Bakhtine [1984, tr. fr.], il existe des genres de discours différents, marqués par la spécificité d'une sphère d'échange. La variété des genres est inépuisable car elle liée à la variété des domaines. Bakhtine définit le genre de discours *premier* (simple) et le genre de discours *second* (complexe) au travers des circonstances d'un échange culturel. Le genre premier est formé au cours d'un échange verbal spontané. Par exemple, la réplique quotidienne est un genre premier formé dans la sphère d'échange de la vie quotidienne. Cette même réplique insérée dans une œuvre littéraire appartient au domaine de la vie littéraire (genre de discours second).

[6] Il est intéressant de citer dans ce contexte le travail récent de C. Poudat [2006] qui a élaboré un système de descripteurs linguistiques du genre de l'article scientifique de revue linguistique. Son travail aboutit à la mise en place d'un véritable observatoire du genre au niveau morphosyntaxique sur des corpus récoltés au cours de l'étude (plus de 220 articles de revues scientifiques parus autour de l'an 2000). Dans ces comptes-rendus d'expérience rigoureusement documentés, Poudat [2006, p. 26] remarque que l'ensemble des corpus mobilisés a fait l'objet d'un lourd traitement d'étiquetage, vérification et annotation spécifique.

[7] Comme le souligne B. Habert [2000], les traits ne doivent pas être trop fins car ils risqueraient alors de déboucher sur un éparpillement d'occurrences rendant impalpables les contrastes (temps des verbes). A l'inverse certains traits sont trop grossiers pour une analyse contrastive. Dans ce contexte, Habert [ibid] propose de réfléchir à la conception des traits structurées de manière à pouvoir utiliser tout ou partie des informations correspondantes.

[8] Lorsque l'on construit une typologie *a priori*, on considère que les genres (au moins certains d'entre eux) possèdent un ancrage social que l'on parvient à repérer par leurs spécificités formelles et discursives.

[9] L'état de l'art sur les différentes méthodes de classification est disponible, par exemple, dans Turenne [2000, p. 25-86] et Jalam [2003, p. 5-19].

[10] La majorité des algorithmes d'apprentissage dépend d'un ensemble de paramètres qu'il est nécessaire de fixer correctement. Par exemple, dans le cas des *arbres de décision*, un des paramètres

possibles correspond au choix du nombre maximum de feuilles de l'arbre qui définit la condition d'arrêt.

**11** L'attribution manuelle de catégories fixées à l'avance peut être remplacée par la répartition des documents d'un corpus d'apprentissage en classes par des méthodes de classification automatique (*clustering*). Dans les recherches en cours, on tente de mobiliser ce type d'approche sur la base de traits linguistiques fins (usage de pronoms, temps, modes, marqueurs lexicaux spécifiques) en essayant de capter des phénomènes de cooccurrence/associations de traits.

**12** Sur l'analyse de la distance intertextuelle, on consultera, par exemple, le numéro thématique de la revue *CORPUS* « La distance intertextuelle. » [Luong, 2003] et, particulièrement la contribution de J-P. Barthélémy, X. Luong et S. Mellet «Prenons nos distances pour comparer des textes, les analyser et les représenter » [Barthélémy *et al.*, 2003]:
http://corpus.revues.org/document25.html#tocto4

**13** Sur ces questions, on consultera, par exemple, [Lamalle et Salem, 2002] ; [Holzem *et al.*, 2005] ; [Beust et Roy, 2006] ; [Lundquist *et al.*, 2006].