Typologie de textes pour le traitement automatique

Maria Zimina-Poirot, Marie-Paule Jacques,
Thierry Poibeau, Françoise Gayral

LIPN - Université Paris 13 et CNRS

Dans la continuité de :

- Journée ATALA « Modéliser et décrire l'organisation discursive à l'heure du document numérique », La Rochelle, juin 2004
- « Colloque international : Discours et document », Caen, juin 2006
- Journées d'étude « Genre textuel/domaine/activité », Toulouse, octobre 2006

Questions partagées

- Les textes ne sont pas des suites de mots mais des discours

- Nécessité d'adapter les grilles d'analyse aux variations que présentent les discours
  - les variations relatives au... genre, type ?

Définir le genre

Convergences :

- fait partie de la compétence du locuteur,
- offre une « grille » pour la production et l'interprétation des textes,
- est lié à une situation de communication,
- dépend de facteurs socio-culturels,
- peut être dénommé *roman, poème, éditorial, essai, thèse, compte-rendu...*

Trosborg 1997, Branca-Rosoff 1999, Moirand 2003

Définir le type

Genre = type *vs* genre ≠ type

Divergences :

- critères divers :
  - intention du locuteur (persuader, ordonner, expliquer...)
  - propriétés cognitives, ex. différenciation et relation des perceptions dans l'espace => description ; différenciation et relation des perceptions dans le temps => narration
  - traits internes des textes

- lien avec des caractéristiques linguistiques

Pas de correspondance type/genre

La question du genre : pertinente aussi bien pour la linguistique que pour le TAL

- articulation genre / caractéristiques linguistiques internes des textes

  hypothèse de corrélations entre le genre du texte et un fonctionnement linguistique

- classement – automatique – des textes selon leur genre

  – tri / rangement des documents (RI, recherche documentaire)

  – élimination (= tri négatif)

  – sélection des traitements

Genre, type et TAL

Un enjeu pour certaines tâches de TAL est l'élaboration de traitements différenciés, adaptés justement au genre (au type ?) du texte à traiter.

Travaux montrent qu'appliquer le même traitement ou la même stratégie de traitement quel que soit le texte peut entraîner des variations de performances.

- Illouz 1999 : étiquetage morpho-syntaxique
- Frérot 2005 : stratégie de rattachement prépositionnel

Typologie inductive

- Travaux de D. Biber

  - regroupements sur des traits linguistiques internes puis interprétation des regroupements

➔ TAL : question de l'homogénéité linguistique des données (TypText)

➔ Linguistique : cerner en quoi des textes réputés de genres différents ou d'un même genre se ressemblent ou se distinguent

Relation genre / traits linguistiques

- Travaux de Malrieu & Rastier (2001), Malrieu (2004)

➔ par quelles caractéristiques linguistiques les genres se distinguent-ils ?

Si en tant que locuteurs nous reconnaissons des genres différents, alors il y a des différences perceptibles au plan linguistique, et si on met en évidence ces différences linguistiques, alors on valide la différence de genres.

- détermination des traits pertinents ;
- mise en évidence de corrélations positives et/ou négatives entre ces traits et le genre.

Classification automatique de textes

- classification thématique : approche lexicale bien au point
- classification en genre : nombreux travaux, se distinguent par :
  - les méthodes utilisées,
  - les traits qui sont pris en compte,
  - les genres à repérer : soit 'fournis' avec les textes (Brown Corpus, BNC), soit définis par les chercheurs.

Kessler, Nunberg et Schütze 1997 ; Argamon 1998 ; Stamatos, Fakotakis et Kokkinakis 2000 ; Dewdney, VanEss−Dykema et MacMillan 2001 ; Finn et Kushmerick 2003.

Les genres du Web

- quels genres ?
- appartenance d'un document à plusieurs genres,
- émergence de genres nouveaux,
- questions spécifiques posées par le Web :
  - place des métadonnées et du balisage pour l'identification du genre
  - quel niveau d'attribution du genre : page, site ou autre ?

Meyer zu Eissen et Stein 2004, Santini 2004.

Automatisation du repérage des traits

- si on veut automatiser, le calcul des traits doit être automatisable,
- est-ce qu'on ne se focalise pas sur des traits qui ont surtout le mérite d'être automatisables, au détriment d'autres qui seraient peut-être plus pertinents ?
- est-ce qu'ils permettent néanmoins de capter les fonctionnements linguistiques dont on veut rendre compte ?

Evaluation... une question récurrente

- pas de référence à l'aune de laquelle évaluer les travaux,
- absence d'unanimité sur une théorie des genres et sur un classement empirique de textes,
- pas d'inventaire des genres, pas d'accord sur les types,
- où placer des oppositions telles que :
  - narratif / argumentatif / descriptif / informatif (genre ? type ? fonction/séquence discursive ?)
  - impliqué / autonome – par rapport à la situation de production - (Bronckart)

Homogénéité *vs* hétérogénéité textuelle

- un document --> un genre : tract syndical, sujet d'examen, histoire drôle, chanson d'amour
- plus gros document ?
- rapport = compte-rendu + étude de cas + description d'un système + données techniques + procédures + ...
- sur quels critères délimiter des segments ? limite de la segmentation ?
- quelle prise en compte de la structuration du document (chapitres, sections, paragraphes) ?

Questions ouvertes (1)

- définition des différentes notions,
- manifestations linguistiques du genre,
- choix et calcul des traits,
- choix des méthodes,
- représentation des textes.

Questions ouvertes (2)

- paradoxe de l'outillage du repérage des traits,
- « viabilité » d'une analyse linguistique poussée,
- applications : conception d'outils intégrant la dimension « genre ».