Analyse stylistique différentielle à base de marqueurs et textométrie

Bénédicte Pincemin, CNRS
ICAR, ENS-LSH, université de Lyon

Journées d'étude « Le style et sa modélisation », Tours, 10-11 décembre 2009.

Plan

- Pourquoi et comment **automatiser**
- **Affinités** de la textométrie avec l'approche proposée
- Questions de **mise en oeuvre**
- **Fonctionnalités** des logiciels et possibilités d'analyse
- Limites actuelles et **pistes** d'innovation

Automatiser : réserves

- Lecture experte : la machine pourrait-elle mieux faire ?
  - L'imprévu (vs. langue contrôlée), qui est de mise en littérature
  - Interprétation :
    - activité non déterministe (par ex. deux lecteurs n'ont pas la même lecture)
    - appropriation (par ex. un même lecteur n'a pas la même lecture à deux moments différents)

Automatiser : des solutions écartées

- Boîte noire :
  - on veut *voir* et *choisir* (et peut-être *éditer* : composer, annoter) les marqueurs, les textes (et leurs éditions), les corpus
- Statique :
  - les statistiques peuvent être plus sensibles au contexte que l'extraction d'informations à base de patrons et d'ontologies
- Tout-automatique (question → réponse)
  - Plutôt : solution ouverte, qui amène à voir le texte sous d'autres angles et à construire une analyse
  - Le calcul produit un *résultat*, le résultat n'est pas encore une *réponse*

Automatiser : des apports

- Scientifiquement, la formalisation permet déjà explicitation (discussion) et rigueur
- Des raisons de faire le pas d'une (semi-)automatisation :
  - Déléguer la procédure machinale
  - Systématicité (cf. Toine)
  - Analyse à grande échelle ? (« appliquer à... ») car mémoire + vitesse
  - Mise à l'épreuve de la formalisation et dynamique de la modélisation (hypothèses rapides et multiples et non pesantes) : le but du jeu n'est pas de tout décrire mais de mieux décrire un aspect donné
- Objectif
  - Parcours efficace (méthodique et outillé) et synthèse
  - Une autre lecture : on ne remplace pas la lecture traditionnelle, mais on l'enrichit

Plan

- Pourquoi et comment **automatiser**
- **Affinités** de la textométrie avec l'approche proposée
- Questions de **mise en oeuvre**
- **Fonctionnalités** des logiciels et possibilités d'analyse
- Limites actuelles et **pistes** d'innovation

Affinités : place du texte

- Textométrie (statistique lexicale/lexicométrie, logométrie)
  - Connaissance du corpus vs. Exploitation d'un gisement d'informations
  - Retour au texte :
    - Affichage premier du texte
    - Contextualisation de résultats
- Il se pratique ailleurs des statistiques linguistiques mais pas nécessairement textuelles (possible en linguistique de corpus)

Affinités : approche différentielle

- Corpus de référence : contraste par rapport à un fond (et multiplicité des fonds possibles)
- Typage : opérationalisation de la structure *même* vs *différent*
- Tris : perception des répétitions et des divergences
- Statistiques : mesurent un écart
- Attention à la hiérarchisation, à l'ordre des résultats, plus qu'aux scores eux-mêmes
- Unités d'analyse :
  - Diversité des unités de départ envisageables
  - Les unités sont en fait reconstruites (ex. segments répétés) ; « dé-ontologie » (Rastier)

cf. (Pincemin 2010 - Topicos)

Plan

- Pourquoi et comment **automatiser**
- **Affinités** de la textométrie avec l'approche proposée
- Questions de **mise en oeuvre**
- **Fonctionnalités** des logiciels et possibilités d'analyse
- Limites actuelles et **pistes** d'innovation

Mise en oeuvre : global ↔ local

- Chercher des indices : local → global ?
- Choix de procédés à rechercher selon le corpus -ni tous, ni univoques : global (expertise humaine, genre) → local
- Mettre en évidence des caractéristiques discriminantes (corpus) et convergentes (complétude du référentiel ?)

Mise en oeuvre : global ↔ local

- Catégories ↔ marqueurs et indices
  - Chercher des indices puis les mettre en relation : local → global
  - Percevoir des indices renforçant une interprétation : global → local
- Description de cette interaction :
  - Convergence d'indices faibles (quantitativement) et dispersés (diversifiés), « réseau de cohérence » (Garric & Léglise 2005)
  - « présomption d'isotopie » (Rastier 1987)

Mise en oeuvre : local ↔ global et réseau de cohérence

...des indices quantitativement faibles peuvent devenir qualitativement remarquables. Parce qu'ils entrent dans un réseau de cohérence, ils peuvent relever de différents niveaux : un même effet discursif peut trouver ses marqueurs à partir de différentes unités, qui, prises individuellement, peuvent être peu nombreuses, mais saisies dans leur ensemble engendrent un marquage significatif.

(Document de travail des journées d'étude, p. 4)

Mise en oeuvre : local ↔ global et réseau de cohérence

[Sur le résultat d'un calcul de spécificités,] chercher des réseaux de cohérence, associer des fréquences pour construire des cohérences. Comme le notent S. Bonnafous et M. Tournier (1995 : 74), « le fréquentiel fait sens dans le constat de convergences expressives à l'intérieur des textes ». Ainsi par exemple, nous serons sensibles à des expressions liées à l'environnement comme *déchets* (+E16), *effet de serre* (+E12), *retraitement des déchets* (+E08), *réchauffement de la planète* (+E06), *toxiques* (+E06), *développement durable* (+E04) etc. Ces différentes formes construisent un **réseau de cohérence** important.

(d'après Garric & Léglise 2005)

Mise en oeuvre : local ↔ global et présomption d'isotopie

[Une conception textuelle de l'isotopie] conduit à un déplacement de problématique. En général, on considère l'isotopie comme une forme remarquable de combinatoire sémique, un effet de la combinaison des sèmes. Ici au contraire, où l'on procède paradoxalement à partir du texte pour aller vers ses éléments, l'isotopie apparaît comme un principe régulateur fondamental. Ce n'est pas la récurrence de sèmes déjà donnés qui constitue l'isotopie, mais à l'inverse la présomption d'isotopie qui permet d'actualiser des sèmes, voire les sèmes.

(Rastier, 1987, *Sémantique interprétative*, pp. 11-12)

Mise en oeuvre : Contextualisations

- Constitution du corpus
  - Faisabilité (disponibilité des textes,...)
  - Frontières : on retrouve la question, entière. Norme, objectivation (explicitation) mais pas neutralité.
- Genre et texte
  - Si le genre détermine la lecture du texte, peut-on avoir un « référentiel texte » défini indépendamment du genre du texte ?
- Passages, contextes : leur multiplicité

Mise en oeuvre : identification d'indices

- Classes fermées, peu ou pas d'homographie ou de polysémie effective (pronoms personnels, conjonctions de coordination,...)
- Étiquetage linguistique (TAL)
- Classes à dominantes et annotation semi-automatique (cohérence, propagation), mémoire d'analyse.

Plan

- Pourquoi et comment **automatiser**
- **Affinités** de la textométrie avec l'approche proposée
- Questions de **mise en oeuvre**
- **Fonctionnalités** des logiciels et possibilités d'analyse
- Limites actuelles et **pistes** d'innovation

Fonctionnalités : moteur de recherche

- Requêtes élaborées et phénomènes linguistiques :
  - Graphie : *donc* [modulo majuscule] ; guillemets, tirets cadratins
  - Morphème : préfixe *re-*
  - Locution (et syntagme) : *il y a* ; *ne … pas*, *ne … rien*
  - Catégorie, trait : *<infinitif>*, *<lexique/affectif>*
  - Construction : *dire que* ; *<verbe/parole> + SN/Pro sujet*
- Mais autres exemples tirés du document de travail :
  - (Mise en page et structure : retours à la ligne ?)
  - *<antonyme>* ; reprises syntaxiques et lexicales
  - *[dialogisme]/[Renvoi lexical], [Renvoi structural]*
  - *[rythme]*

Fonctionnalités : une typologie

- Données
- Lecture, retour au texte
- Synthèses
  - Relevés : Vocabulaire, Mesures
  - Positions, répartition
  - Associations : Séquences, Cooccurrences, Analogies
- Conduite de l'analyse : progression des traitements, qualification des résultats

cf. Pincemin, Heiden, Lay, Leblanc, Viprey – soumission à JADT 2010

Fonctionnalités : attestations qualitatives

- Moyen : VOCABULAIRE (dictionnaire, index, liste, t-gen...)
- Usage :
  - formes attestées
    - Lexicalisations canoniques ou non
    - Lexicalisation dispersée/diffuse ou non
  - absences, creux :
    - Un exemple tiré du document de travail : [narration] → pas de <déictiques>, pas de <personnes 1 et 2>

Fonctionnalités : Répartition, disposition

- Au fil du texte (positionnement ou densité)
  - Moyens : TEXTE, (global/macro) DÉROULEMENT, (local/micro) CONCORDANCE
  - Usages : rythmes, constructions en miroir,...
- Par rapport à une structuration en parties
  - Moyens : DISTRIBUTION (graphe, histogramme, spécificités...)
  - Rq. : les parties peuvent être plus ou moins continues : épisode, locuteur ou foyer énonciatif, position dans la structure (ex. pied, rime), type de phrase,... cf. cooccurrence

Fonctionnalités : attestations quantitatives

- Dans l'absolu : fréquences
  - 0 : non attesté
  - 1 (hapax) : non répété
  - seuils ?
- Relativement au corpus :
  - fréquences relatives : intuitives
  - spécificités : modélisation statistique
- Perspectives :
  - base des spécificités (cf. Mayaffre JADT 2006)
    - ex. calculer la spécificité d'un verbe par rapport à tous les verbes et non tous les mots, pour découpler de l'incidence discours nominal vs. verbal
  - Mesures : cf. Cordial, mais interprétation linguistique et cohérence d'ensemble

Fonctionnalités : Mesures – exemple des variables Cordial

- Données (extrait – plusieurs centaines de mesures)
  - % d'articles définis par rapport à l'ensemble des mots
  - % d'article définis par rapport à l'ensemble des déterminants
  - % de déterminants par rapport à l'ensemble des mots
  - % d'articles définis par rapport à l'ensemble des articles
  - % Articles définis masculins singuliers par rapport aux types grammaticaux
  - % Articles définis féminins singuliers par rapport aux types grammaticaux
  - % Articles pluriels invariants en genre par rapport aux types grammaticaux
- Discussion
  - Dépendances et équilibre du jeu
  - Interprétation : paradigme, syntaxe, rythme, trait...

Fonctionnalités : Associations

- Cooccurrences, classification ou cartographies de mots en fonction de leurs contextes
- observer des contextes (EXTRAITS), rapprocher des contextes (ANALOGIES), mesurer et sélectionner des associations fortes (COOCCURRENCE, ANALOGIES)

Fonctionnalités : Données, Analyse

- Comparer des observations
- Textométrie sur corpus étiquetés (Pincemin JADT 2004)
  - Construire les propriétés
  - Étapes : sélection / calcul / visualisation

Plan

- Pourquoi et comment **automatiser**
- **Affinités** de la textométrie avec l'approche proposée
- Questions de **mise en oeuvre**
- **Fonctionnalités** des logiciels et possibilités d'analyse
- Limites actuelles et **pistes** d'innovation

Pistes

- Répétition et effets d'accumulation
- Pluralité et coexistence de marqueurs
  - Plusieurs marqueurs actualisés pour une même unité (ex. les → <actualisation>, <universalisation>)
  - Sèmes : des propriétés ensemblistes ?
  - Un indice a une réalisation ponctuelle ; mais comment définir la réalisation d'un stylème ou d'un marqueur ? (quantité et concentration d'indices, couverture des marqueurs,...)

Pistes

- Visualiser
  - ex. ThemeEditor : une seule couleur par mot (celle qui correspond à l'isotopie la plus développée)
- Modélisation des textes

Plan

- Pourquoi et comment **automatiser**
- **Affinités** de la textométrie avec l'approche proposée
- Questions de **mise en oeuvre**
- **Fonctionnalités** des logiciels et possibilités d'analyse
- Limites actuelles et **pistes** d'innovation