LIDILEM

UNIVERSITÉ Grenoble Alpes

JADT 2018. International Conference on Statistical Analysis of Textual Data (Roma)

Are the expressions relative to space and time relevant to separate novel genres ?

Olivier KRAIF & Julie SORBA, Univ. Grenoble Alpes, LIDILEM

Wednesday, June, 13

Introduction

- Project ANR-DFG PhraseoRom (2016-2020) :

→ Aim : the specific phraseological units of the modern novel (English, French and German)

Hypotheses

• Hypothesis 1: “how the subjective impression of ‘literariness’ arising from fictional works is at least based on the statistically significant use of highly specific words and lexico-grammatical configurations”. (Siepmann, 2015: 362)

• Hypothesis 2: « chaque genre comprend un certain nombre de sous-ensembles, des séries fondées sur la réutilisation de composantes identiques » (Boyer, 1992: 91).

Background

PhraseoRom Corpora

• French corpora: 110 millions of words, novels published since 1950.

Genres		Tokens	Texts
Thriller & detective stories	POL	17 859 351	194
Science Fiction novels	SF	13 173 618	151
Historical novels	HIST	14 868 273	113
General Literature novels	GEN	34 334 554	444
Fantasy novels	FY	13 323 976	105
Sentimental novels	SENT	9 802 410	112

Methods and Tools

• Text mining method: *arbres lexico-syntaxiques récurrents* or *recurring trees* (ALR, Kraif, 2016).

→ A previous study on scientific texts corpora shows first positive results (Tutin & Kraif, 2016).

- SCP-12572 : Sans doute **faut-il y voir** un effet de la dissociation, peut-être plus marquée en économie qu'ailleurs, entre une discipline - instrument de connaissance et une discipline - instrument de pouvoir.
- SOC-12253 : **Il faut** à nouveau **y voir**, mais pas seulement, cette nécessaire adaptation à la complexité d'un système d'enseignement que l'expérimentation contribue à accentuer encore.
- ANT-2214 : La chose est remarquable, car la jeune fille jouit en pays vouté, depuis aussi longtemps qu'on s'en souvienne (**il** n'**y faut voir** aucune trace d'une quelconque « modernité »), d'une grande liberté dans le choix de son époux.

Methods and Tools

• Secondly: comparison of the ALR's results obtained on our corpora with those obtained by n-gram method (Salem, 1987).

→ Do ALR and n-gram method yield different results for sub-genre identification ?

Phraseological Units

• Previous studies on phraseological units strongly related to the themes specific to a genre:

→ e.g. *crime scene* in the detective novels (Kraif, Novakova & Sorba, 2016)

• Present study: phraseological units less related to the salient themes of the novels (like love, science, crime etc.):

→ time and space expressions (setting)

Literature Review

- How the textual genres could be characterised:

→ Lefer, Bestgen & Grabar (2016) examined the 2-4 words n-grams in 3 textual genres (journalistic, scientific and parliamentary debates).

→ Kraif, Novakova & Sorba (2016) examined the ALR's in POL and SF novels (qualitative study, salient themes).

→ Chambre & Kraif (2017) concluded that ALR could properly classify the 98% of the texts of our corpora on the basis of features semantically belonging to settings of the novels (*e.g*. POL *le numéro de portable, à travers le pare-brise, démarrer en trombe, etc.)*

Methodology

Methodology in detail - 1

- Balanced corpora for the present study: 8 millions words samples of the 4 textual sub-genres POL, SF, HIST, GEN:

Genres	Authors	Texts	Tokens
POL	46	69	8 008 395
SF	36	75	8 001 582
HIST	38	70	8 015 933
GEN	46	69	8 008 395

Methodology in detail - 2

- The ALR's method:

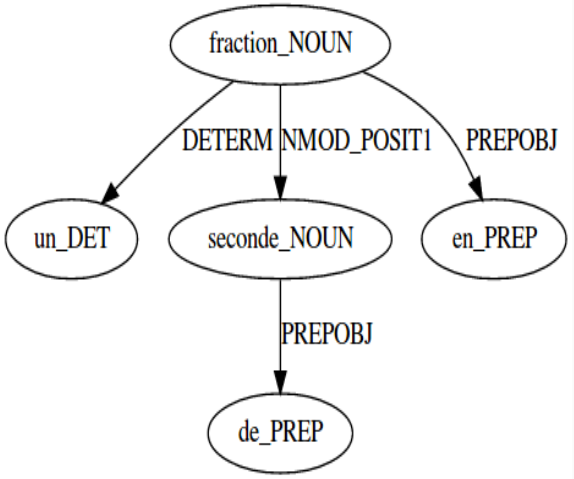
→Dependency relations (XIP, Aït-Mokhtar et al., 2002)

- Criteria for the choice of the ALR:

→Frequency ≥ 10 occurrences

→ Dispersion ≥ 10 authors & 3 sub-genres

→Size ≥ 3 nodes and ≤ 8 nodes



Methodology in detail - 2

• The n-gram's method:
→ Sequences constitued of the lemmata (not of the inflected forms)

• Same criteria used for the choice of the n-grams:
→Dispersion ≥ 10 authors & 3 sub-genres
→Size ≥ 3 nodes and ≤ 8 nodes

• Specific issue : combinatory explosion ! (solution: no new n-gram after the first 80 novels + keyword filtering)

Methodology in detail - 3

- Secondly, the keywords filtering technique based on the observation of the extracted ALR:

→ keywords referring to the space: *cave, salon, hôpital, immeuble, bâtiment, camp, restaurant, village, route, rue, quai, chaussée, terrasse, ministère, parc, bureau, carlingue, maison, toit, chambre, hôtel, palais, rez-de-chaussée, entrée, pont, escalier, chemin, place, salle, jardin, seuil, cour, couloir, colline, sentier, sol, rive, rivage, plage, rivière, mont, montagne, mer, océan, lac, bois, forêt, espace, endroit, coin, pays, continent, frontière, direction, cap, sud, est, nord, ouest, confins, mètre, kilomètre, année-lumière, hectare, acre, loin, proche, près de, au bord de, orée, distance.*

Methodology in detail - 3

- Secondly, the keywords filtering technique based on the observation of the extracted ALR:

→ keywords referring to the time: *matin, soir, soirée, après-midi, nuit, jour, temps, fois, moment, instant, toujours, jamais, parfois, souvent, autrefois, jadis, tôt, tard, longtemps, brièvement, immédiatement, subitement, tout à coup, tout de suite, aujourd'hui, demain, hier, lendemain, maintenant, heure, minute, seconde, journée, semaine, mois, an, année, décennie, siècle, millénaire, printemps, été, automne, hiver.*

Methodology in detail - 4

• Lastly, the selected ALR and n-grams are put into an automatic classification system in order to:

→Test if our *a priori* categories are coherent and if they objectively correlate some criteria.

→ Identify these criteria as a set of semantically relevant features for categorising the textual genres.

Results and Discussion

Results and Discussion

• Global results using the 6000 more frequent ALR (Weka, SVM/SMO):

- P=74 % (10-folds cross validation), Kappa=0,65 (strong agreement)
- Best precision for SF (93,1%) and POL (79,5%). Most of the confusions are with GEN
- GEN got lowest P=59,6 %

Results and Discussion

- Using the 1000 more frequent ALR for TIME :
  - P=48.8 %, Kappa=0.31 → weak agreement
- Using the 1000 more frequent ALR for SPACE :
  - P=59.6 %, Kappa=0.46 → moderate agreement
- For SPACE +TIME (2000 ALR):
  - P=61.4 %, Kappa=0.48 → moderate agreement
  - for POL P=69 %, for GEN P=55,9 %

Results and Discussion

- Selection of the 54 more discriminant features (Weka, BestFirst). Most of them pertain to SPACE:
  - SPACE: 33
  - TIME: 17
  - noise: 4

- Only one feature specific to GEN : *chemin de traverse*

Results and Discussion

- Features specific to HIST:
  - ‘political power’: *la place forte, de son palais, salle du palais, salle du château, pénétrer dans la grande salle*
  - ‘sea’: *sur la mer, de la mer*
  - ‘long time period’: *au bout de quelques mois*, *règne de X années*, *avoir le temps*
  - ‘absolute or relative dates’: *du Ne siècle, venir le lendemain, à trois heures de l’après-midi*

Results and Discussion

- Features specific to POL:
  - ‘rooms’: *de la salle de bain, vers la salle de bain, entrer dans le bureau, vers le bureau, dans le coin*
  - ‘urban places’: *aller à l’hôtel, passer à l’hôpital, à l’hôpital*
  - ‘time’: *8 heures, 21 heures*
  - ‘short time laps’ : *une vingtaine de secondes*

Results and Discussion

- Features specific to SF:

  - ‘very long period’: *milliers d’années, de mille ans*
  - ‘very short time laps’: *une fraction de seconde, un centième de seconde*
  - ‘distances’: *dizaines de mètres, centaine de mètres, plusieurs centaines de mètres*
  - ‘sideral space’: *dans l’espace, à travers l’espace, être dans l’espace, voyager dans l’espace, flotter dans l’espace, espace-temps*
  - ‘ground’: *sur le sol, sous-sol*

Results and Discussion

- Comparison with n-grams → significantly better
  - SPACE: P=66.7% (vs 59.6%)
  - TIME: P=58.3% (vs 48.8%)
  - SPACE+TIME: P=64.1% (vs 61.4%)
    → Difficult to interpret

Results and Discussion

- Best features are very similar :

le chambre de, le cour de, à le cour, dans le espace, le salle de bain, de le espace, dans son bureau, de le immeuble, le maison et, à le hôtel de, centaine de mètre, sur le bureau, sur le place de, le palais de, dans le grand salle, de bureau de, de le salle de bain, sur son bureau, cour de France, en route pour, dans mon bureau, dans tout le direction, un dizaine de mètre, de son pays, à le rue, dans le sous-sol, quitter le salle, dans un restaurant, sur le rivage, mètre plus bas, vers le bureau, route vers le, dizaine de mètre de, un kilomètre de, à ministère de, dans le espace et, de un montagne, le espace et le…

Conclusion and Perspectives

Conclusion

• Phraseology gives good clues for sub-genre identification.

• Even in a restricted semantic field (e.g. SPACE), marked sub-genres (POL, SF, HIST) are well classified.

• n-grams yield better results → need more investigations.

Perspectives

- Deeper comparison of ALR vs n-gram.
- Exploration of other semantic fields.
- Study of other kind of phraseological objects (e.g. motifs, cf. Legallois, Charnois & Poibeau, 2016)

Thank you !

Grazie !

Merci !