JLC 2017

LIDILEM

UNIVERSITÉ Grenoble Alpes

Identification de traits spécifiques du roman policier et de science-fiction

Journées Internationales de la Linguistique de corpus
-
5 juillet 2017 - Grenoble

Judith Chambre & Olivier Kraif

Projet PhraseoRom - ANR 15-FRAL-009

AGENCE NATIONALE DE LA RECHERCHE ANR

Introduction

- Contexte de l'étude
  - Projet PhraséoRom (ANR-DFG) I. Novakova, D. Siepmann
  - Corpus de romans contemporains en trois langues : anglais, allemand, français
  - Comparaison des sous-genres littéraires à travers le prisme de la phraséologie (étude des motifs récurrents)
  - Sous-genres étudiés : policier, sentimental, science-fiction, fantasy, historique… ainsi que la littérature dite générale (environ 12M de mots au moins par sous-genre pour chaque langue)

Introduction

- La notion de genre

« chaque genre comprend un certain nombre de sous-ensembles, des séries fondées sur la réutilisation de composantes identiques : réapparition d'un même protagoniste ou, du moins, recours à un typologie à peu près fixe de personnages ; fonctions déterminées attribuées à chacun d'entre eux ; constantes stylistiques ; utilisation d'un décor propre ; techniques spécifiques d'agencement des épisodes ; etc. » (Boyer, 1992, p.97)

Introduction

- Objectif : identifier des traits spécifiques au roman policier et de science-fiction
- Corpus : 6 sous-corpus de 8M mots
  - POL anglais : 47 auteurs, 69 œuvres
  - SF anglais : 47 auteurs, 67 œuvres
  - GEN anglais : 46 auteurs, 69 oeuvres
  - POL français : 46 auteurs, 69 œuvres
  - SF français : 36 auteurs, 75 œuvres
  - GEN français : 46 auteurs, 69 oeuvres

Introduction

- Hypothèse de recherche :

  Phraséologie (au sens étendu)

  → bon indice pour l'identification des genres

- Exemple

  « se laisser tomber sur le bord du lit »

Introduction

- Hypothèse de recherche :

  Phraséologie (au sens étendu)

  → bon indice pour l'identification des genres

- Exemple

  « se laisser tomber sur le bord du lit »

  → très spécifique du sous corpus Arlequin !

Méthodologie

Méthodologie

- Regroupement des textes en blocs par auteur
- Classification des genres associés aux auteurs *a priori* sur critères éditoriaux
- Approche *corpus-driven*

  → Extraire une série d'indices et identifier *a posteriori* lesquels sont les plus discriminants à l'aide de classifieurs supervisés

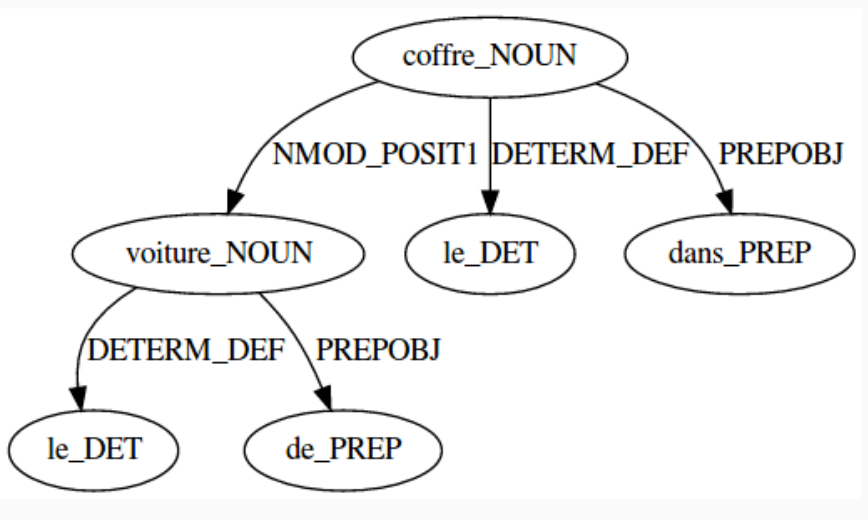
Méthodologie

- Types d'indices
  - Catégories morpho-syntaxiques
    - exemple : fréquences relatives des noms, des adjectifs...
  - Traits morphologiques
    - exemple : mode et temps des verbes, etc.
  - Ponctuation
    - exemple : fréquence relative de la ponctuation, fréquence des points par rapport à tous les signes de ponctuation, etc.

Méthodologie

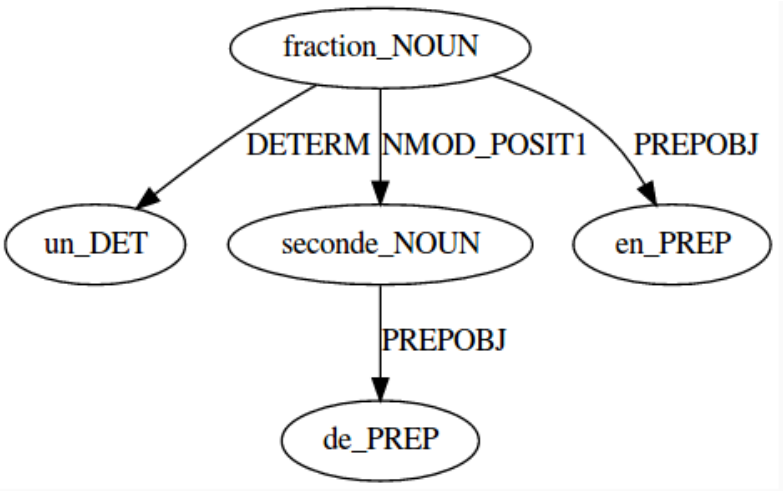
- Lexique
  - Extraction des 2000 noms et des 2000 verbes les plus fréquents dans l'ensemble du corpus
  - Calcul de la fréquence de chacun des lemmes pour chaque auteur
- Expressions polylexicales
  - Extraction automatique d'arbres lexicosyntaxiques récurrents (ou ALR cf., Kraif, Tutin, Diwersy 2014)
  - Dispersion ≥ 10 (nombre de blocs où l'ALR apparaît)

Méthodologie



« dans le coffre de la voiture »

<l=voiture,c=NOUN,#1>&&<l=le,c=DET,#2>&&<l=de,c=PREP,#3>&&<l=coffre,c=NOUN,#4>&&<l=le,c=DET,#5>&&<l=dans,c=PREP,#6>::(NMOD_POSIT1,4,1) (PREPOBJ,1,3) (DETERM_DEF,1,2) (PREPOBJ,4,6) (DETERM_DEF,4,5)



« en une fraction de seconde »

<l=fraction,c=NOUN,#1>&&<l=un,c=DET,#2>&&<l=seconde,c=NOUN,#3>&&<l=de,c=PREP,#4>&&<l=en,c=PREP,#5>::(PREPOBJ,1,5) (NMOD_POSIT1,1,3) (DETERM,1,2) (PREPOBJ,3,4)

Méthodologie

- Algorithmes de classification
  - Classification naïve bayesienne
  - Réseaux bayésiens
  - SVM (Séparateur à vaste marge) / SMO (Sequential minimal optimization)

Résultats

Résultats

Types d'indices	Classifieurs	% d'instances bien classifiées	% d'instances mal classifiées
POS+traits mots outils	LibSVM	67.9688 %	32.0313 %
POS+traits mots pleins	NaiveBayes	65.625 %	34.375 %
Ponctuation	SM0	53.125 %	46.875 %
POS et ponctuation	SMO	69.5313 %	30.4688 %
Lexique (+ Standardize)	LibSVM	82.8125 %	17.1875 %
ALR	SMO	89.0625 %	10.9375 %

Résultats

- Ponctuations

	Fréquence		
	POL	**SF**	**GEN**
Points	451 757	417 985	351 570
Points d'interrogation	60 545	43 743	40 425
Virgule	457 913	472 318	542 559

Résultats

– Répartition des mots outils

	Fréquence		
	POL	**SF**	**GEN**
Adjectifs possessifs	149 422	152 941	156 884
Quantifieurs	31 577	36 275	30 526
Pronom sujet 3e pers. Sg	184 761	145 334	164 860
Pronom sujet 2e pers. Pl	33 015	25 912	22 643
Pronom sujet 3e pers. Pl	21 858	28 049	23 575
Pronom On	25 995	17 921	27 557
Pronoms le, la, lui, l'	78 763	63 535	70 413
Pronom réfléchi se	97 203	102 114	88 747
Conjonctions de coordination	172 422	190 730	212 748

Résultats

- Répartition des parties du discours
  - mots pleins

	Fréquence		
	POL	SF	GEN
Noms propres	128 447	65 843	100 591
Verbes	1 322 885	1 238 589	1 219 695
Verbes (indicatif)	715 321	670 348	648 384
Verbes (futur)	17 247	19 841	15 492
Verbes (subjonctif imparfait)	2 087	4 100	8 241
Adjectifs	312 665	373 612	352 564

Résultats

- Lexique : des résultats prévisibles...

	Fréquence		
	POL	SF	GEN
97 arme_n	1734	**2205**	852
147 espace_n	581	**2309**	603
152 police_n	**2450**	538	470
234 planète_n	145	**2201**	146
247 flic_n	**1927**	214	273
301 univers_n	214	**1480**	412
368 énergie_n	299	**1188**	311
578 vaisseau_n	80	**1130**	53
840 combinaison_n	198	**611**	87
935 robot_n	68	**728**	15
958 engin_n	167	**553**	70
1011 onde_n	126	**534**	85
1298 galaxie_n	8	**550**	9
1346 sphère_n	31	**461**	46
1473 sas_n	25	**439**	19
1810 coque_n	30	**275**	72
1993 circuit_n	78	**228**	34

Résultats

- Lexique : des résultats prévisibles ... ou pas !

	Fréquence		
	POL	SF	GEN
60 voiture_n	**4006**	835	1632
176 café_n	**1790**	271	1097
195 téléphone_n	**1943**	236	592
268 dossier_n	**1642**	375	300
290 surface_n	474	**1322**	389
319 système_n	373	**1 454**	211
331 patron_n	**1175**	187	602
483 métal_n	312	**956**	204
620 cellule_n	**444**	**589**	165
874 communication_n	267	**496**	106
1199 structure_n	85	**482**	55
1910 suspect_n	**278**	25	52
1979 complexe_n	42	**260**	41
2579 garer_v	**612**	103	168
3050 enquêter_v	**258**	43	55
3212 injecter_v	**124**	**135**	21
3536 sélectionner_v	58	**109**	14

Résultats

- ALR : 84 expressions discriminantes sélectionnées (impact de 58 %)

  - 39 ALR spécifiques de POL
  - 26 ALR spécifiques de SF
  - 7 ALR spécifiques de POL et SF (vs GEN)
  - 6 ALR spécifiques de GEN

Résultats

- ALR spécifiques à POL : quelques classes très productives
  - voiture

    de la voiture, dans la voiture, à travers le pare-brise, la voiture de police, sortir de la voiture, la boîte à gant, démarrer en trombe
  - téléphone, communication

    le numéro de portable, laisser un message, noter le numéro, passer un coup de fil, sur son portable, appeler la police, si on a du nouveau, tenir au courant
  - expressions temporelles

    à huit heure, vingt et une heure, au bout de X minutes
  - police

    le policier qui, la gendarmerie de, dans la police, la brigade criminelle, à l'inspecteur

Résultats

- ALR spécifiques à POL : restent des expressions liées aux déroulement d'une enquête

  faire le rapprochement, dans ses pensées, dans le dossier, un léger accent, être en danger, le meurtre a, disparaître de la circulation, il est le témoin, se prendre la main entre, entrer dans le bureau, d'un signe de tête, sortir de prison, jeter un coup, pour un con, il ouvre le tiroir, dans la poche

Résultats

- ALR spécifiques à SF : également des classes homogènes
  - expressions spatiales très variées

    par la voie, dans le territoire, dans la sphère, dans l'espace, dans toute les directions, la zone de, être dans l'espace, de la galaxie, de la planète, planète du système, plusieurs centaines de mètres
  - artefacts technologiques

    défiler sur l'écran, de la machine, dans le circuit, plaque de métal, un vaisseau de
  - rapports de pouvoir

    donner l'ordre, prendre le pouvoir, le plus puissant
  - rapport à la connaissance

    la solution du problème, commettre une erreur, être le cas, de la science
  - autre : *d'être vivant, de la drogue, de la science fiction*

Résultats

- Quelques expressions spécifiques communes

  prendre le risque
  attendre le résultat
  appuyer sur la touche
  commettre une erreur
  être le cas
  presser la détente
  d'aucune utilité

Résultats

- Très peu d'expressions spécifiques à GEN (et de faibles fréquences)

  trois heure de l'après-midi

  venir le lendemain

  billet de mille

  le soleil fait

  l'argent que

  la mer est

Conclusion et perspectives

Conclusion

- Des indices textométriques simples purement surfaciques peuvent apporter une information intéressante (à 10 point des indices lexicaux non standardisés) :
  - ponctuation
  - parties du discours et traits morphosyntaxiques

- Les ALR sont plus discriminants que les mots isolés
  - Beaucoup de collocations
  - Moins d'ambiguïtés
  - Certains agissent comme de vrais marqueurs génériques

Perspectives

- Analyser les résultats pour l'anglais
- Interpréter finement les indices surfaciques (ponctuations, etc.) par un retour au texte
- Classer plus finement les expressions spécifiques par genre
- Elargir l'étude à l'ensemble des sous-genres du corpus

Merci pour votre attention !