Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Straight Talk

Automatic Recognition of Direct Speech in Nineteenth-century French Novels

CLiGS

Daniel Schlör, Stefanie Popp, Christof Schöch, Ulrike Henny,
Annelen Brunner, José Calvo Tello
(CLiGS Group, University of Würzburg, Germany)

DH2016, Kraków, July 13, 2016

Straight Talk

Automatic Recognition of Direct Speech in Nineteenth-century French Novels

1. Introduction
2. Data
3. Method
4. Results
5. Conclusion

1. Introduction

The CLiGS group

- CLiGS = Computational Literary Genre Stylistics
- Junior Research Group, Department of Literary Computing, University of Würzburg, Germany
- French and Spanish Studies and Computer Science / Text Mining

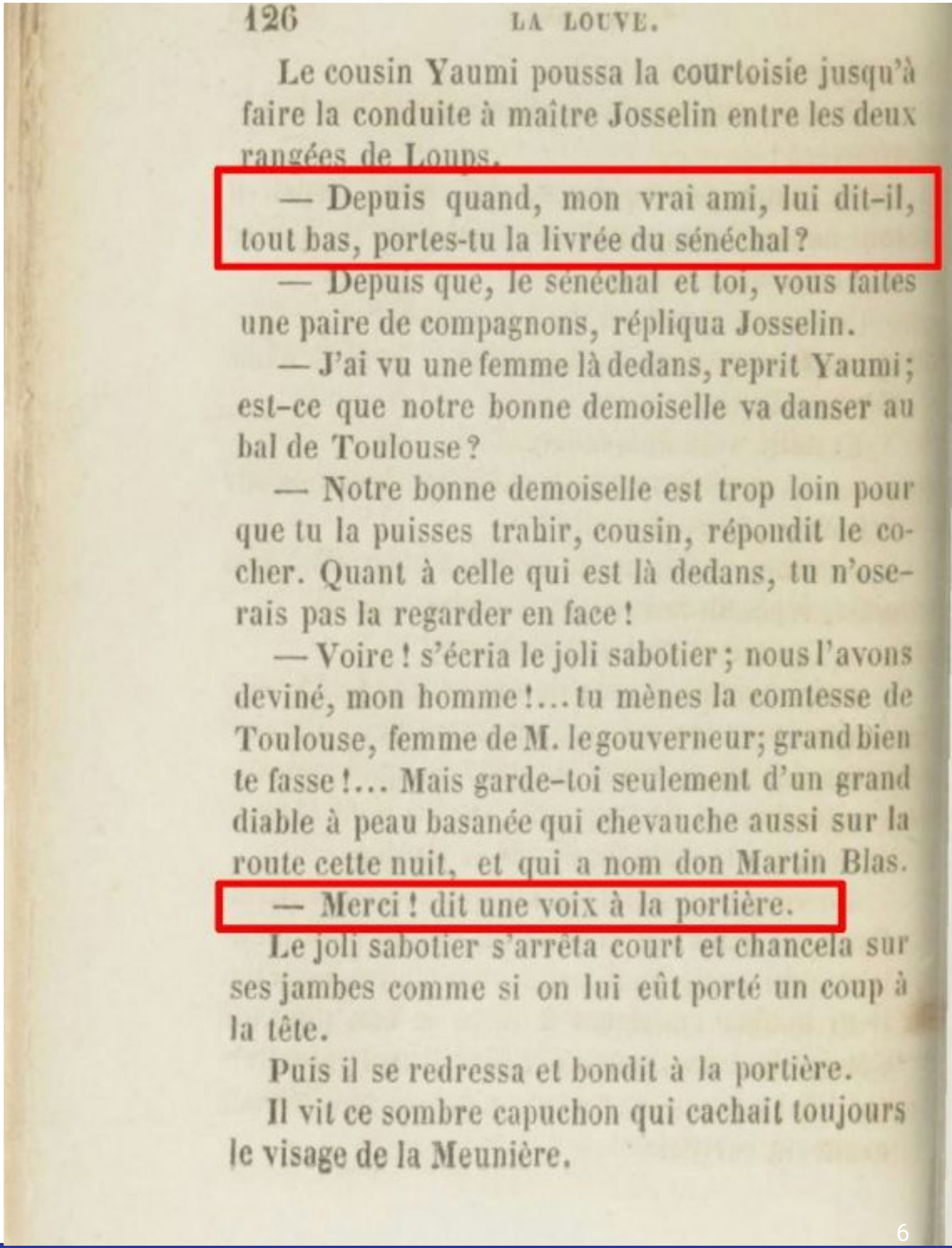
- http://cligs.hypotheses.org
- http://github.com/cligs/

Aims of this study

- Automatically identify direct speech in novels (using linguistic markers vs. using typography)
- Assess proportion of direct speech for subgenres and decades
- Enhance subtlety of analyses in stylometry
- Prepare for subdivision of narrator speech by text type

Starting point

- direct speech and thought presentation
- lack of systematic typographical boundaries (in French)



126 LA LOUVE.

Le cousin Yaumi poussa la courtoisie jusqu'à faire la conduite à maître Josselin entre les deux rangées de Loups.

— Depuis quand, mon vrai ami, lui dit-il, tout bas, portes-tu la livrée du sénéchal ?

— Depuis que, le sénéchal et toi, vous faites une paire de compagnons, répliqua Josselin.

— J'ai vu une femme là dedans, reprit Yaumi; est-ce que notre bonne demoiselle va danser au bal de Toulouse ?

— Notre bonne demoiselle est trop loin pour que tu la puisses trahir, cousin, répondit le cocher. Quant à celle qui est là dedans, tu n'oserais pas la regarder en face !

— Voire ! s'écria le joli sabotier ; nous l'avons deviné, mon homme !... tu mènes la comtesse de Toulouse, femme de M. le gouverneur; grand bien te fasse !... Mais garde-toi seulement d'un grand diable à peau basanée qui chevauche aussi sur la route cette nuit, et qui a nom don Martin Blas.

— Merci ! dit une voix à la portière.

Le joli sabotier s'arrêta court et chancela sur ses jambes comme si on lui eût porté un coup à la tête.

Puis il se redressa et bondit à la portière.

Il vit ce sombre capuchon qui cachait toujours le visage de la Meunière.

2. Data

Corpus 1

- 127 novels
- 1840-1889
- 40 random chapters annotated manually
- sentence contains direct speech: yes/no

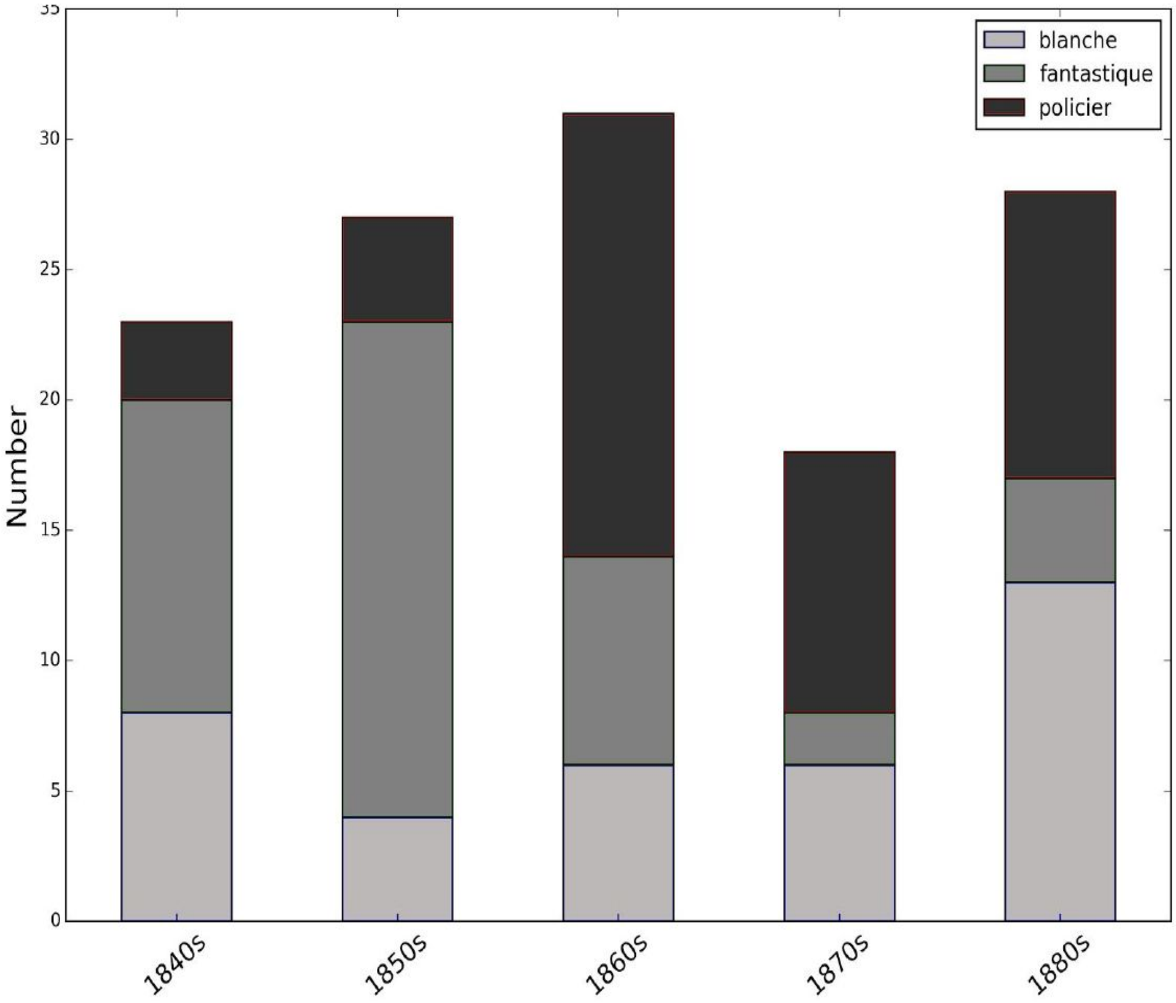
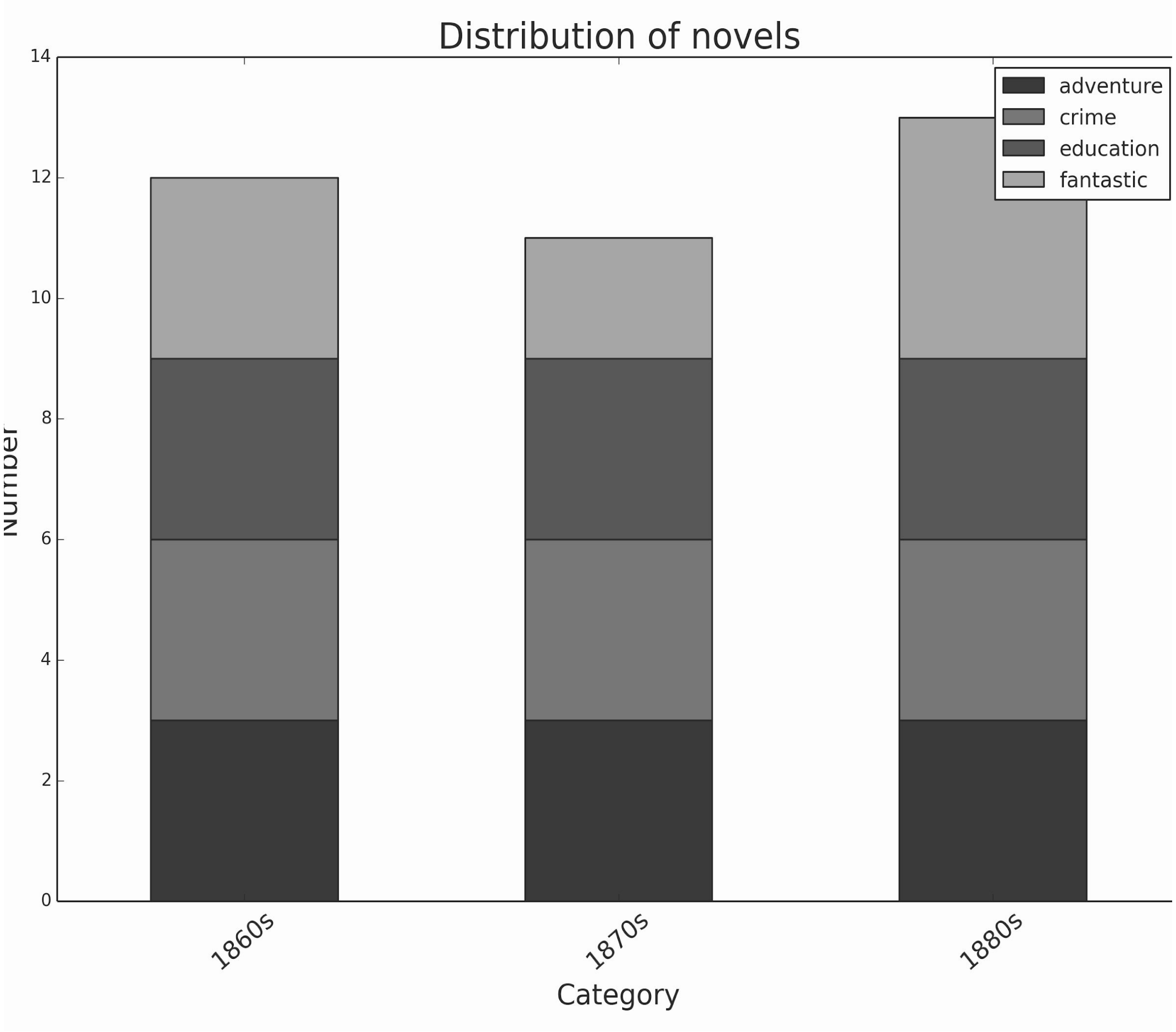


Figure 2: Distribution of novels per subgenre and decade.

Corpus 2

- 36 novels
- 1861-1889
- 4 subgenres
- balanced for subgenre



3. Method

Overview

Approach: Machine-learning

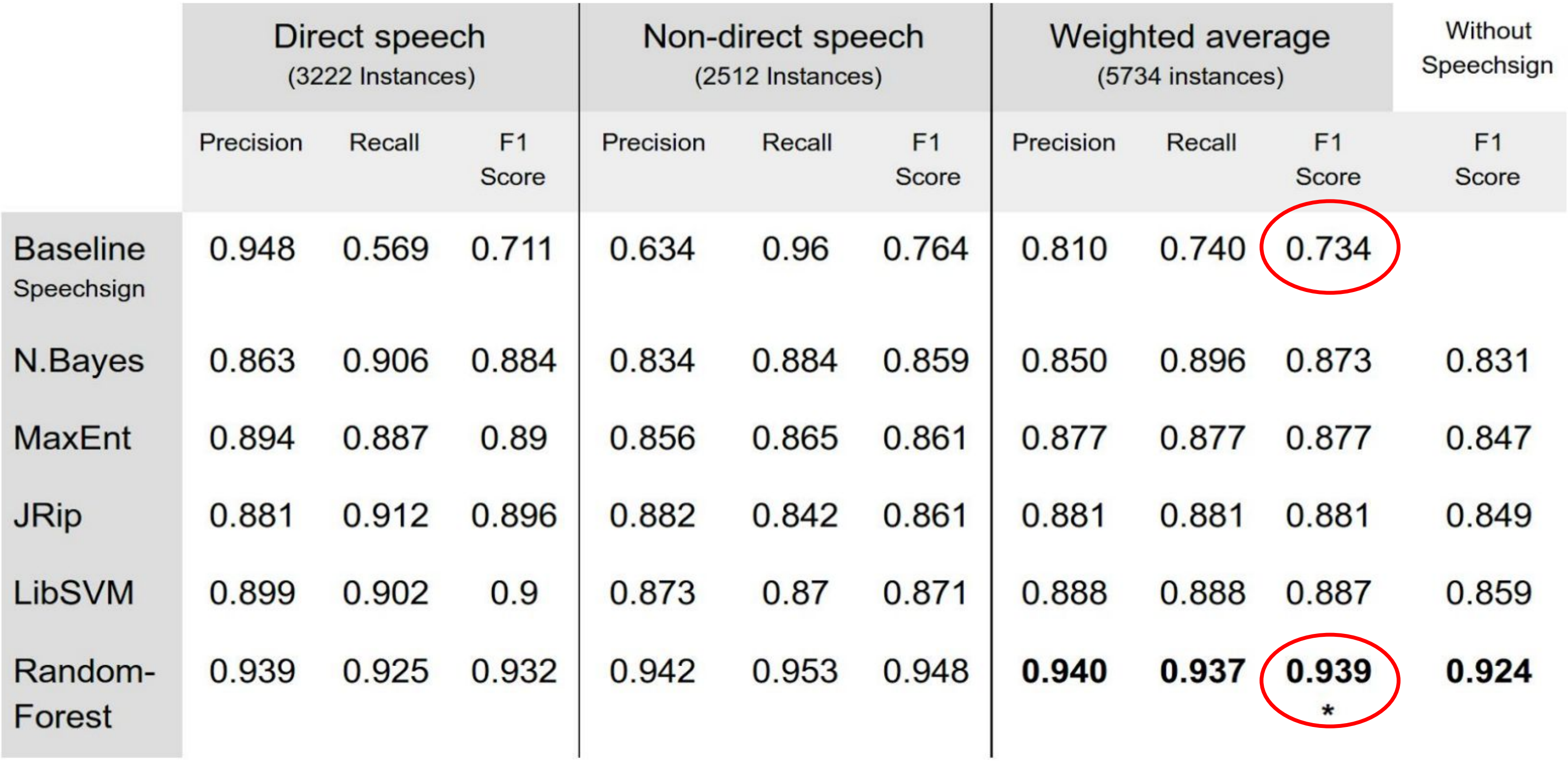
Steps:

- Manual label-annotation of a partial corpus:
  Sentence contains direct speech
- Feature generation
- Learning relation between labels and features
- Evaluation
- Application (automatic classification) on complete corpus
- Analysis of distribution of direct speech:
  (decade and subgenre)

Features Types

- 81 Features
- Different feature-categories:
  - Char-based: Speech-sign, exclamation-mark, ...
  - Lexical: deictic expressions, interjections, ...
  - Semantic: Lexical verb-category (WordNet)
  - Morphological: Part-Of-Speech, tense, lemmas, ...
  - Syntactical: Numbers of commas, sentence-length, ...

Performance on annotated partial-corpus



	Direct speech (3222 Instances)			Non-direct speech (2512 Instances)			Weighted average (5734 instances)			Without Speechsign
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	F1 Score
Baseline Speechsign	0.948	0.569	0.711	0.634	0.96	0.764	0.810	0.740	0.734	
N.Bayes	0.863	0.906	0.884	0.834	0.884	0.859	0.850	0.896	0.873	0.831
MaxEnt	0.894	0.887	0.89	0.856	0.865	0.861	0.877	0.877	0.877	0.847
JRip	0.881	0.912	0.896	0.882	0.842	0.861	0.881	0.881	0.881	0.849
LibSVM	0.899	0.902	0.9	0.873	0.87	0.871	0.888	0.888	0.887	0.859
Random-Forest	0.939	0.925	0.932	0.942	0.953	0.948	**0.940**	**0.937**	**0.939** *	**0.924**

Table 1: Performance (10-fold cross-validation on the gold standard)

Applying the trained model to unannotated corpus

- Complete corpus automatically annotated with trained model
- 100 sentences per document randomly sampled and manually revised
- 15.1% false positives for direct speech
- 16.1% false positives for non direct speech
- F1 Score: 0.84
- Problems identified

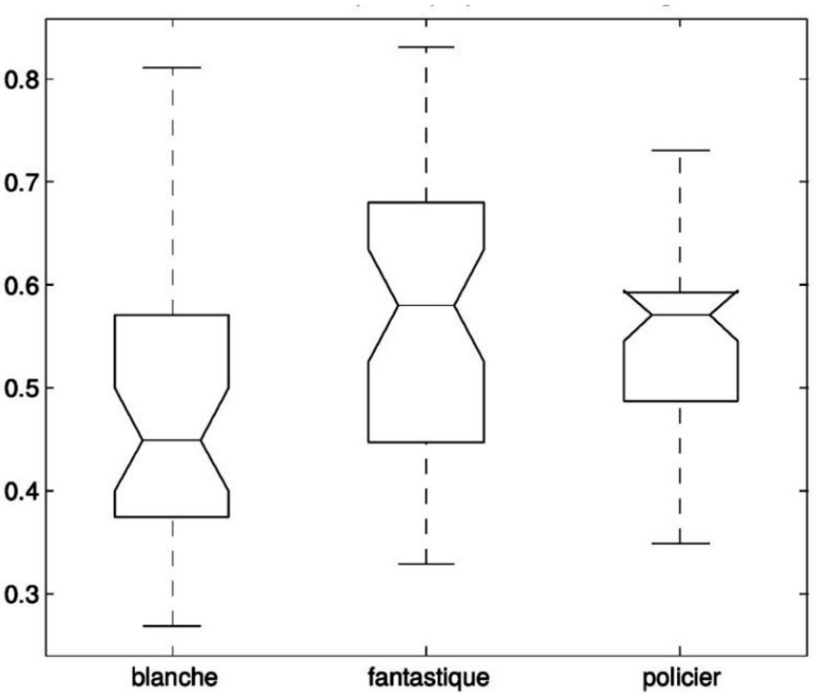
  E.g. Sentence splitting policy (colon)

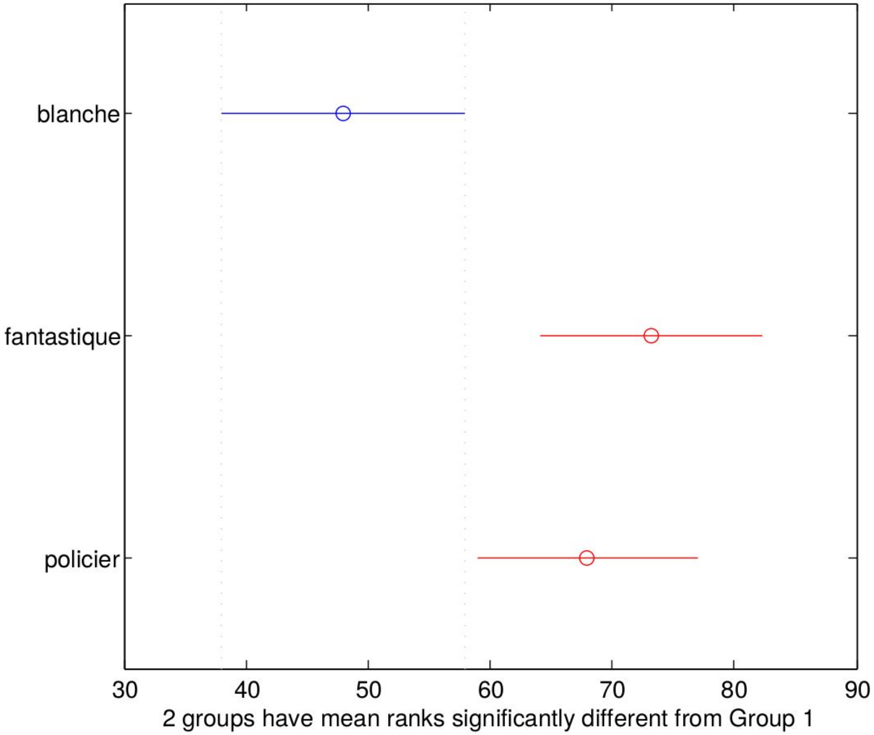
Impact of sentence-splitting strategy

- Impact due to problem definition:
  - More instances
  - More instances classified as non-direct speech (40.2% splitting by , vs. 61% sentence based)
- Worse performance
- Same performance ranking of algorithms
- Benefit of sequential methods (CRFs) (each Macro-F1)
  - MaxEnt: 0.779
  - SVM: 0.782
  - CRF (w=5): 0.823
  - CRF (w=Sen): 0.834

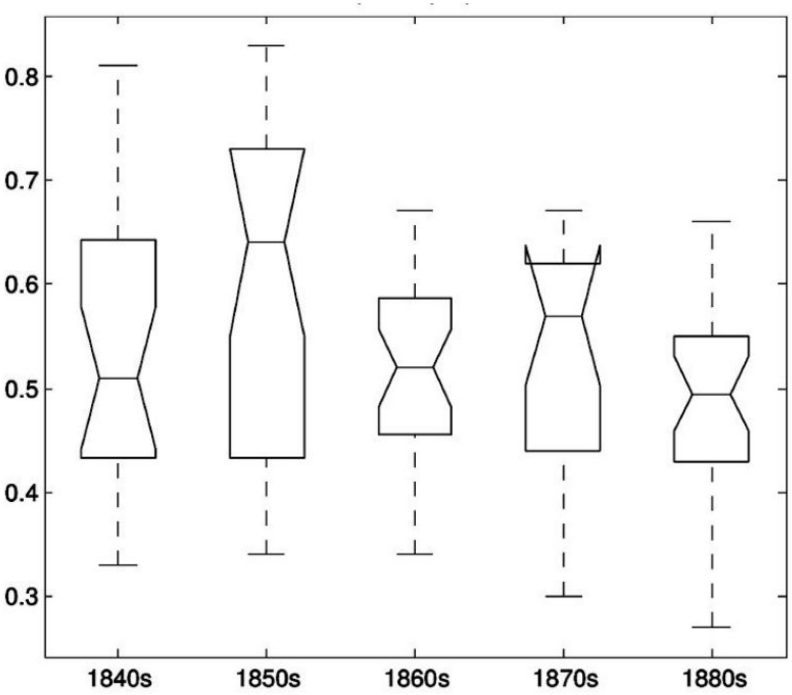
4. Results

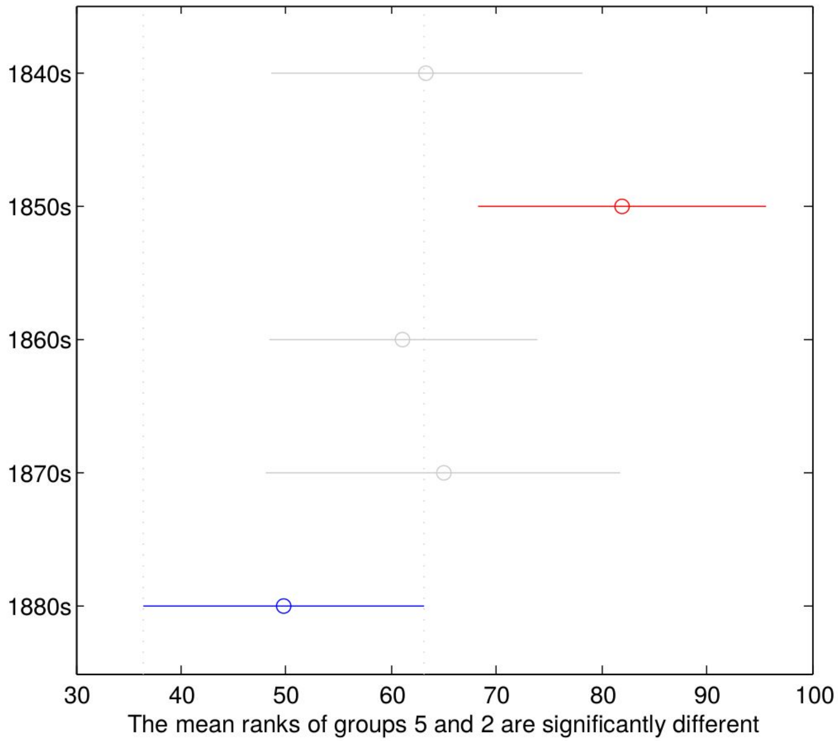
Corpus 1: Proportion of direct speech (by genre)





Corpus 1: Proportion of direct speech (by decade)





Significance and corpus

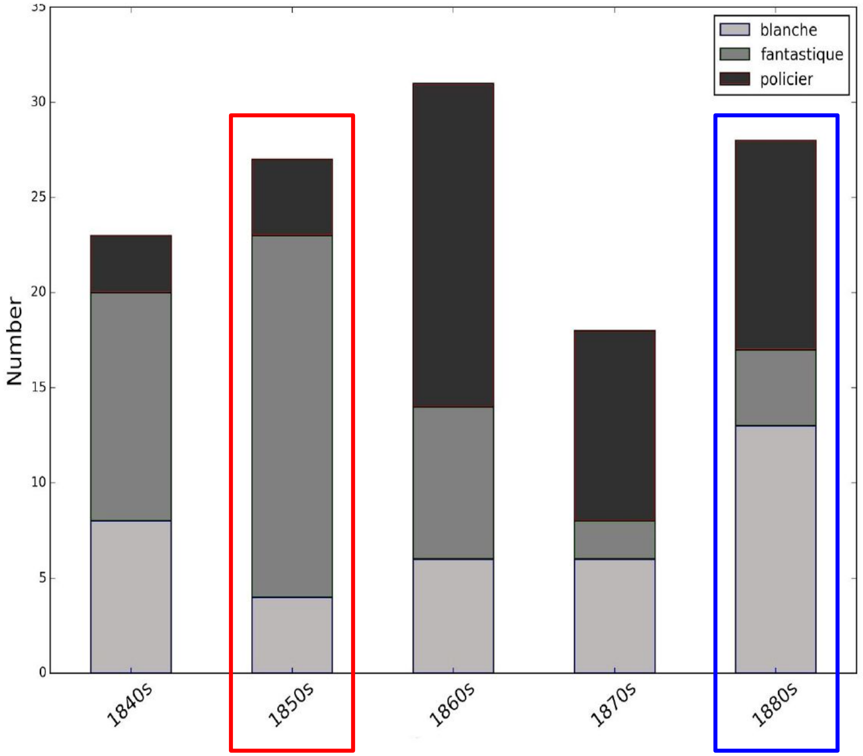
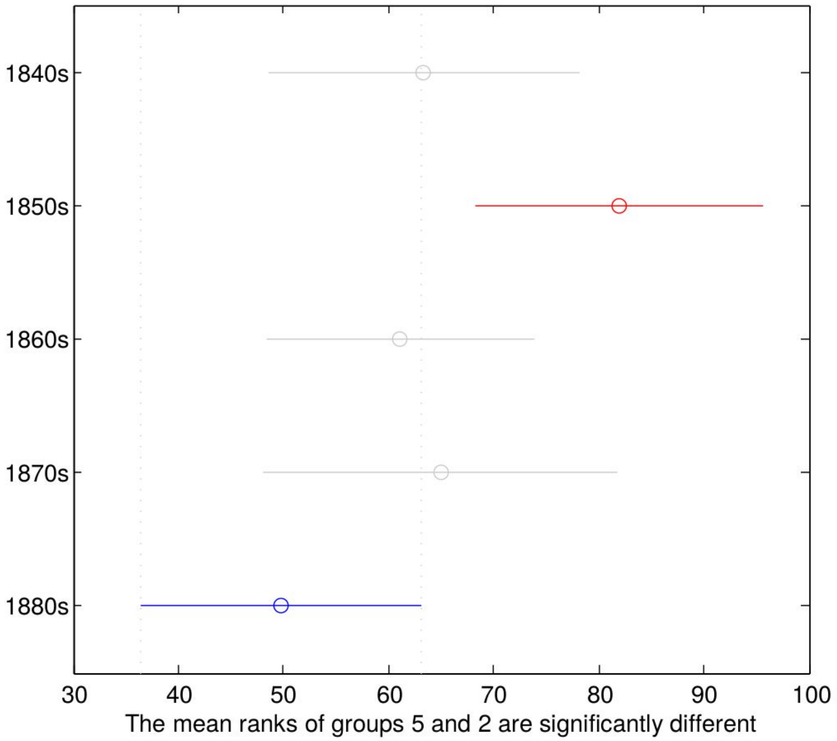
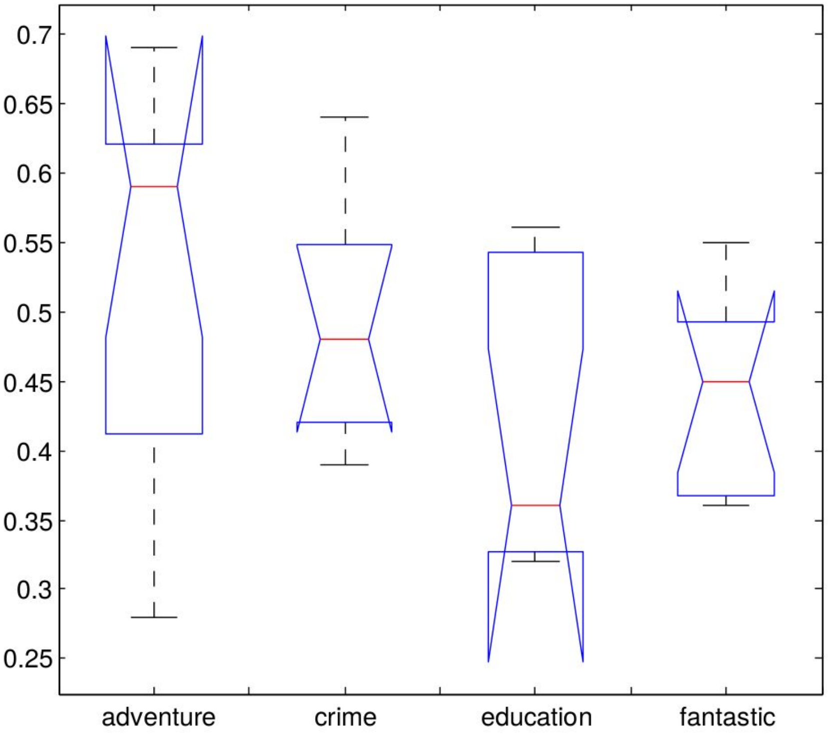
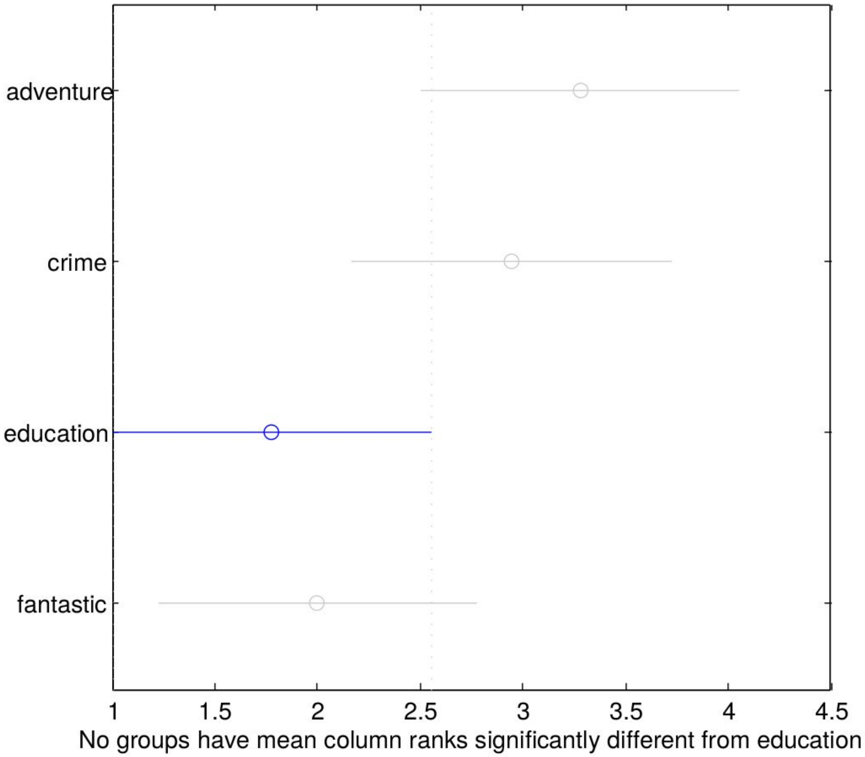


Figure 2: Distribution of novels per subgenre and decade.

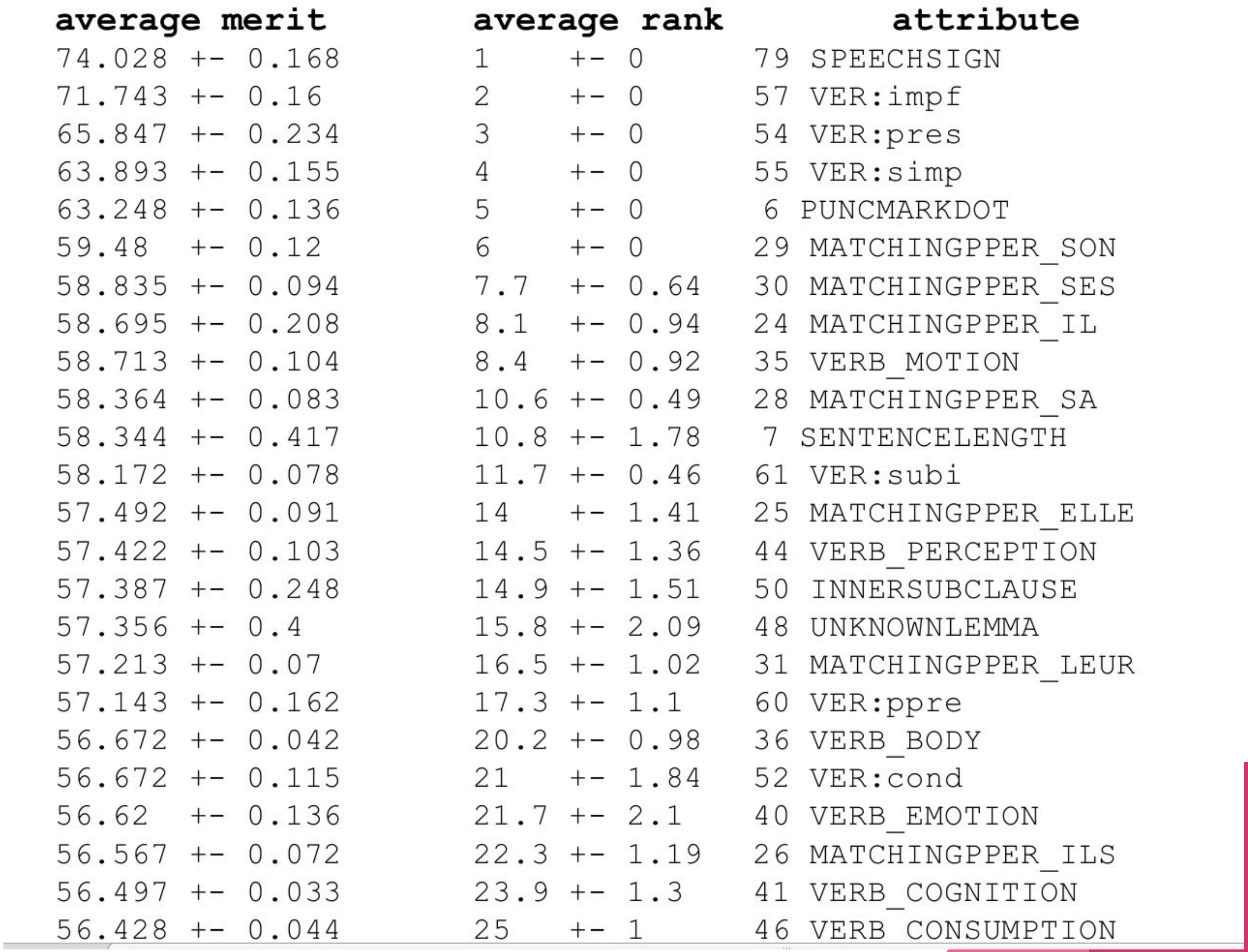


Corpus 2: proportions of direct speech (by genre)



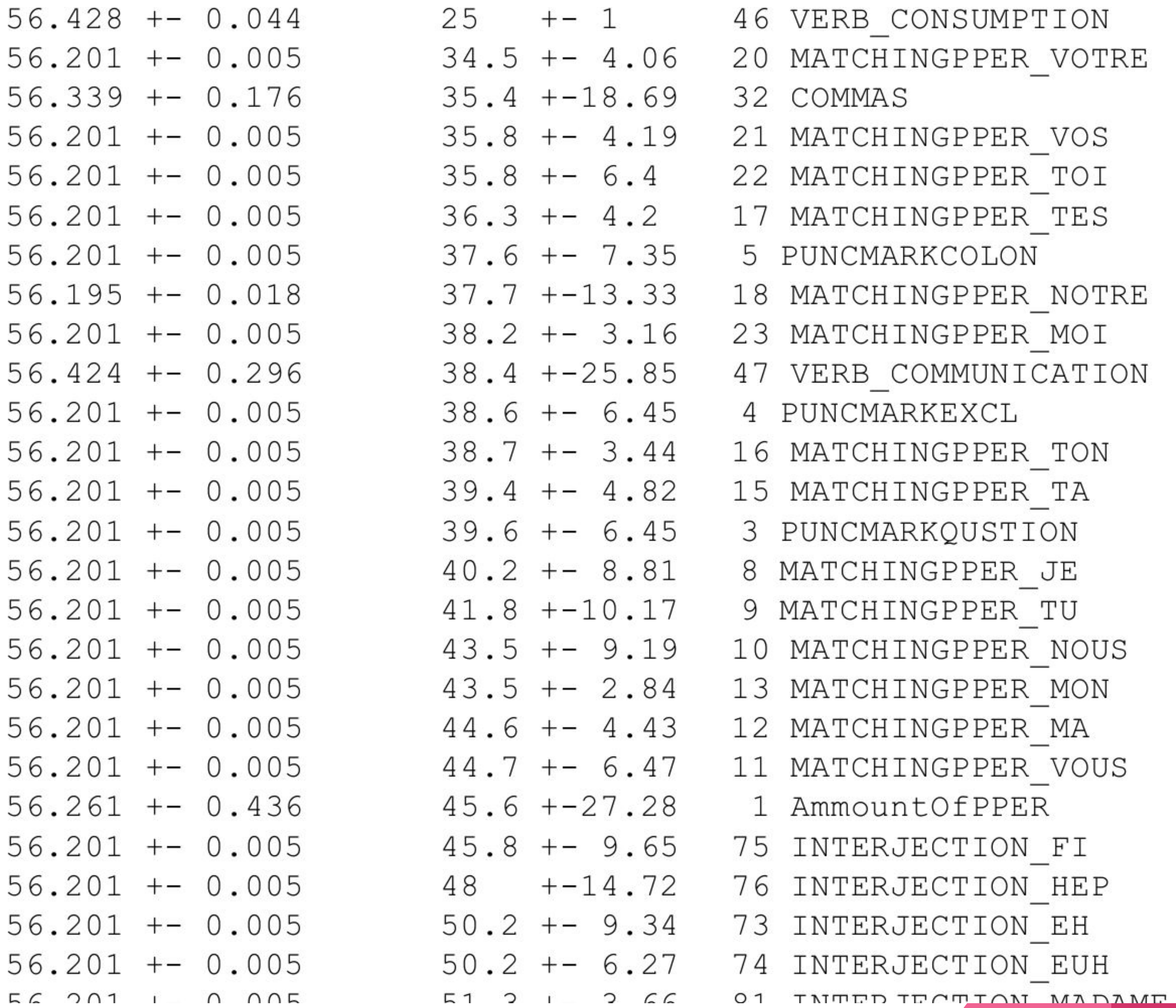


Features used (sorted by rank; part 1)



```
average merit          average rank          attribute
74.028 +- 0.168        1    +- 0       79 SPEECHSIGN
71.743 +- 0.16         2    +- 0       57 VER:impf
65.847 +- 0.234        3    +- 0       54 VER:pres
63.893 +- 0.155        4    +- 0       55 VER:simp
63.248 +- 0.136        5    +- 0        6 PUNCMARKDOT
59.48  +- 0.12         6    +- 0       29 MATCHINGPPER_SON
58.835 +- 0.094        7.7  +- 0.64    30 MATCHINGPPER_SES
58.695 +- 0.208        8.1  +- 0.94    24 MATCHINGPPER_IL
58.713 +- 0.104        8.4  +- 0.92    35 VERB_MOTION
58.364 +- 0.083        10.6 +- 0.49    28 MATCHINGPPER_SA
58.344 +- 0.417        10.8 +- 1.78     7 SENTENCELENGTH
58.172 +- 0.078        11.7 +- 0.46    61 VER:subi
57.492 +- 0.091        14   +- 1.41    25 MATCHINGPPER_ELLE
57.422 +- 0.103        14.5 +- 1.36    44 VERB_PERCEPTION
57.387 +- 0.248        14.9 +- 1.51    50 INNERSUBCLAUSE
57.356 +- 0.4          15.8 +- 2.09    48 UNKNOWNLEMMA
57.213 +- 0.07         16.5 +- 1.02    31 MATCHINGPPER_LEUR
57.143 +- 0.162        17.3 +- 1.1     60 VER:ppre
56.672 +- 0.042        20.2 +- 0.98    36 VERB_BODY
56.672 +- 0.115        21   +- 1.84    52 VER:cond
56.62  +- 0.136        21.7 +- 2.1     40 VERB_EMOTION
56.567 +- 0.072        22.3 +- 1.19    26 MATCHINGPPER_ILS
56.497 +- 0.033        23.9 +- 1.3     41 VERB_COGNITION
56.428 +- 0.044        25   +- 1       46 VERB_CONSUMPTION
```

Features used (sorted by rank; part 2)



```
56.428 +- 0.044       25   +- 1      46 VERB_CONSUMPTION
56.201 +- 0.005       34.5 +- 4.06   20 MATCHINGPPER_VOTRE
56.339 +- 0.176       35.4 +-18.69   32 COMMAS
56.201 +- 0.005       35.8 +- 4.19   21 MATCHINGPPER_VOS
56.201 +- 0.005       35.8 +- 6.4    22 MATCHINGPPER_TOI
56.201 +- 0.005       36.3 +- 4.2    17 MATCHINGPPER_TES
56.201 +- 0.005       37.6 +- 7.35    5 PUNCMARKCOLON
56.195 +- 0.018       37.7 +-13.33   18 MATCHINGPPER_NOTRE
56.201 +- 0.005       38.2 +- 3.16   23 MATCHINGPPER_MOI
56.424 +- 0.296       38.4 +-25.85   47 VERB_COMMUNICATION
56.201 +- 0.005       38.6 +- 6.45    4 PUNCMARKEXCL
56.201 +- 0.005       38.7 +- 3.44   16 MATCHINGPPER_TON
56.201 +- 0.005       39.4 +- 4.82   15 MATCHINGPPER_TA
56.201 +- 0.005       39.6 +- 6.45    3 PUNCMARKQUSTION
56.201 +- 0.005       40.2 +- 8.81    8 MATCHINGPPER_JE
56.201 +- 0.005       41.8 +-10.17    9 MATCHINGPPER_TU
56.201 +- 0.005       43.5 +- 9.19   10 MATCHINGPPER_NOUS
56.201 +- 0.005       43.5 +- 2.84   13 MATCHINGPPER_MON
56.201 +- 0.005       44.6 +- 4.43   12 MATCHINGPPER_MA
56.201 +- 0.005       44.7 +- 6.47   11 MATCHINGPPER_VOUS
56.261 +- 0.436       45.6 +-27.28    1 AmmountOfPPER
56.201 +- 0.005       45.8 +- 9.65   75 INTERJECTION_FI
56.201 +- 0.005       48   +-14.72   76 INTERJECTION_HEP
56.201 +- 0.005       50.2 +- 9.34   73 INTERJECTION_EH
56.201 +- 0.005       50.2 +- 6.27   74 INTERJECTION_EUH
56.201 +- 0.005       51.3 +- 3.66   81 INTERJECTION_MADAME
```

5. Conclusion

Results

- Good classification result (F1-Score 0.94 resp. 0.84)
- Quite large proportion of direct speech (61%) on average
- Proportion over decade: no significant variations (α=0.01)
- Genre: blanche vs. policier and fantastic

Challenges

- Sentence segmentation: precision and granularity
- Take insertions (“dit-elle”) into consideration
- Features related to the position in the sentence / paragraph
- Corpus structure: larger and more balanced

“Thank you!”, they said at the end of their presentation.

—Merci!, ils ont dit à la fin de leur présentation.

http://cligs.hypotheses.org

http://github.com/cligs