Direkte Rede im französischen Roman

Automatische Erkennung und gattungsabhängige Verteilungen

<philtag n="13"/>, Universität Würzburg, 25.-26. Feb. 2016

Christof Schöch, Stefanie Popp, Daniel Schlör, José Calvo Tello, Ulrike Henny

Direkte Rede im französischen Roman

Automatische Erkennung und gattungsabhängige Verteilungen

Gliederung

1. Kontext (CLiGS)
2. Ausgangslage
3. Methode
4. Ergebnisse
5. Fazit

1. Kontext: CLiGS-Nachwuchsgruppe

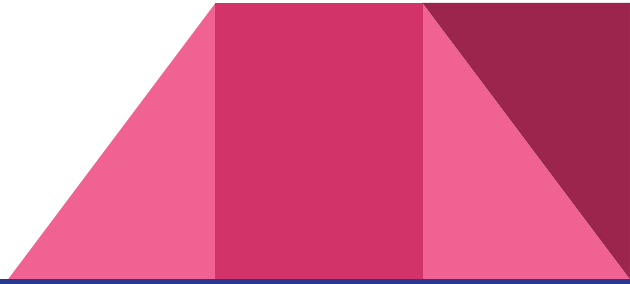
CLiGS

- CLiGS = Computergestützte literarische Gattungsstilistik
- Nachwuchsgruppe am Lehrstuhl für Computerphilologie
- Romanistische Literaturwissenschaft und Informatik/Text Mining
- Mentoren: Fotis Jannidis, Andreas Hotho, Brigitte Burrichter

textbox, toolbox, tmw

- textbox: vier projektbezogene Textsammlungen (Teilmenge)
- toolbox: Python-Modul mit diversen kleineren Funktionen
- tmw: Python-basierter Workflow für topic Modeling

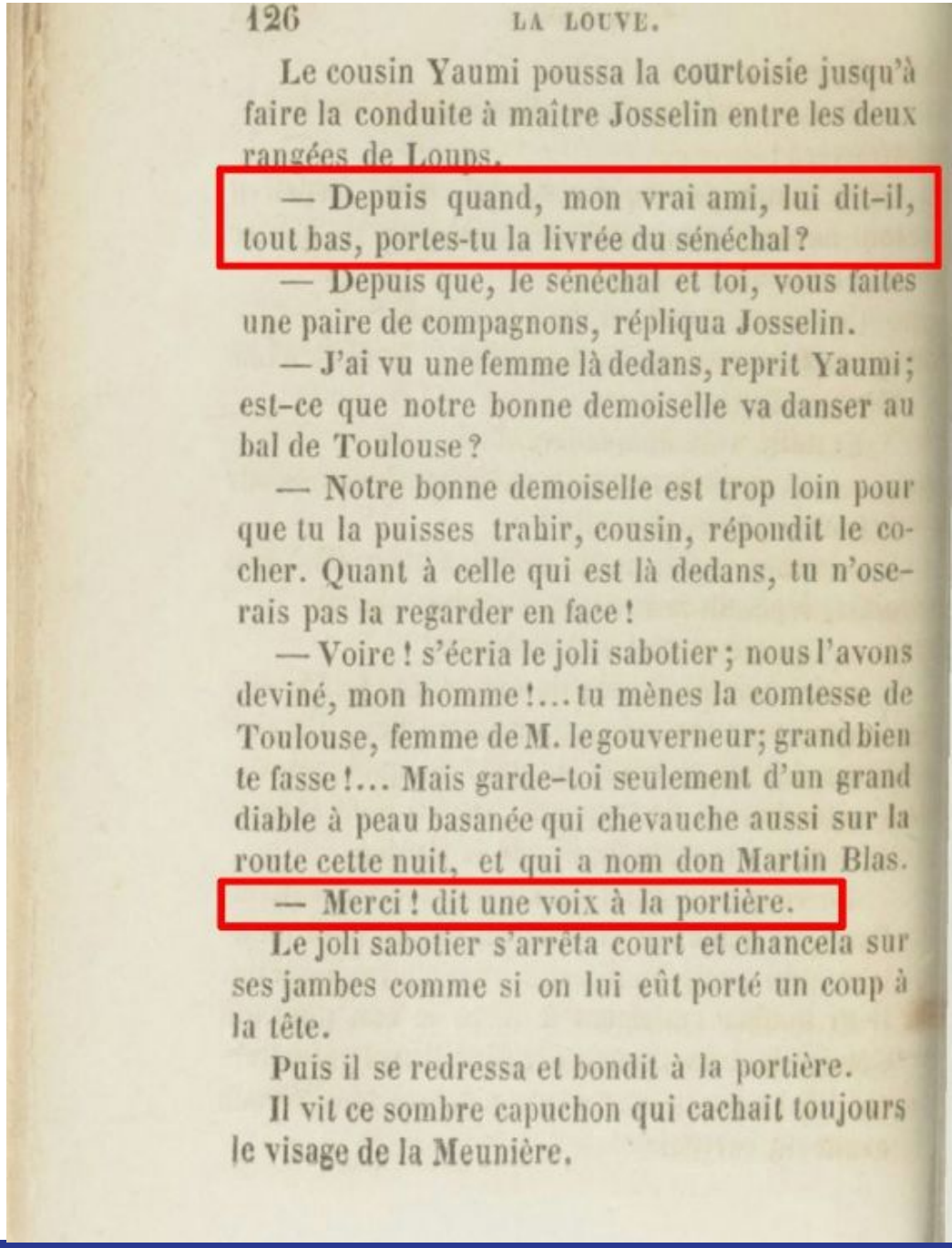
github.com/cligs



2. Fragestellung und Korpora

2.1 Fragestellung

- verschiedene Repräsentationen für gesprochene Sprache und Gedanken
- französische Literatur: keine einheitliche Typographie



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.

Das große Korpus

- 127 Romane aus den Jahren 1840-1889
- 40 Kapitel manuell annotiert
- Satz enthält direkte Rede: ja/nein

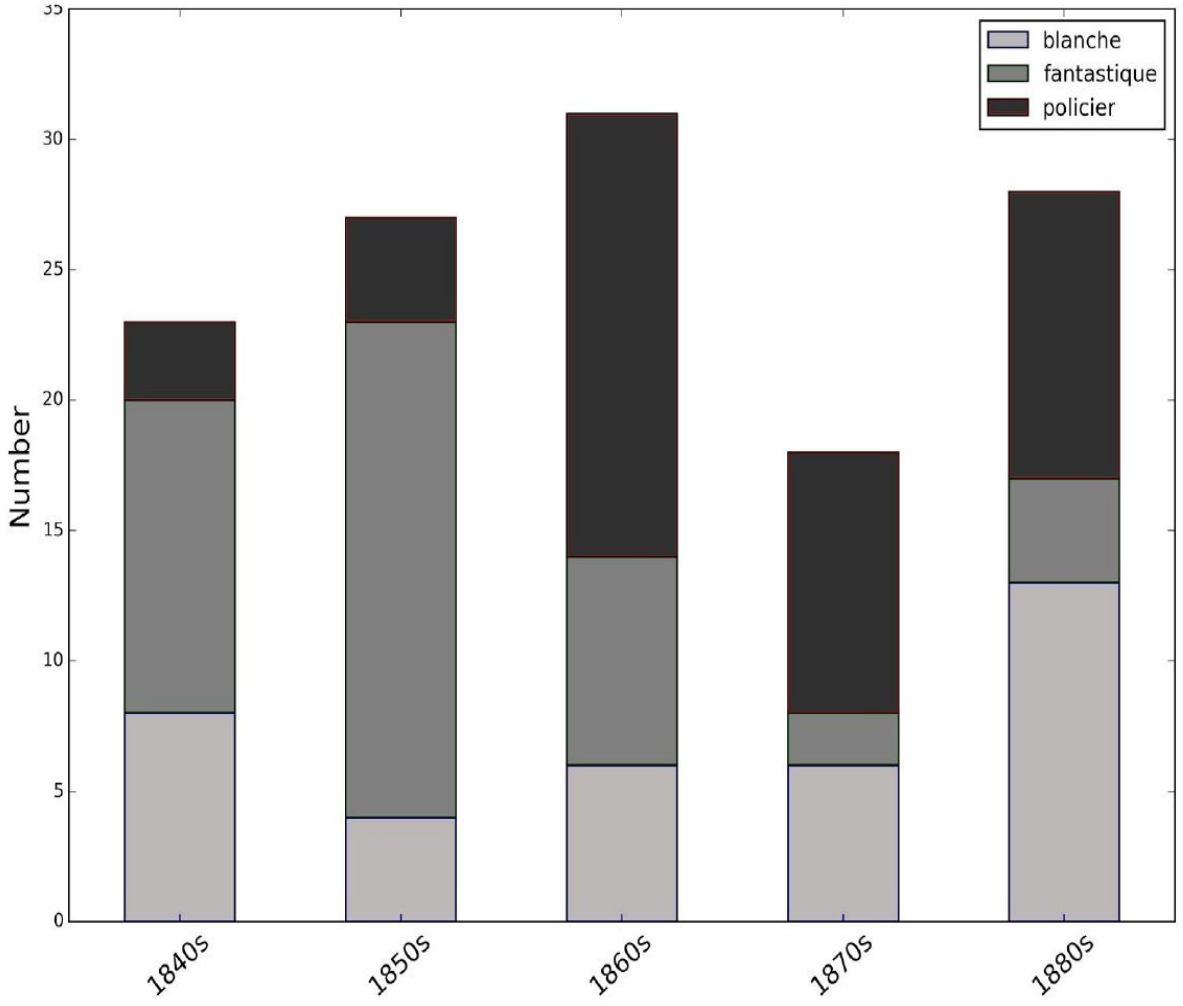
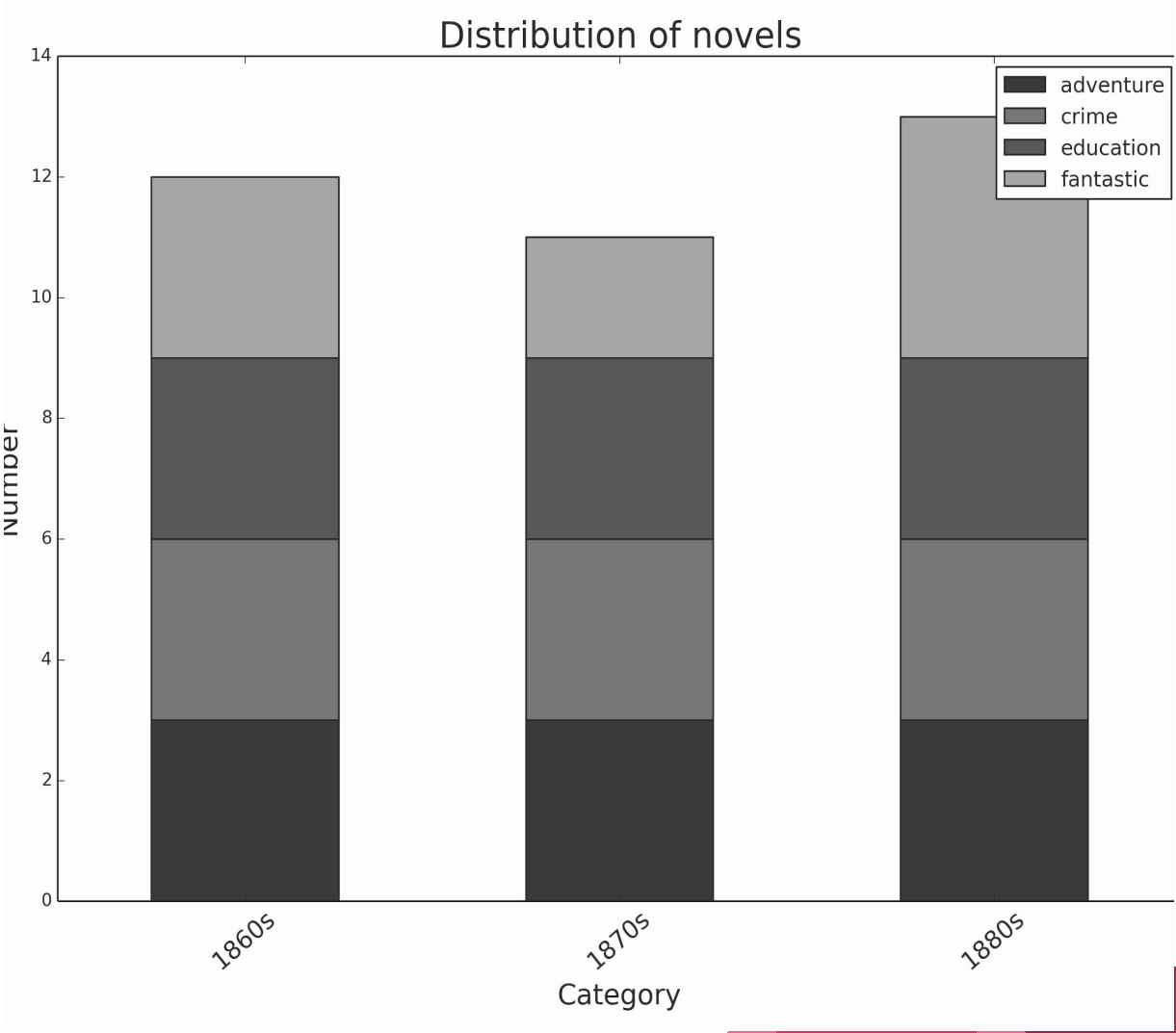


Figure 2: Distribution of novels per subgenre and decade.

Das textbox-Korpus

- 36 Romane aus den Jahren 1861-1889
- vier Untergattungen
- balanciert



3. Methode

Überblick über das Vorgehen

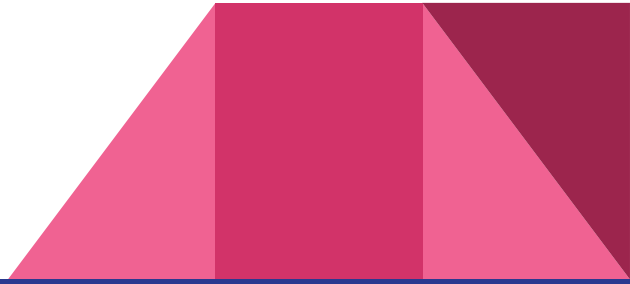
Grundlegender Ansatz: Maschinelles Lernen

Einzelne Schritte:

- Manuelle Annotation eines Teilkorpus (Satz enthält direkte Rede: ja/nein)
- Generieren von relevanten Merkmalen
- Algorithmus “lernt” Beziehung zw. Merkmalen u. Klassen
- Evaluieren der Performanz
- Anwendung der gelernten Beziehung auf übriges Korpus
- Analyse der Verteilungen: Jahrzehnt und Untergattung

Die Typen von Features

- 81 Features modelliert
- verschiedene Kategorien:
  - Zeichenbasiert: Redezeichen, Ausrufezeichen, ...
  - Lexikalisch: Deiktische Ausdrücke, Interjektionen, ...
  - Semantisch: Verbkategorie (WordNet)
  - Morphologisch: Part-Of-Speech, Zeitform, Lemmata, ...
  - Syntaktisch: Anzahl an Kommas, Satzlänge, ...



Performanz auf dem annotierten Teilkorpus

	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**

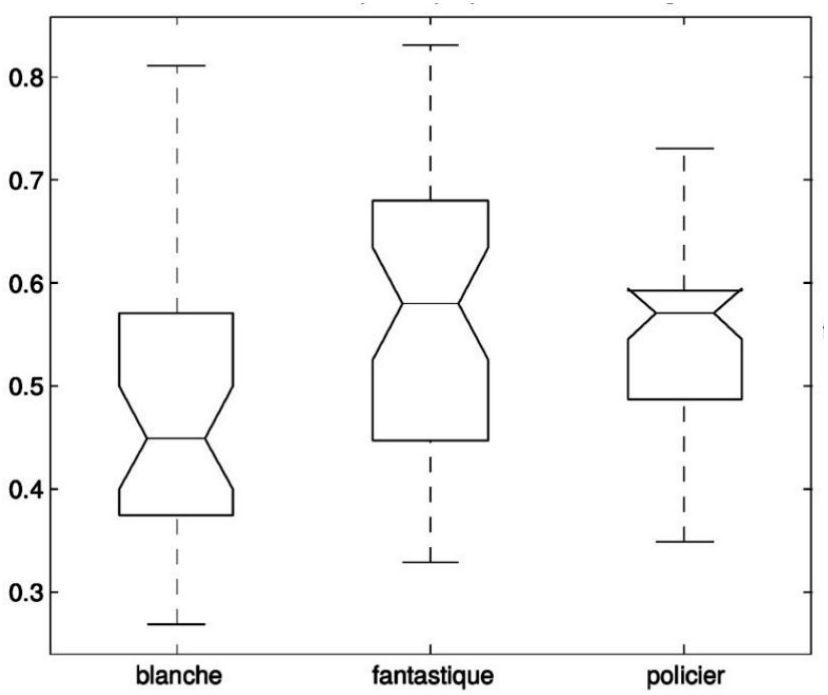
Table 1: Performance (10-fold cross-validation on the gold standard)

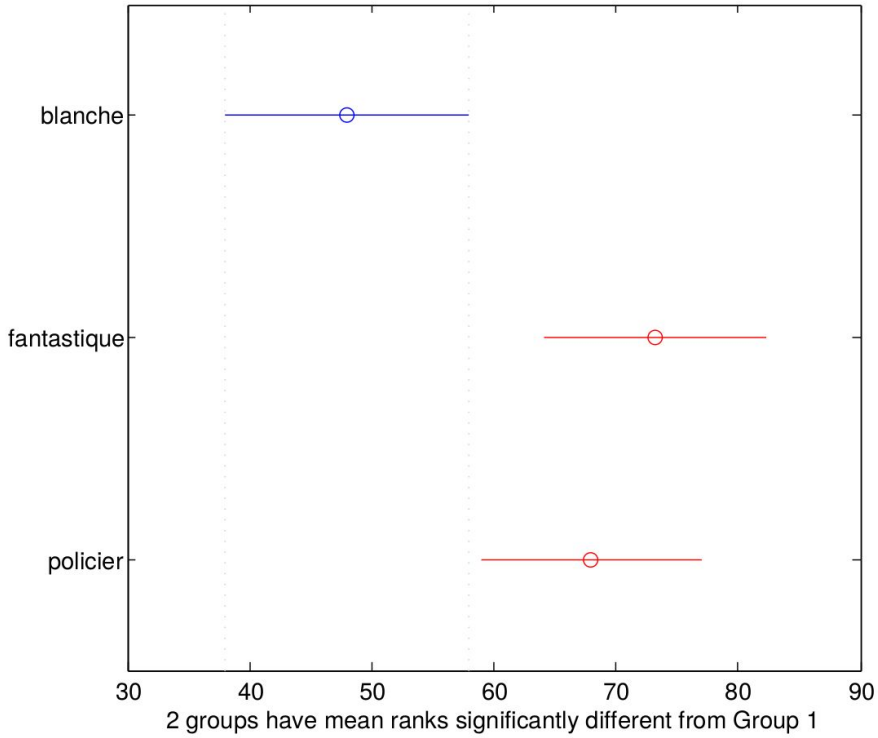
Anwendung des Modells auf das unannotierte Korpus

- Vollständiges Korpus mit gelerntem Modell automatisch annotiert
- 100 Sätze je Dokument zufällig gesamplet
- 15.1% false positives für Direkte Rede
- 16.1% false positives für Erzählerrede
- F1 Score: 0.84
- Probleme bei der Satzsegmentierung (Doppelpunkt) identifiziert

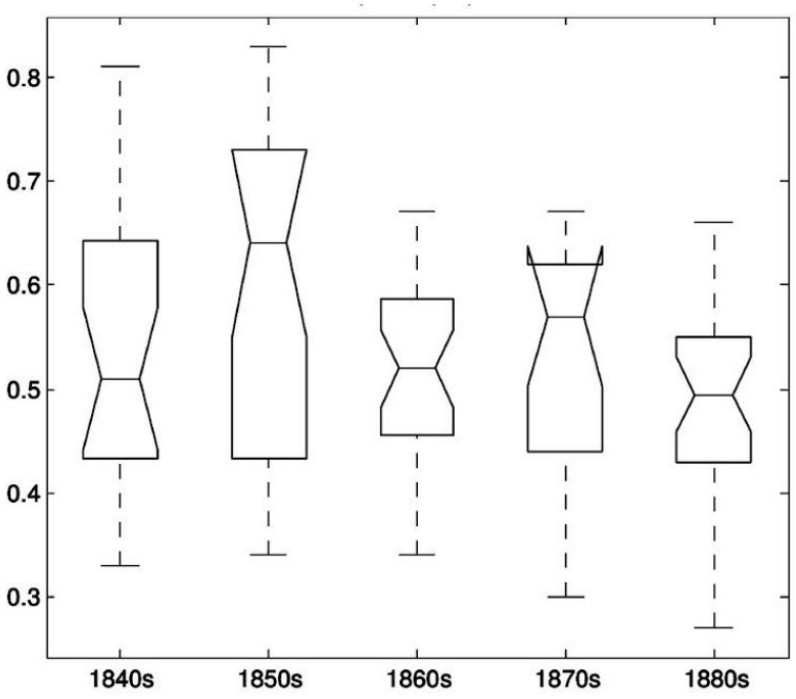
4. Ergebnisse

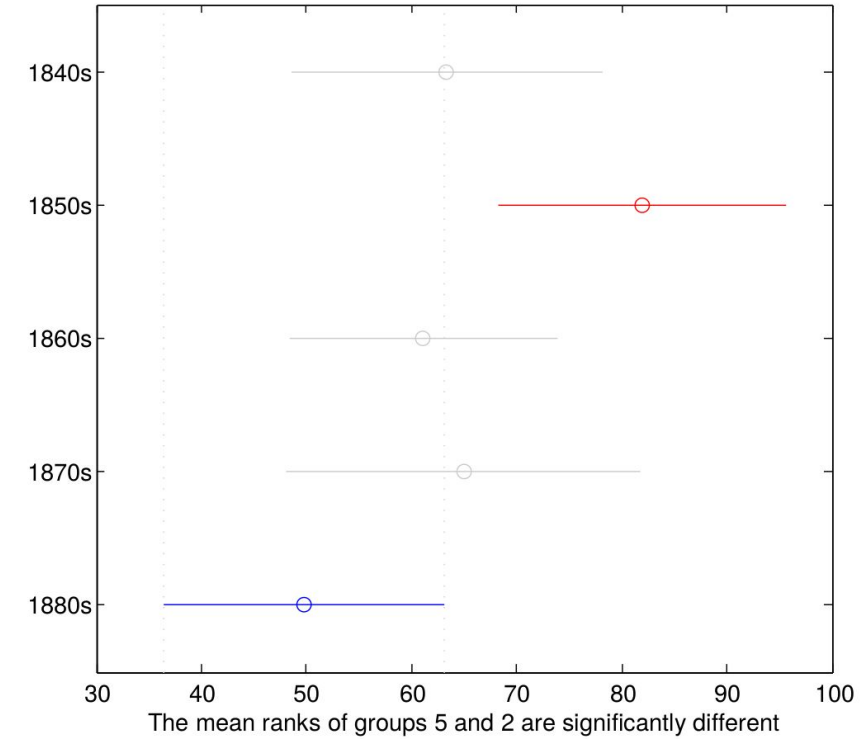
Anteil von direkter Rede nach Genre





Anteil von direkter Rede nach Dekade





Signifikanz und Korpus

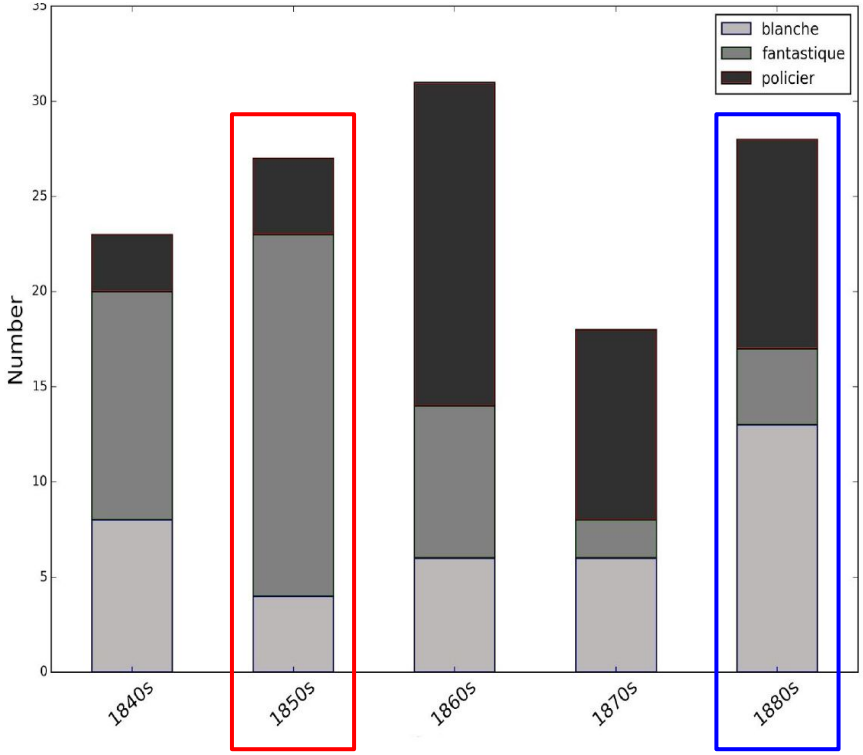
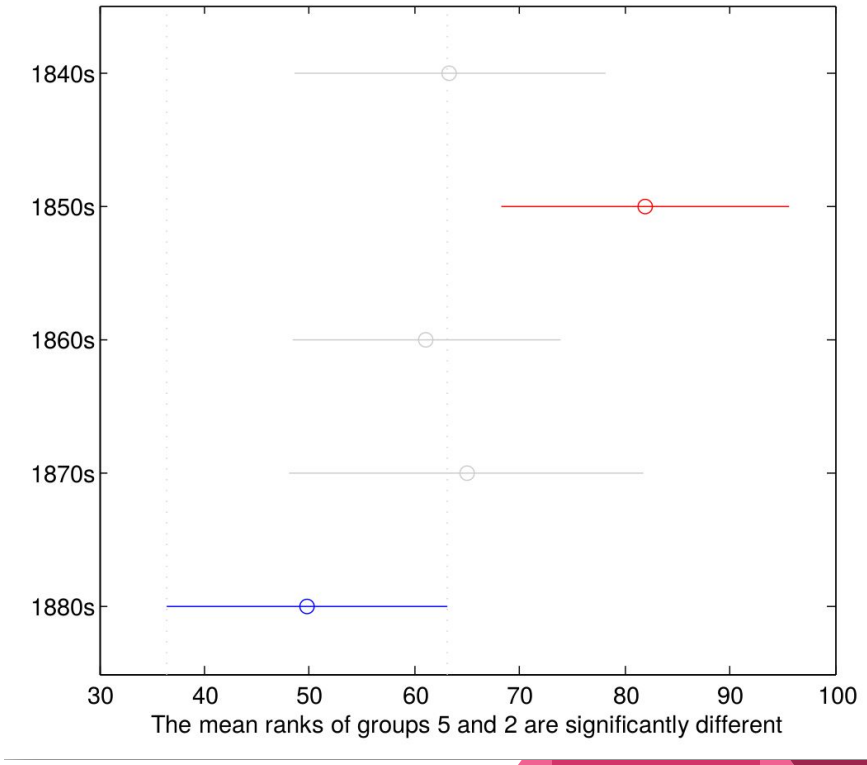
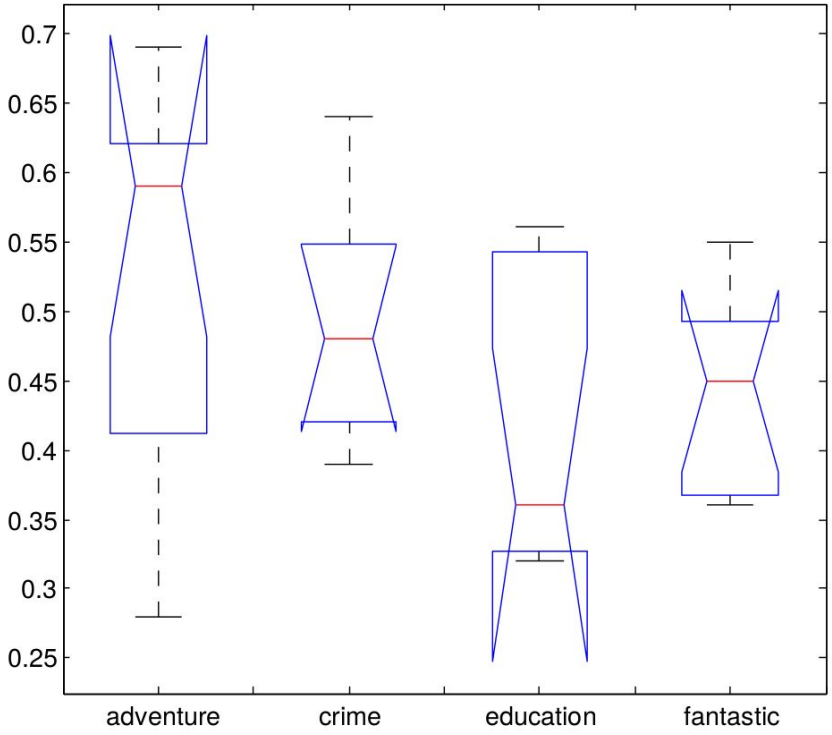
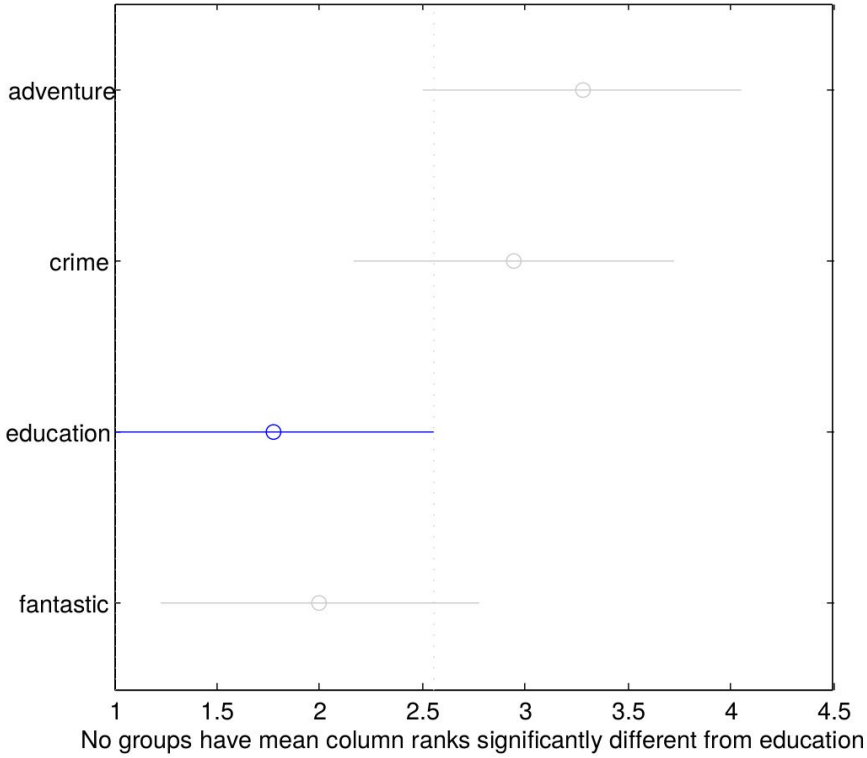


Figure 2: Distribution of novels per subgenre and decade.



textbox-Korpus: Untergattungen

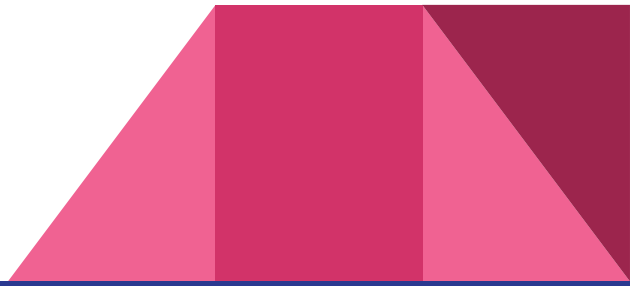




Fazit

Ergebnisse

- Sehr gute / gute Erkennungsrate (0.94 bzw. 0.84) mit dem besten Algorithmus
- Durchschnittlich recht hoher Anteil von Sätzen mit direkter Rede (61%)
- Dekade: keine signifikanten Unterschiede
- Gattung: blanche vs. policier und fantastic



Herausforderungen

- Satzsegmentierung: Präzision und Granularität
- Einschübe berücksichtigen
- Merkmale, die mit der Position im Satz / Absatz zusammenhängen
- Korpuszusammenstellung: balanciert und umfangreich