

Automatische Erkennung von Figuren in deutschsprachigen Romanen

F. Jannidis, M. Krug, I. Reger, M. Toepfer, L. Weimer, F. Puppe
(Universität Würzburg)

Kontext

- Korpusbasierte Geschichte des deutschsprachigen Romans von 1500 bis 1945
- Kumulative Bibliographie der erschienenen Romane
- Romankorpus (rd. 1650 Titel)
- Entwicklung von Gattungen, narrativen Techniken, Motiven

Von *Named Entities* zum Figurennetzwerk

- Interaktion als gemeinsames Vorkommen in einem Absatz

Eduard, Ottilien

Charlotte, Eduard

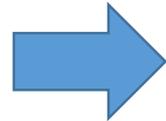
Charlotte, Eduard

Ottilie, Eduard

Eduard, Charlotte

Ottilie, Charlotte

...



Charlotte, Ottilie, 102

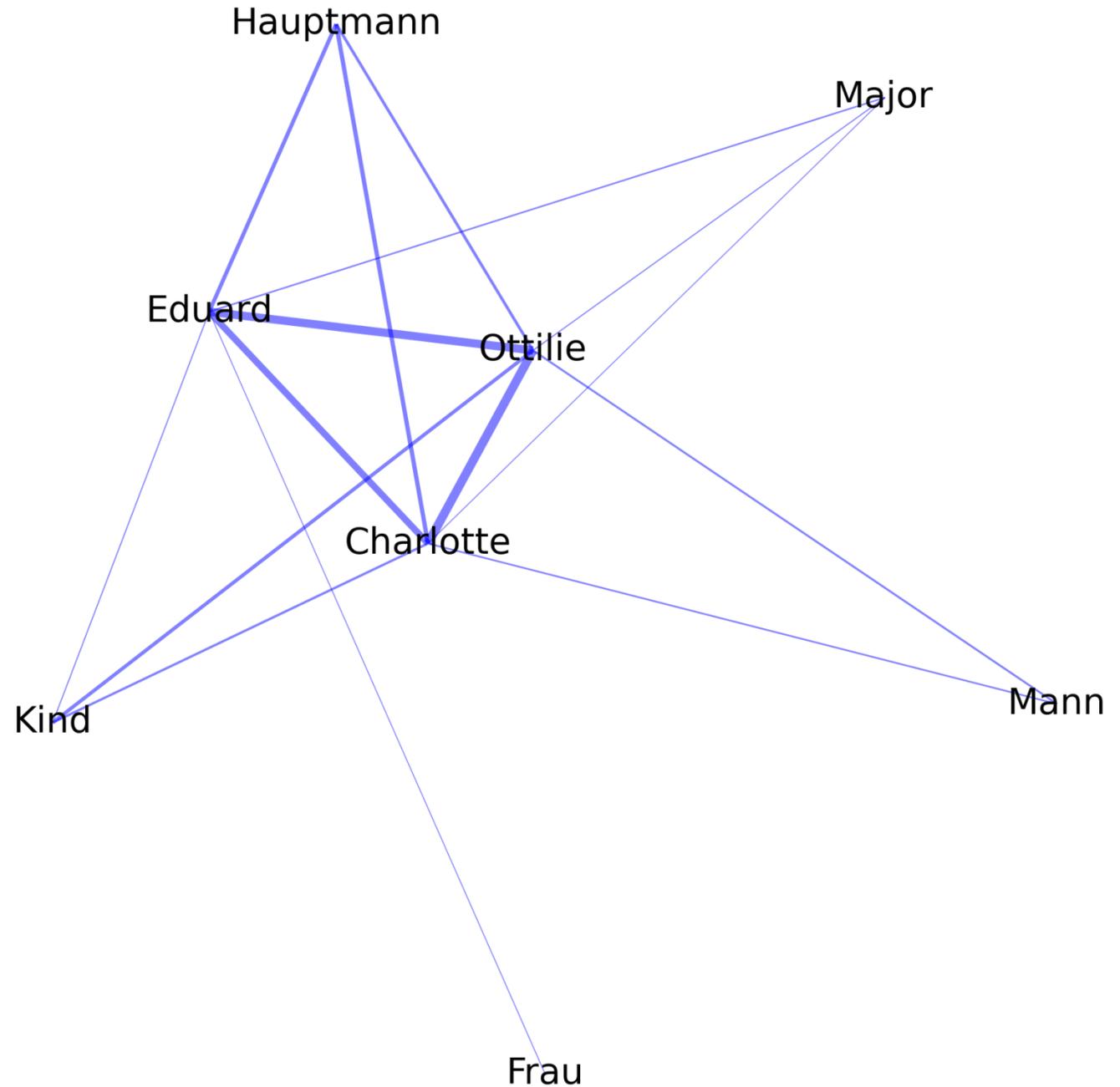
Eduard, Ottilie, 96

Charlotte, Eduard, 77

Charlotte, Hauptmann, 50

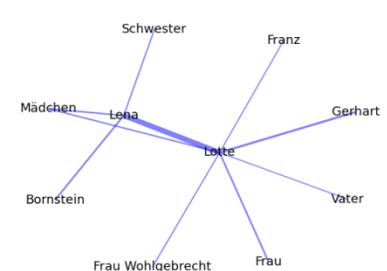
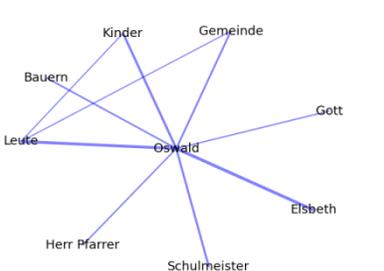
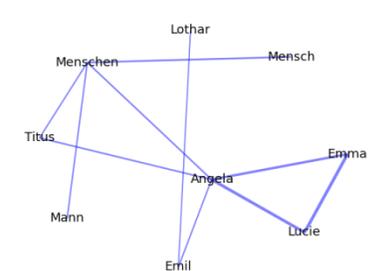
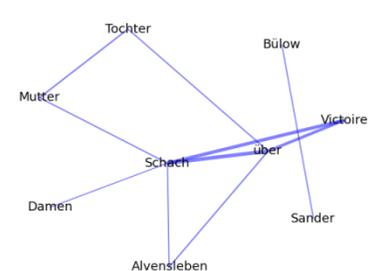
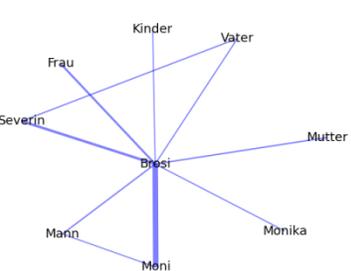
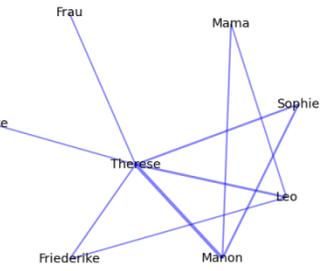
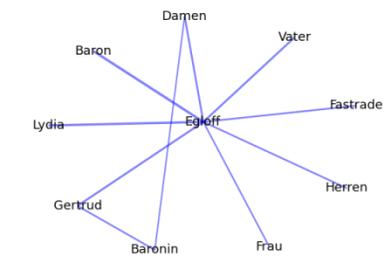
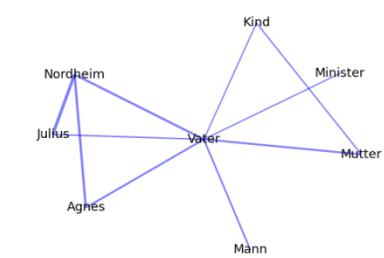
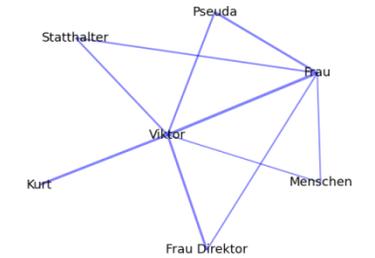
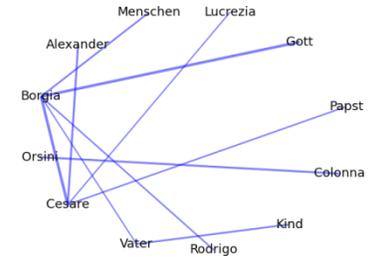
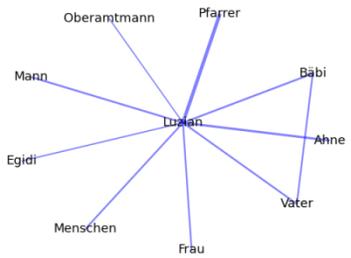
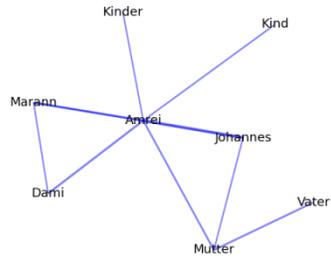
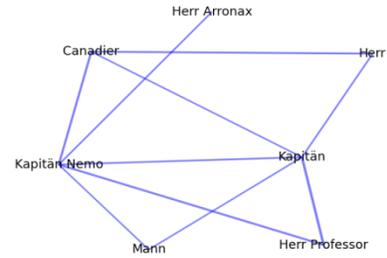
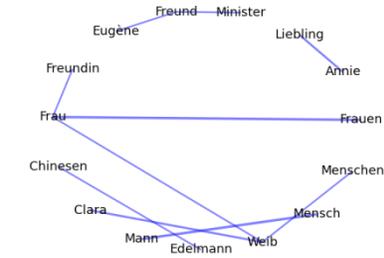
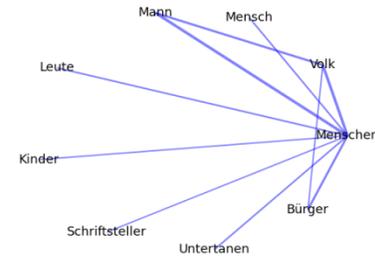
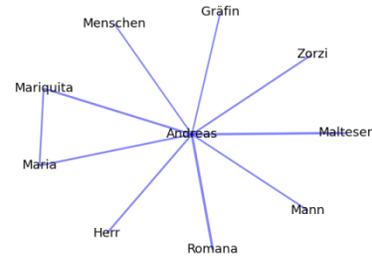
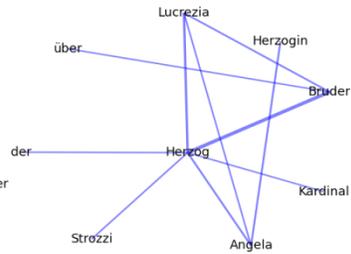
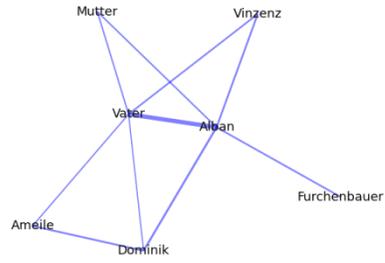
...

Figurennetzwerke



Goethe:
Wahlverwandtschaften

Figurennetzwerke



Stanford NER

Eduard – so nennen wir einen reichen Baron im besten Mannesalter – **Eduard** hatte in seiner Baumschule die schönste Stunde eines Aprilmitttags zugebracht, um frisch erhaltene Pfropfreiser auf junge Stämme zu bringen. Sein Geschäft war eben vollendet; er legte die Gerätschaften in das Futteral zusammen und betrachtete seine Arbeit mit Vergnügen, als der Gärtner hinzutrat und sich an dem teilnehmenden Fleiße des Herrn ergetzte.

»Hast du meine Frau nicht gesehen?« fragte **Eduard**, indem er sich weiterzugehen anschickte.

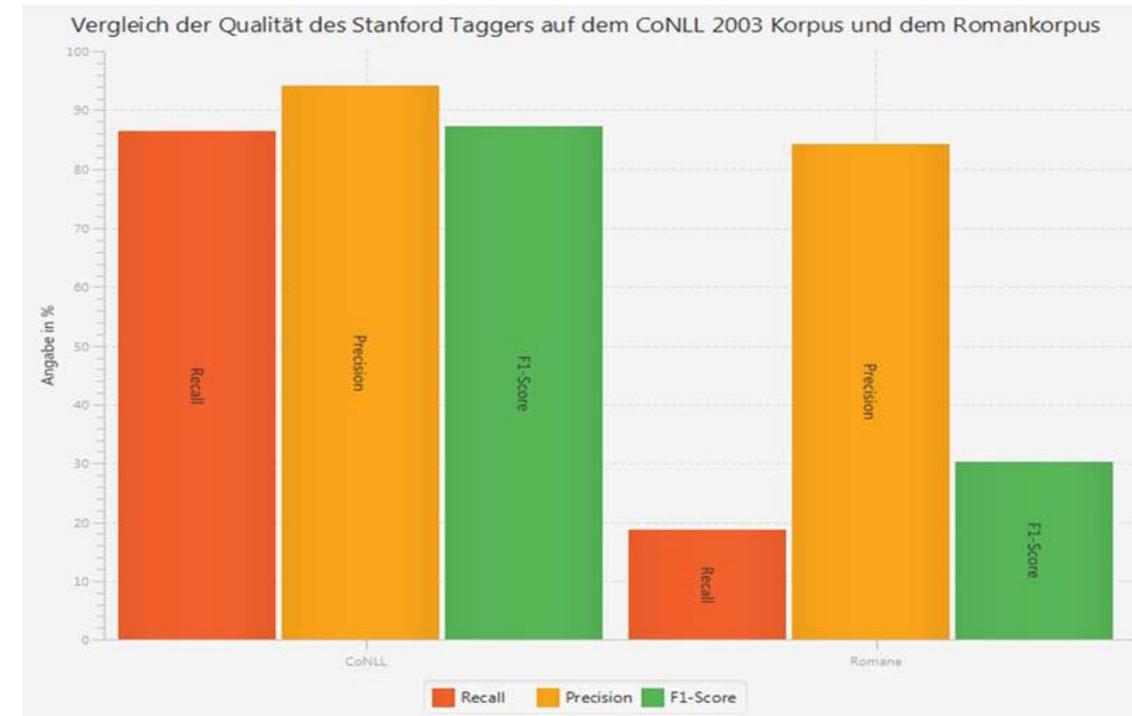
Erweiterte Definition von Named Entity

Eduard – so nennen wir einen reichen **Baron** im besten Mannesalter – **Eduard** hatte in seiner Baumschule die schönste Stunde eines Aprilmittags zugebracht, um frisch erhaltene Pfropfreiser auf junge Stämme zu bringen. Sein Geschäft war eben vollendet; **er** legte die Gerätschaften in das Futteral zusammen und betrachtete seine Arbeit mit Vergnügen, als der **Gärtner** hinzutrat und sich an dem teilnehmenden Fleiße des **Herrn** ergetzte.

»Hast du meine **Frau** nicht gesehen?« fragte **Eduard**, indem **er** sich weiterzugehen anschickte.

Erweiterte Definition von Named Entity

- Tatsächliche Namen
- weitere Figurenreferenzen: Appellativa
 - Berufsbezeichnungen
 - (Adels-)Titel
 - Verwandtschaftsbezeichnungen
 - Äußerlichkeiten: „der Schwarzhaarige“
- StanfordNER: F1-Score von nur 31%



Erstellung eines eigenen Trainingskorpus

- Kooperation mit Informatikern
- Je 130 zusammenhängende Sätze aus 85 Romanen
- Manuelle Annotation
- Unterstützung durch Annotationstool
 - Graphische Benutzeroberfläche
 - Regelbasierte Vorschläge
 - Beschleunigung des Annotationsvorgangs
 - Gleichzeitige Annotation von Entitäten und Koreferenzen möglich

Features für NER*

1. Current Word: das Wort an Position i
2. Previous Word: das Wort an Position $i-1$
3. Next Word: das Wort an Position $i+1$
4. Word Shape: für Groß/Kleinschreibung oder Zahlen
5. Part-Of-Speech Tags (POS-Tags) an den Positionen i , $i-1$ und $i+1$, die mit Hilfe des TreeTaggers [Schmid 1995] bestimmt wurden.
6. Präfix bzw. Suffix, das aus den ersten oder letzten 2 Zeichen besteht.

* analog zum Stanford-Parser.

Neue getestete Features

7. Gazeteers: Listen bestehend aus rd. 5200 männlichen, 3400 weiblichen Vornamen, 160 Adelstiteln, Anreden und 8700 Berufen
8. Semantische Felder (GermaNet)
9. Satzsubjekt (Mate-Dependency Parser)
10. Compound-Words (SFST)
11. Head-Lemma
12. LDA-Cluster (Nähe zu 250 Clustern)
13. Word2Vec-Cluster

Resultate

Verfahren	Precision in %	Recall in %	F1-Score in %	Unterschied zur Baseline (F1-Score) in %
Baseline (Features 1- 6)	95.12	79.60	86.66	+0
Baseline + (7)	95.73	79.28	86.70	+0.04
Baseline +(8)	94.53	81.74	87.65	+0.99
Baseline + (9)	94.96	79.74	86.67	+0.01
Baseline + (10)	95.07	81.00	87.45	+0.79
Baseline + (11)	95.03	79.63	86.63	-0.03
Baseline + (12)	96.47	77.83	86.13	-0.53
Baseline + (13)	94.97	85.28	89.84	+3.18
Baseline + (7),(8),(10),(13)	94.86	85.60	89.98	+3.32

Word2Vec

- Ein Algorithmus für Maschinelles Lernen auf der Grundlage von Neuronalen Netzwerken
- Eingabe: ein sehr großes Textkorpus
- Ausgabe: Vektoren für jedes Wort
- Verfahren:
 - Erstellt Wortliste für das Trainingskorpus
 - Lernt dann eine Vektorrepräsentation für jedes Wort

$\text{vector}(\text{"King"}) - \text{vector}(\text{"Man"}) + \text{vector}(\text{"Woman"})$
results in a vector that is closest to the vector representation of the word
Queen

Nearest words to "France"

Word	Cosine distance
spain	0.678515
belgium	0.665923
netherlands	0.652428
italy	0.633130
switzerland	0.622323
luxembourg	0.610033

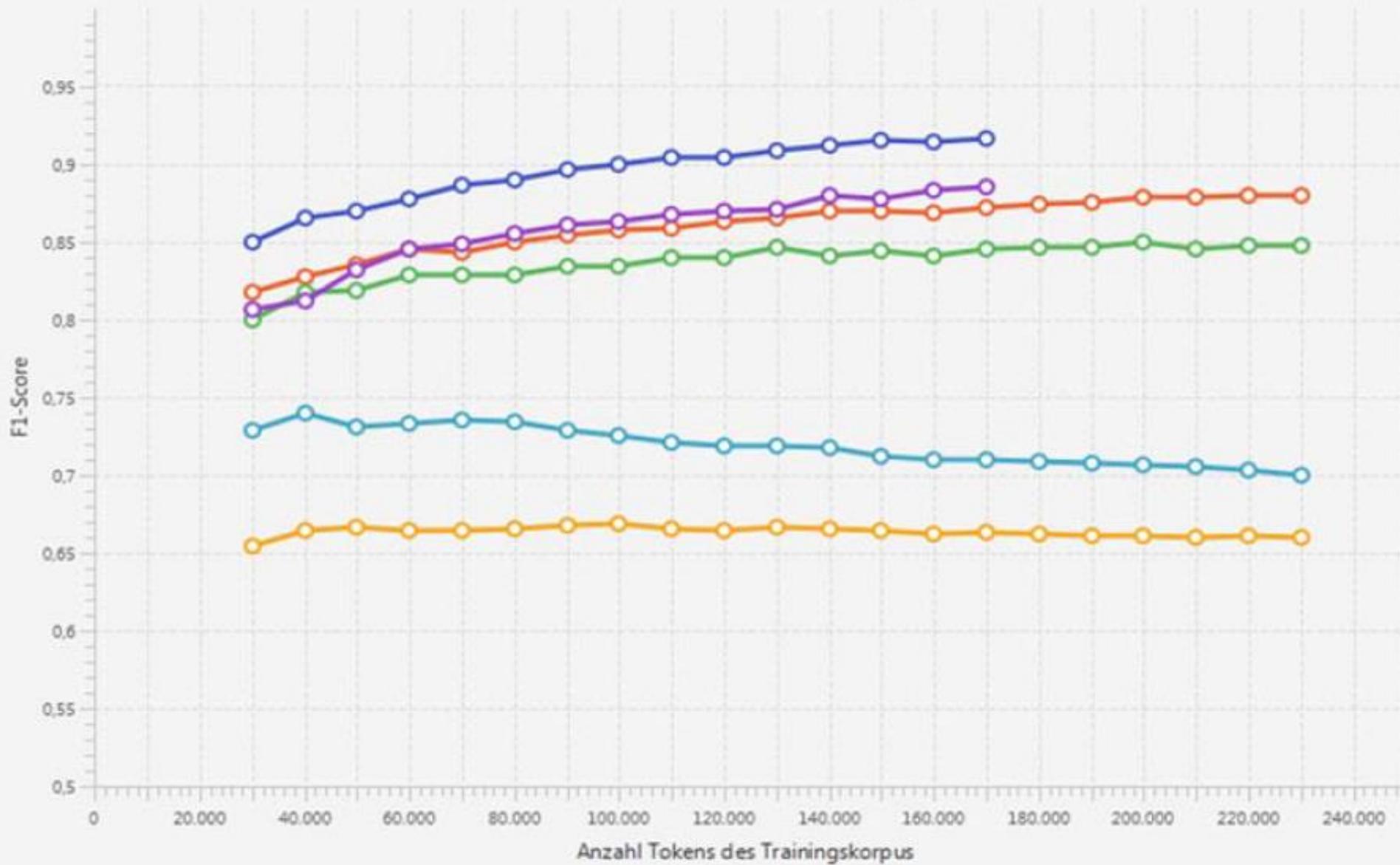
Verwendetes Feature: Word2Vec Cluster

- Ermittlung der Vektoren für jedes Wort aufgrund von Word2Vec
- Clustering der Wort-Vektoren mittels k-means
Clusteranzahl ab 250 (relativ konstant bis 1000)

Zweites Experiment: Größe des Trainingskorpus

- Domain-Anpassung als Hauptproblem der textwissenschaftlichen Verwendung von NLP-Werkzeugen
- Optimierung der Domain-Anpassung zwischen F1-Maximierung und pragmatischen Einschränkungen

NE-Erkennung im Verhältnis zur Trainingsmenge



∅ Satzlänge Romane: 24,2
∅ Satzlänge Zeitung: 16,3

- MaxEnt (Romane)
- Decision-Tree (Romane)
- CRF (Romane)
- Naive-Bayes (Romane)
- CRF (CoNLL)
- MaxEnt (CoNLL)

Zusammenfassung

- Domain Adaption
 - Figurenreferenzen -> Named Entities + Appellativa
 - Erstellung eines eigenen Korpus - Kann deutlich kleiner sein
- Algorithmus-Verbesserung
 - Verwendung von Word2Vec + Features des Stanford-Parsers

Software, Korpus

- Software als Teil von DKPro Core verfügbar; siehe: <https://github.com/MarkusKrug/NERDetection/>
- Trainingskorpus wird noch überarbeitet: Publikation voraussichtlich im Sommer/Frühherbst; siehe:
- <http://www.kallimachos.de/doku.php/kallimachos:leserlenkung:start>