Thomas Bartz, Christian Pölitz

Routineaufgaben bei der Nutzung von Korpora: Disambiguieren, Klassifizieren, Annotieren mit KobRA-Verfahren

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

4. KobRA-Tagung, 30.10.2015, BBAW, Berlin

Technische Universität Dortmund
Institut für deutsche Sprache und Literatur und Lehrstuhl für Künstliche Intelligenz

Gliederung

1. Korpus-basierte Sprachanalyse: Forschungsprozess und Aufgaben
2. KobRA-Verfahren für die Korpus-basierte Sprachanalyse
3. Was leisten die Verfahren aktuell: Nutzwert
   Studien aus dem Bereich Lexikographie/historische Semantik
4. Zusammenfassung

Korpus-basierter Forschungsprozess

Wann ändert sich die prototypische Bedeutung von *toll*?

1 Wie häufig ist *weil* ohne Verbletztstellung auf Wikipedia-Diskussionsseiten

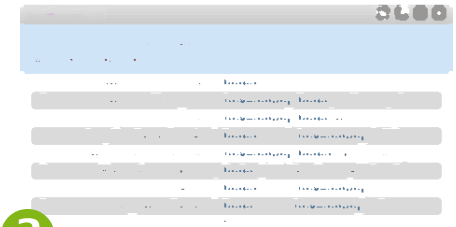
Wie häufig werden Stützverbgefüge in unterschiedlichen Textsortenbereichen verwendet?

Forschungsfrage

2 **Suchanfrage**

Korpus-Schnittstelle

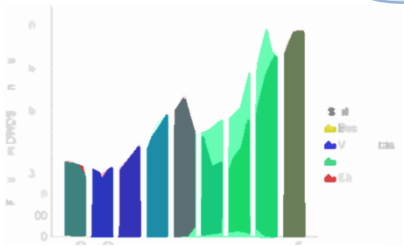
- Form- und Phrasensuche, Abstands- und Verknüpfungsoperatoren
- Reguläre Ausdrücke
- Linguistische Informationen: Lemmata, Wortarten, Syntax



3 **Analyse der KWIC-Trefferliste**

Annotations-/Analysesoftware

- Aussortieren falsch positiver Treffer
- Annotieren und Klassifizieren
- Disambiguieren



4 **Visualisierung/ Exploration**

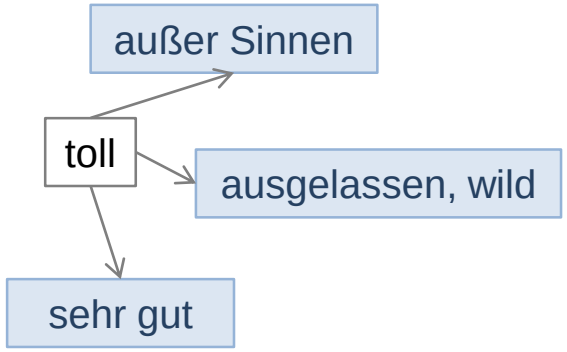
Statistik-/Visualisierungssoftware

- Überführen in Häufigkeitstabellen
- Berechnung relativer Häufigkeiten/statistischer Kennwerte
- Erzeugen von Visualisierungen

Aussortieren falsch positiver Treffer 3

Aufgabenstellung

Trefferlisten zu einem sprachlichen Phänomen enthalten häufig noch viele falsch positive Treffer, die vor weitergehenden Analysen ausgesondert werden müssen.



KWIC-Snippets aus dem diachronen Korpus TüBa-D/DC (‚Projekt Gutenberg')

... in de Koje wa dat Soltwater twee	**toll**	af an de Planken klappt. Dat is ...
… der letzten Woche ad sidera	**tollere**	vultus, d.h., die Nase so zu …
… Es ward nicht aufgetan.	**Toller**	telegraphierte aus dem …
… dem alten verbummelten Lehrer	**Toll**	nicht einmal als etwas …
… Mythologie. Worauf sich das	**toller**	beziehet, ist ungewiß; denn …
… zu still die Welt. Manchmal zu	**toll**	.
… von Angouleme. Talleyrand	**toll**	? Ich weiß nicht.

Sprache

Namen

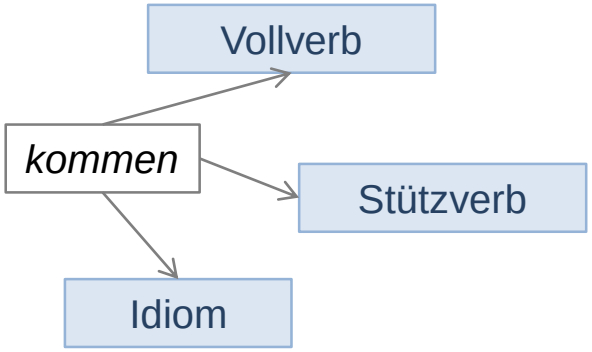
Metasprache

Kontext

Annotieren und Klassifizieren ❸

Aufgabenstellung

Sprachliche Phänomene lassen sich häufig nicht zielgenau erheben. Trefferlisten müssen daher oft nachklassifiziert/annotiert werden.



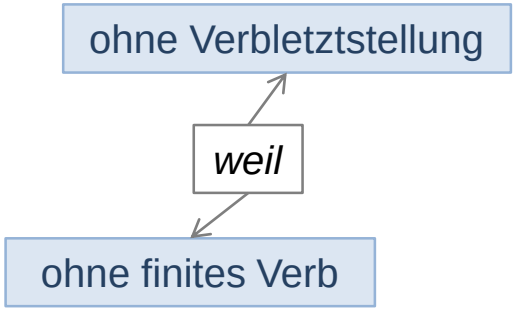
KWIC-Snippets aus dem DWDS-Kernkorpus des 20. Jh. (BBAW)

… Portwein erstickte. Es	**kamen**	Die Damen Buddenbrook, die …	X		
… führt. **Zum Einsatz**	**kommen**	hier fünf Dieseltriebwagen vom …		X	
… darüber. Diese Worte	**kamen**	ihm einfach **über die Lippen** …			X
… Ruhe. Frau Brigitte	**kam**	noch einmal herauf und holte …			
… ich, Ingenieur? Was	**kommt**	mir gerüchteweise **zu Ohren**? …			
… Luft **in Berührung**	**kamen**.	Einzig die Knie blühten dunkel …			
… endlich **zum Stehen**	**kam**.	Rote Leuchtbomben, aus denen …			

Annotieren und Klassifizieren 3

Aufgabenstellung

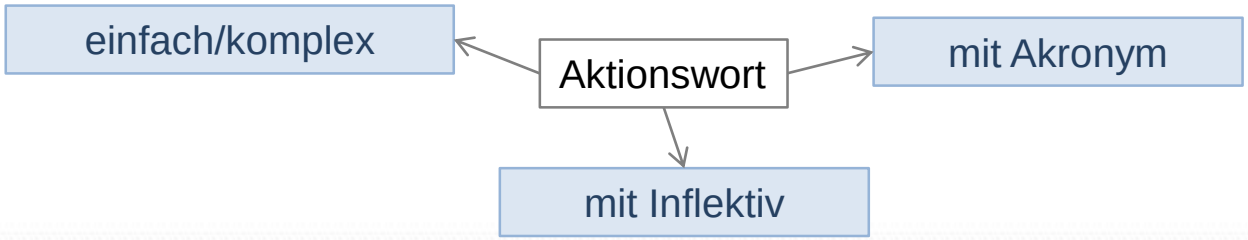
Häufig erfordert auch ein Analyseschritt eine feinere Klassifizierung der Trefferlisten nach vorgegebenen Merkmalen.



KWIC-Snippets aus Wikipedia-Korpus (IDS), Diskussionsseiten

			kein Verbletzt	kein finites Verb
... So, eigentlich überflüssig,	**weil**	eine Wiederholung, allerdings ...		X
... wenn auch nur ein Anfang -	**weil**	hier geht es ja nicht um die …	X	

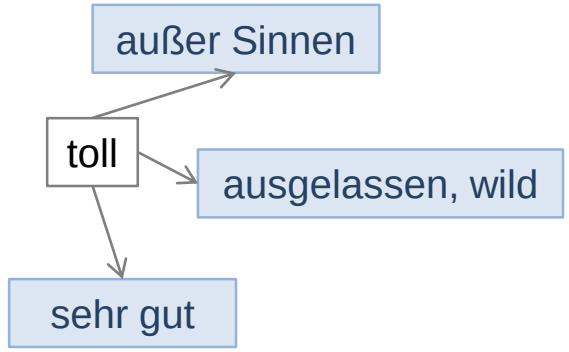
	Aktionswort	Akronym	komplex	Inflektiv
… ***ganzganzfiesgrins*** da ergibt sich ne tolle gelegenheit für …	X		X	X
… und da ist irgendwo hier verschollen bei mir ***g***. Und was …	X	X		
… Kram doch in deutscher Sprache veröffentlichen ***wunder*** …	X			X



Disambiguieren ❸

Aufgabenstellung

Gesuchte Ausdrücke kommen häufig in verschiedenen Bedeutungen vor. Trefferlisten müssen daher meist erst disambiguiert werden.



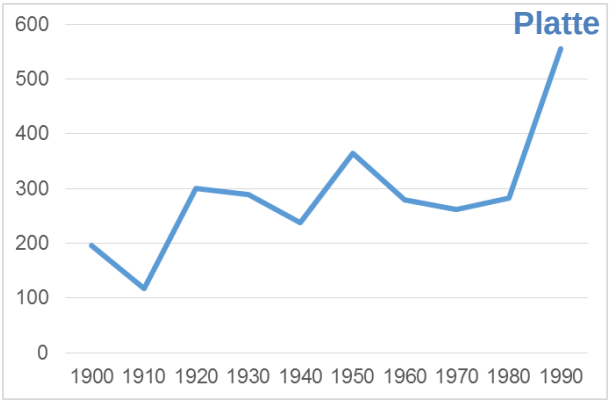
KWIC-Snippets aus dem diachronen Korpus TüBa-D/DC („Projekt Gutenberg‘)

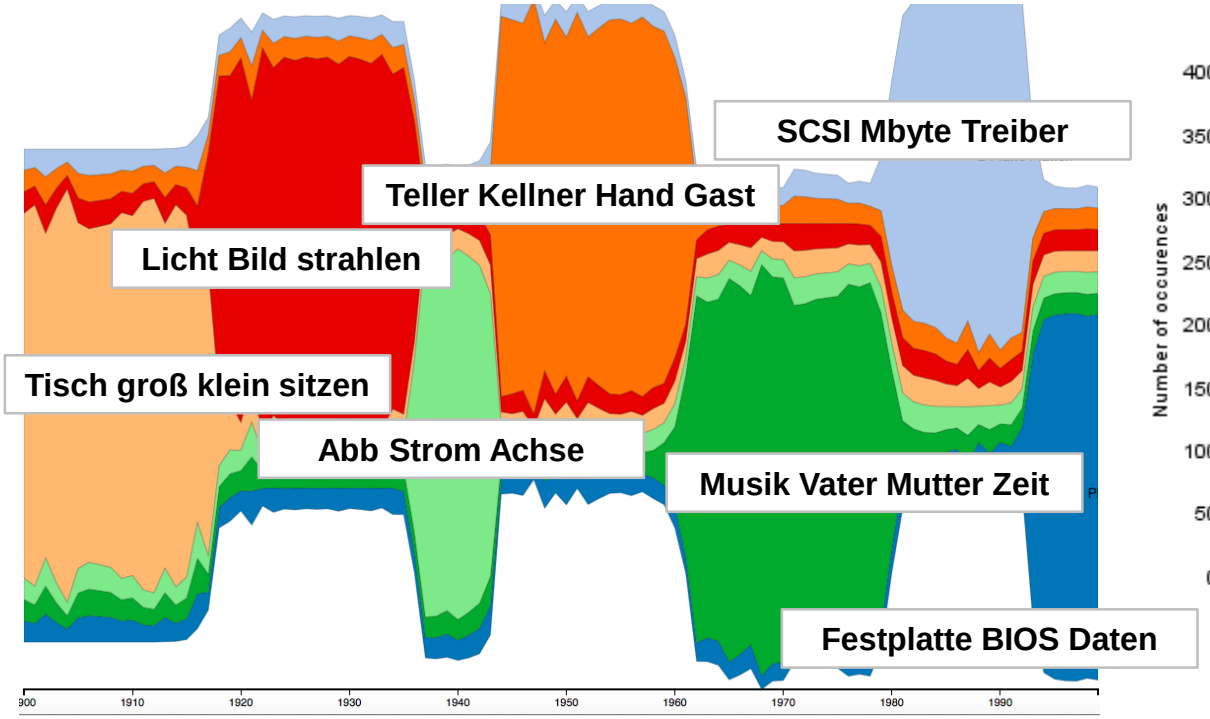
… Billy, den er als kleines Fohlen in	**toller**	Hetze mit dem Lasso …	wild
… Nun waren sie	**toll**	vor Freude, daß sie …	ausgelassen
… Sechzig Centimes? Du bist wohl	**toll**?	Für das Geld kriege …	außer Sinnen
… Außerdem wird es mit den	**tollen**	Aufstiegschancen …	sehr gut
… Es war, als wollte sich alles wie	**toll**	in einen Abgrund stürzen …	
… geschickt. Ihre Gebieterin war	**toll**	vor Liebe, berauscht vor Freude …	
… Witwer; auf der Insel wurden	**tolle**	Geschichten über sein Eheleben …	
… kennt er doch beide? das wäre ja	**toll**	und unvernünftig. Allein wenn es …	

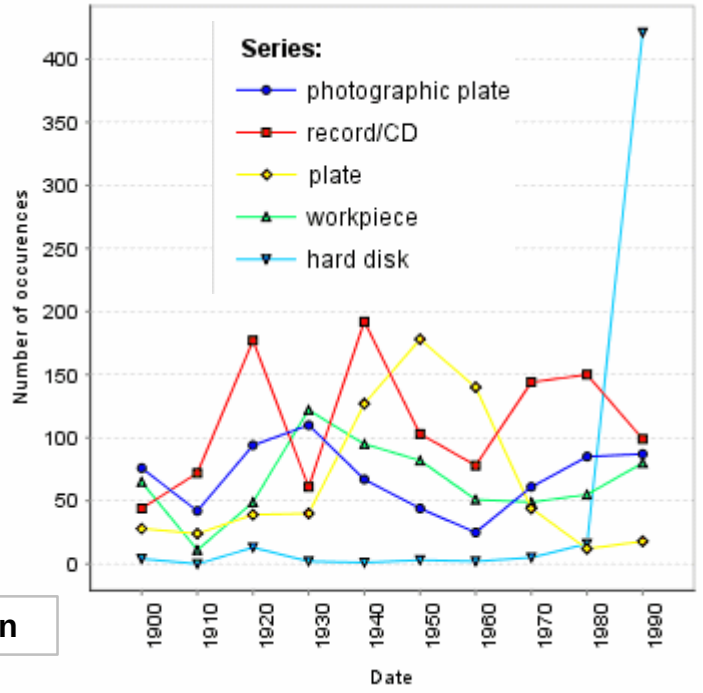
Erzeugen von Visualisierungen 4

Aufgabenstellung

Bei der Visualisierung von Häufigkeitsverteilungen müssen die verschiedenen Bedeutungen eines Ausdrucks berücksichtigt werden.



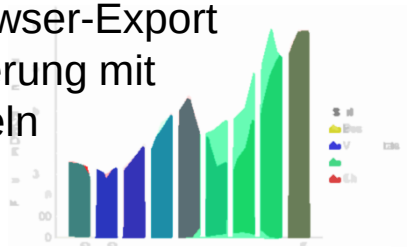




Korpus-basierter Forschungsprozess mit KobRA-Verfahren

Exportoperatoren/ Visualisierung

- CSV-, Excel®-Export etc.
- DFR-Browser-Export
- Visualisierung mit Bordmitteln



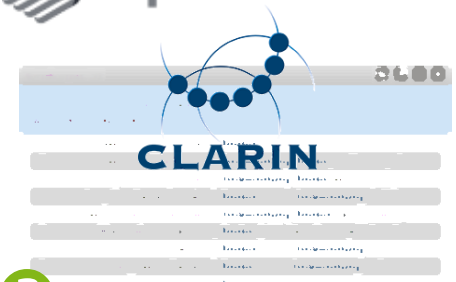
1

Forschungsfrage

2

Suchanfrage

rapidminer



Anfrage-/Importoperatoren

- Abfrage von CLARIN-Ressourcen
- CSV-, Excel®-, (TEI-)XML-Import etc.

4

Visualisierung/ Exploration

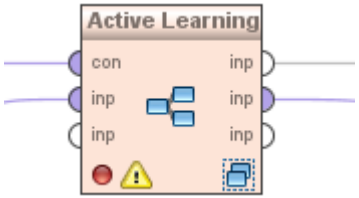
3

Analyse der KWIC-Trefferliste

Annotationsoperator

- Erstellung von Trainingsdaten
- Evaluation

Data-Mining-Operatoren



- Klassifikation
- Disambiguierung

Operatoren für die Vorverarbeitung

- linguistische Annotation, z. B. POS-Tags, Syntax

Was leisten die Verfahren?

Gütemaß: F_1-Score auf Basis manuell annotierter Stichproben

Gewichtetes harmonisches Mittel aus Präzision (Precision) und Ausbeute (Recall), Präzision und Ausbeute werden gleich gewichtet

$$F_1 = 2 \times \frac{\text{Präzision} \times \text{Ausbeute}}{\text{Präzision} + \text{Ausbeute}}$$

Überprüfte Gütefaktoren

Klassifikation

- Menge der (annotierten) Datensätze
- Größe der KWIC-Snippets
- Repräsentation der Daten:
  - Bag-of-Words
  - POS/Syntax
  - Distributionelle Merkmale (‚Expertenmerkmale')

Disambiguierung

- Menge der Datensätze
- Größe der KWIC-Snippets (Kontext)
- Abgefragtes Wort/Wortart
- Menge der Bedeutungen
- Korpus: synchron/diachron
- Sprache

Merkmal-Kategorie	Merkmale: Das Stützverb ...	Beispiele
B1 (Wortart)	ist ein Vollverb (*VVFIN*)[6]	
	oder tritt in einem Satz als Partizip (*VVPP*) zusammen mit einem Hilfsverb (*VAFIN*) auf	1.1.2 1.2.2
	oder tritt in einem Satz als Infinitiv (*VVINF*) zusammen mit	1.1.5

Stützverbgefüge: Nur im ‚Juristendeutsch‘?

Annotieren und Klassifizieren

KWIC-Snippets aus dem DWDS-Kernkorpus des 20. Jh. (BBAW)

			Vollverb	Stützverb
… Portwein erstickte. Es	**kamen**	Die Damen Buddenbrook, die …	X	
… führt. **Zum Einsatz**	**kommen**	hier fünf Dieseltriebwagen vom …		X

	Belletristik	Gebrauchsliteratur	Wissenschaft	Zeitung
Anfrage kommen 1% manuell analysiert	71.399	26.068	23.924	33.703
F_1 Bag-of-Words	72,4	65,4	74,9	67,5
F_1 POS-Tags	74,5	71,1	76,9	67,7
F_1 Distributionelle Merkmale	71,6	67,5	69,4	63,7
Anfrage bringen 1% manuell analysiert	18.006	14.301	12.653	19.669
F_1 Bag-of-Words	69,7	67,6	72,6	70,3
F_1 POS-Tags	67,1	67,2	76,8	71,7
F_1 Distributionelle Merkmale	65,4	66,9	63,5	65,5

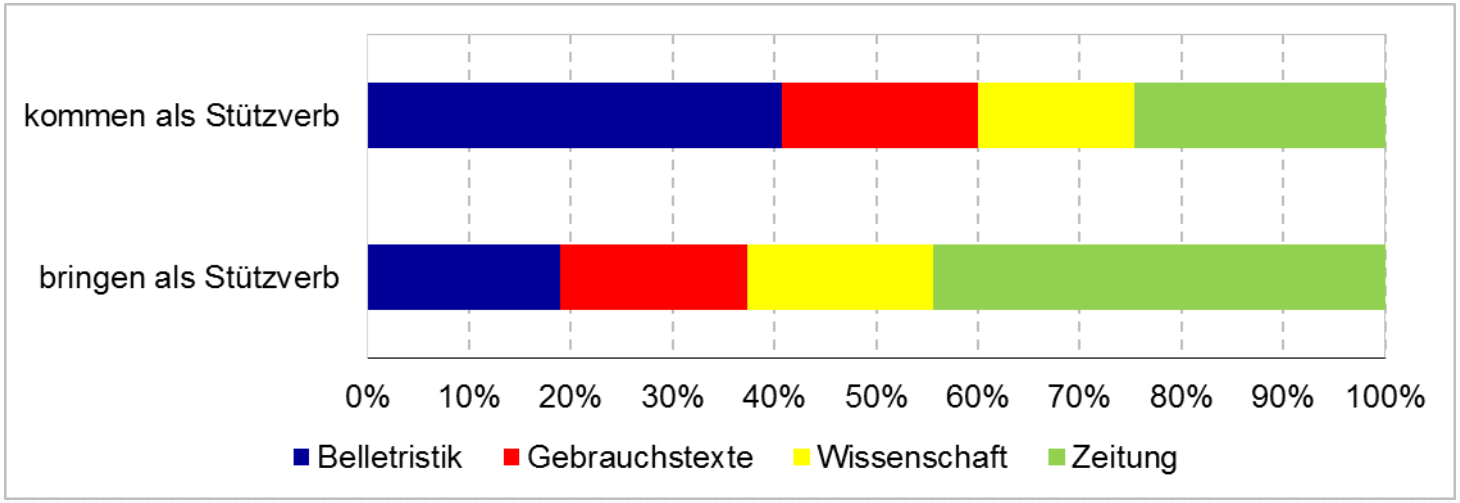
Stützverbgefüge: Nur im ‚Juristendeutsch‘?

Annotieren und Klassifizieren

KWIC-Snippets aus dem DWDS-Kernkorpus des 20. Jh. (BBAW)

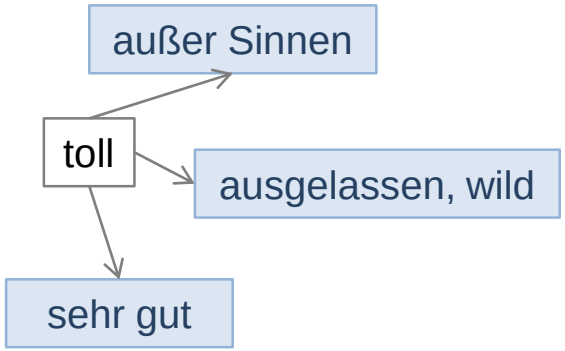
			Vollverb	Stützverb
… Portwein erstickte. Es	**kamen**	Die Damen Buddenbrook, die …	X	
… führt. **Zum Einsatz**	**kommen**	hier fünf Dieseltriebwagen vom …		X

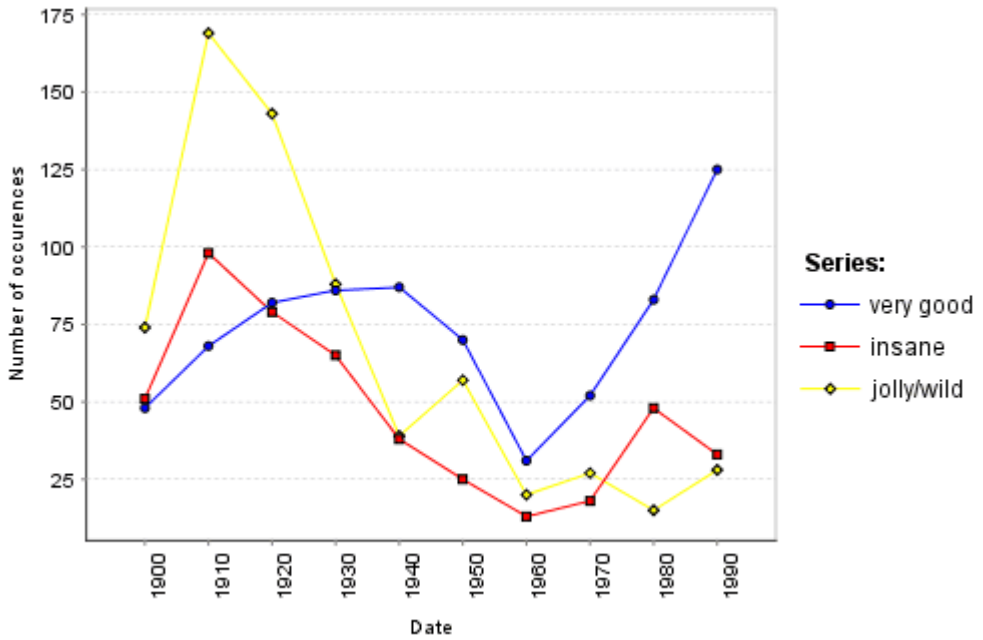
	Belletristik	Gebrauchsliteratur	Wissenschaft	Zeitung
***kommen* als Stützverb**	15.179	7.187	5.723	9.181
***bringen* als Stützverb**	3.131	3.046	3.025	7.354



toll: Umwertung der Bedeutung?

Disambiguieren





KWIC-Snippets aus dem DWDS-Kernkorpus des 20. Jh. (BBAW)

	Snippets	*außer Sinnen*	*ausgelassen, wild*	*sehr gut*
Anfrage toll with $p=ADJ*	2.106	702	772	632
F_1 Kontext 20 Wörter		0,519	0,571	0,167
F_1 Kontext 30 Wörter		0,714	0,615	0,632
F_1 Kontext 40 Wörter		0,625	0,667	0,500
IAA 30% manuell disambiguiert	0,76			

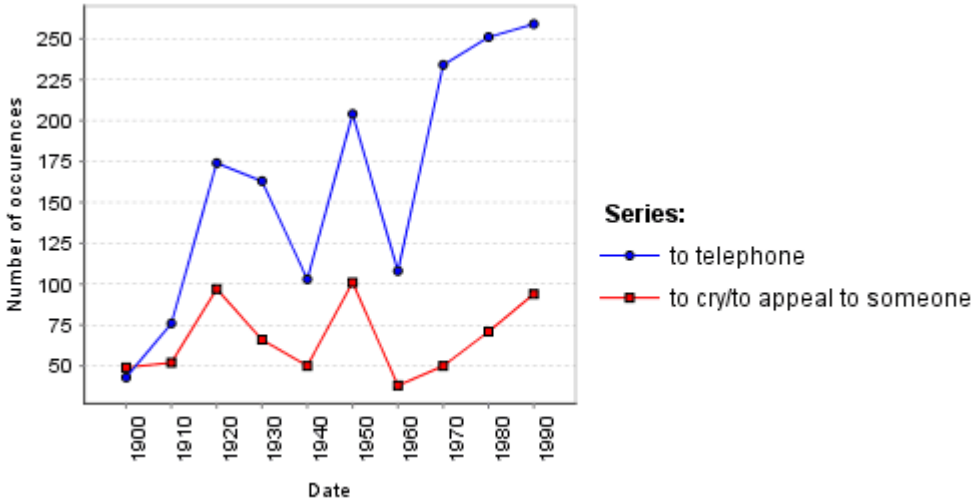
anrufen: Verschiebung der prototypischen Bedeutung?

Disambiguieren

jmdn. bitten, auf sich aufmerksam machen

anrufen

telefonieren

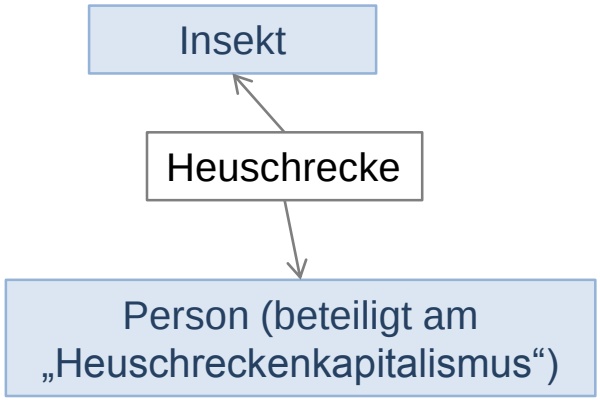


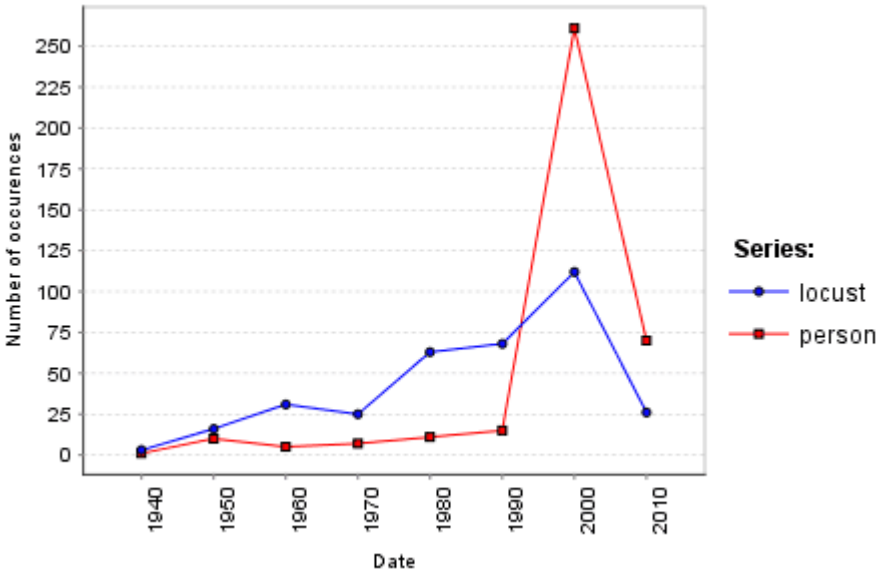
KWIC-Snippets aus dem DWDS-Kernkorpus des 20. Jh. (BBAW)

	Snippets	*bitten etc.*	*telefonieren*
Anfrage anrufen with $p=VV*	2.958	1.183	1.775
F_1 Kontext 20 Wörter		0,727	0,667
F_1 Kontext 30 Wörter		0,800	0,800
F_1 Kontext 40 Wörter		0,909	0,889
IAA 30% manuell disambiguiert	0,97		

Heuschrecke: Metapher als Quelle für Neubedeutungen?

Disambiguieren



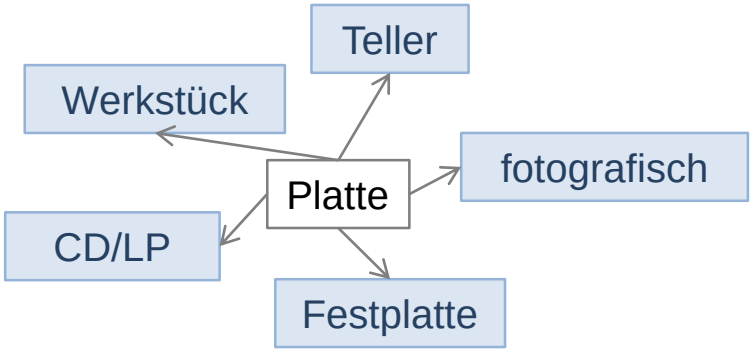


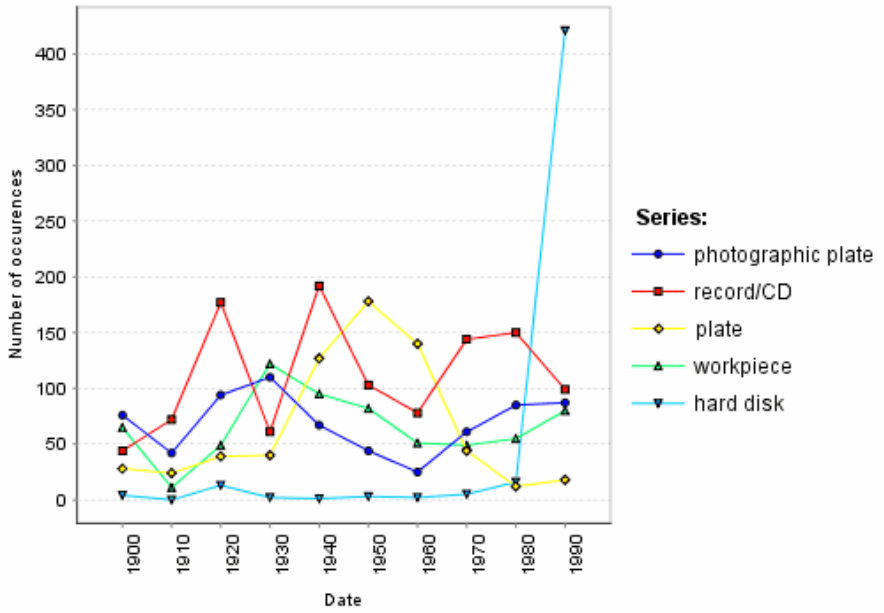
KWIC-Snippets aus dem ZEIT-Korpus (BBAW)

	Snippets	*Insekt*	*Person*
Anfrage Heuschrecke	693	139	554
F_1 Kontext 20 Wörter		0,857	0,842
F_1 Kontext 30 Wörter		0,800	0,933
F_1 Kontext 40 Wörter		0,667	0,727
IAA 30% manuell disambiguiert	0,98		

Platte: Technische Entwicklung/Bedeutungsentwicklung

Disambiguieren



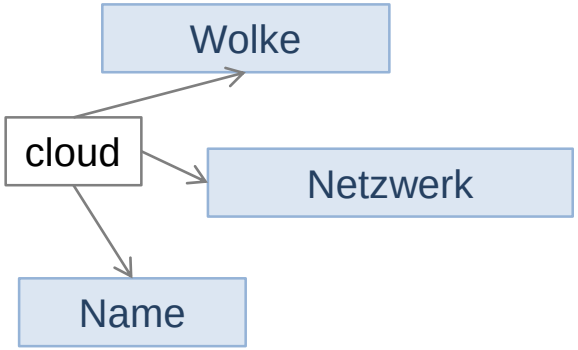


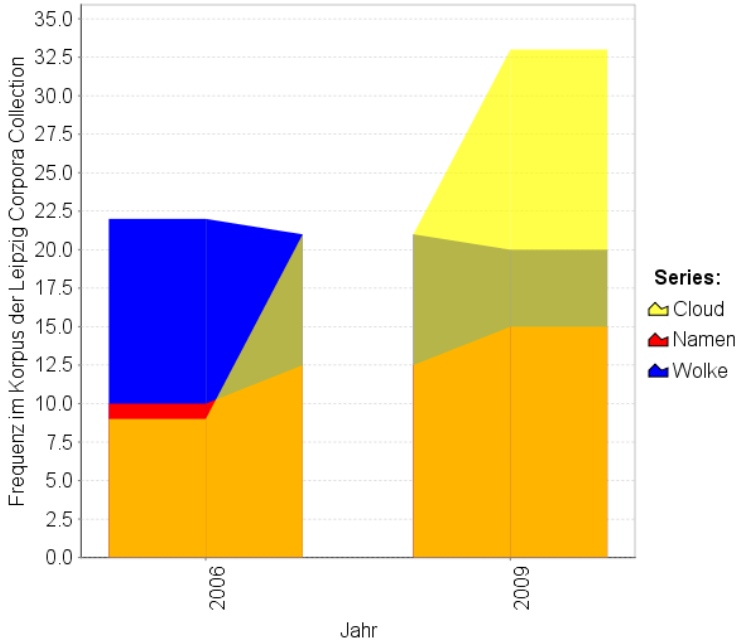
KWIC-Snippets aus dem DWDS-Kernkorpus des 20. Jh. (BBAW)

	Snipp.	*Werkstück*	*Teller*	*fotografisch*	*CD/LP*	*Festpl.*
Anfrage Platte with $p=NN && ! @Platt && ! $l=Platter	2.008	373	387	357	440	451
F_1 Kontext 20 Wörter		0,800	0,800	0,667	0,287	0,857
F_1 Kontext 30 Wörter		0,984	0,875	0,500	0,381	0,983
F_1 Kontext 40 Wörter		0,733	0,600	0,750	0,353	0,800
IAA 30% manuell disamb.	0,98					

cloud: Beispiel aus dem Englischen

Disambiguieren





KWIC-Snippets aus der Leipzig-Corpora-Collection (englisch)

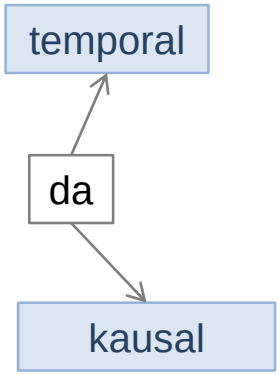
	Snippets	*Wolke*	*Netzwerk*	*Name*
Anfrage cloud	1.486	966	280	240
F_1 Kontext 20 Wörter		0,526	0,500	0,471
F_1 Kontext 30 Wörter		0,783	0,631	0,615
F_1 Kontext 40 Wörter		0,467	0,545	0,684
IAA 30% manuell disambiguiert	0,92			

da: Ab wann auch kausal?

Disambiguieren

KWIC-Snippets aus dem diachronen Korpus TüBa-D/DC (‚Projekt Gutenberg')

temporal	… losgebrochen,	**da**	sie eben das Haus erreicht …
kausal	… abschlagen,	**da**	ich an der Sache beteiligt bin …



	Snippets	***temporal***	***kausal***
Anfrage da with $p=KOUS	123.496	67.923	55.573
F_1 Kontext 20 Wörter		0,471	0,556
F_1 Kontext 30 Wörter		0,353	0,529
F_1 Kontext 40 Wörter		0,400	0,611
IAA 30% manuell disambiguiert	0,75		

Zusammenfassung

- Mithilfe der erprobten Verfahren in RapidMiner lassen sich **Routineaufgaben bei der korpusbasierten Sprachanalyse** wie z.B. das Klassifizieren oder Disambiguieren von KWIC-Snippets eines gesuchten Ausdrucks **mit akzeptabler Genauigkeit** automatisieren.
  - Die Genauigkeit wird **begünstigt** durch KWIC-Snippets mit einem **Kontext mittleren Umfangs** (ca. 30 Wörter); die Menge der Datensätze in einem Bereich zwischen 500 und 3000 zeigte keine Auswirkung auf das Ergebnis.
  - Aus linguistischer Sicht überraschend: Die evaluierten Verfahren erzielten mit einer **Bags-of-Words-Repräsentation** der Daten im Schnitt bessere Ergebnisse als mit einer Repräsentation, die weitere sprachliche Merkmale wie Wortartenzuordnungen oder Syntax berücksichtigt.
- Die Einbettung in die Umgebung RapidMiner ermöglicht auch die **unmittelbare aufgabenbezogene Annotation, Evaluation und Visualisierung der Ergebnisse**.
- Dadurch ist die Voraussetzung geschaffen, dass sich Korpus-Nutzer **schneller auf die eigentlich interessanten linguistischen Fragestellungen konzentrieren** können.

Literatur (Auswahl)

Bartz, T., Pölitz, C., Morik, K., Storrer, A. (2014). Using Data Mining and the CLARIN Infrastructure to Extend Corpus-based Linguistic Research. In: Jan Odijk (Ed): *Selected Papers from the CLARIN 2014 Conference*, Linköping: Linköping University Electronic Press, 1–13.

Bartz, T., Pölitz, C. and Radtke, N. (2013). *Automatische Klassifikation von Stützverbgefügen mithilfe von Data-Mining*. Technischer Bericht, Technische Universität Dortmund.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of Machine Learning Research,* 3: pp. 993–1022.

Goldstone, A. (o.J.): dfr-browser. Take a MALLET to disciplinary history.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines*. Dissertation, Dordrecht: Kluwer.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. and Euler, T. (2006): YALE: Rapid Prototyping for Complex Data Mining Tasks. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*. Philadelphia, USA.

Herzlichen Dank