
Beteiligte Projekte

- Ausbau und Pflege der Korpora geschriebener Sprache
- Methoden der Korpusanalyse und -erschließung

Beteiligte Mitarbeiter

- Dipl.-Ing. Cyril Belica
- Dr. Harald Längen
- Dr. Marc Kupietz
- Dipl.-Inf. Rainer Perkuhn

Kontakt

Dr. Harald Längen
Programmbereich Korpuslinguistik
Institut für Deutsche Sprache
Postfach 10 16 21
D-68016 Mannheim

Telefon: 0621/1581-0
Fax: 0621/1581-200
E-Mail: korpuslinguistik@ids-mannheim.de



 INSTITUT FÜR
DEUTSCHE SPRACHE

R 5, 6–13
D–68161 Mannheim
www.ids-mannheim.de

Mitglied der

Leibniz-Gemeinschaft

Das Institut für Deutsche Sprache (IDS) ist die zentrale Einrichtung zur Erforschung und Dokumentation der deutschen Sprache in ihrem Gebrauch und in ihrer neueren Geschichte. Es gehört zu den 86 außeruniversitären Forschungseinrichtungen und Serviceeinrichtungen für die Forschung der Leibniz-Gemeinschaft.

 INSTITUT FÜR
DEUTSCHE SPRACHE

Programmbereich Korpuslinguistik

MDCA

*Methodik
multidimensionaler
Korpusanalysen*

Viele sprachliche Phänomene – von einzelnen Wörtern bis hin zu komplexen syntaktischen Strukturen – treten nicht mit einer gleichmäßigen Verteilung auf, sondern bevorzugt in bestimmten situativen Kontexten (i.w.S.). Solche distributionellen Präferenzen sind in vielen Fällen eine intrinsische Eigenschaft des Phänomens selbst (und nicht etwa der Datengrundlage) und daher Anlass für vertiefende linguistische Untersuchungen.

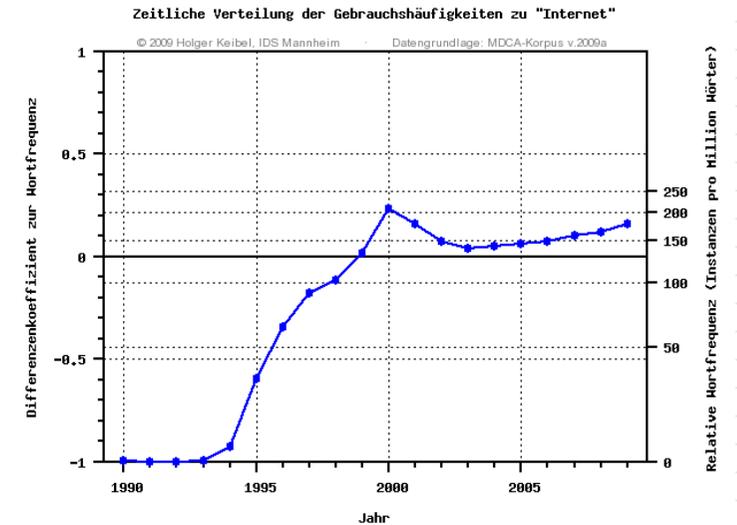
In diesem Forschungsschwerpunkt werden Verfahren erforscht, mit deren Hilfe überprüft werden kann, ob die Gebrauchshäufigkeit eines untersuchten Sprachphänomens entlang verschiedener situativer Dimensionen (wie Zeit, Ort, Genre, Thema oder Stil) eine auffällige Verteilung aufweist, die für eine gegebene linguistische Fragestellung relevant sein könnte. Die laufenden Forschungsarbeiten konzentrieren sich auf die Dimensionen *Zeit* und *Thema* sowie insbesondere auf deren Interaktion.

Relevante Forschungsaspekte

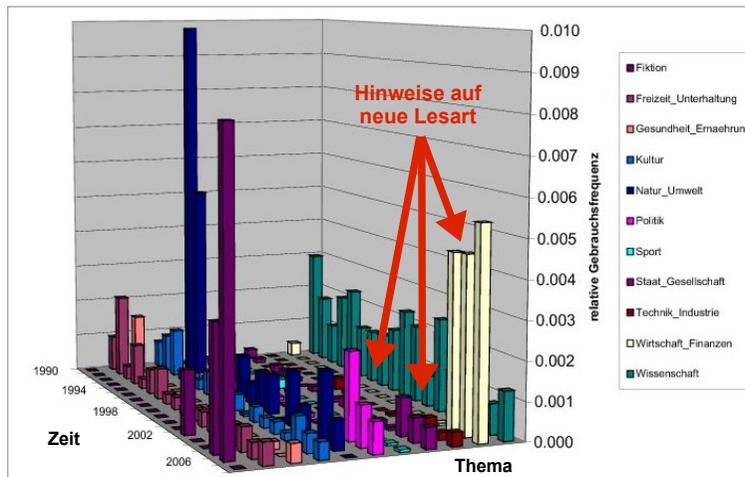
- Typologie möglicher Dimensionen: linear geordnet vs. hierarchisch vs. unstrukturiert
- universelle und dimensionsspezifische Analyseverfahren
- ein- und mehrdimensionale Analysen
- Kontrolle von anderen für die Vorkommensverteilung relevanten Einflussfaktoren (z.B. Textlängeneffekte, Sättigungseffekte)
- Exploration und Evaluation in konkreten Analyseszenarien (z.B. halbautomatische Detektion von Neologismen)

Stand: r1943 / r1501 - 2012-04

manuelle Auswahl	ID	Wort	Freq.	Eigennamen-Schätzer	Anteil vorherrsch. Region	vorherrsch. Region	LLR 2000er vs. 1990er	M-Score	spätestes Aufkomme
x	4744	Energieausweis	1090	3.72%	0.57798	DE-SW	1020456	331.35	20
x	4745	Vorratsdatenspeicherung	833	9.18%	0.38295	DE-O	791310	507.05	20
	4746	jön	566	11.11%	0.71731	DE-N	537673	344.72	20
x	4747	Mehrgenerationenhaus	545	35.88%	0.62448	DE-SW	524122	167.87	20
	4748	InBev	453						20
	4749	Welterbestatus	375						20
	4750	Chorverbands	374						20
	4751	Nierzwicki	310						20
	4752	kulturnotizen	303						20
	4753	Mecom	294						20
x	4754	Kompetenzagentur	258						20
x	4755	Verbraucherportal	237						20
x	4756	Breitbandversorgung	237						20
x	4757	Bespitzelungsaffäre	219						20
x	4758	Kreativwirtschaft	214						20
x	4759	Stevia	204						20
x	4760	Umweltplakette	194						20
	4761	Ehrenamtskarte	178						20
	4762	Hypothekenfinanzierer	177						20
	4763	Ambrosie	174						20
x	4764	Energieausweise	170						20
x	4765	Waterboarding	168						20
	4766	Peilomat	160						20
	4767	Crafter	157	36.21%	0.72611	DE-N	149143	96.06	20
	4768	Zafón	156	48.28%	0.30128	DE-SW	148193	95.45	20
	4769	Jaissle	153	86.05%	0.68627	DE-SW	126010	30.80	20
x	4770	Molekularl							20
	4771	wikipedia							20
	4772	Welterbeti							20
x	4773	Raucherkn							20
	4774	Aveo							20
	4775	Prinovis							20
	4776	Pooths							20
	4777	Geschäftsl							20
	4778	Entdeckerl							20
	4779	stefanie							20
	4780	Maatsch							20
	4781	Culcay							20
	4782	Forschung							20
x	4783	Blogosphä							20
	4784	Vettels							20
x	4785	Internetzei							20
x	4786	Bürgermetr							20
	4787	Nadals							20
	4788	Dreifeldsp							20
	4789	Bürgerwerkstatt	111	6.90%	0.47 / 48	DE-SW	105445	68.09	20
x	4790	blog	110	48.15%	0.39091	DE-O	104495	67.49	20
	4791	Fünfparteiensystem	105	0.00%	0.55238	DE-O	64518	10.23	20



Zeit-Thema-Interaktion der Gebrauchshäufigkeiten zu „Heuschrecke“



	33	90.59	20
	33	85.72	20
	14	82.68	20
	76	26.75	20
	14	80.25	20
	14	79.04	20
	27	39.21	20
	27	39.21	20
	27	39.21	20
	14	75.39	20
	50	14.59	20
	34	74.78	20
	15	71.74	20
	78	23.31	20
	77	13.74	20
	26	34.65	20
	35	17.02	20
	37	10.94	20
	105445	68.09	20
	104495	67.49	20
	64518	10.23	20