Projekte

- Ausbau und Pflege der Korpora geschriebener Sprache
- Methoden der Korpusanalyse und -erschließung
- Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen

Mitarbeiter

- Dipl.-Ing. Cyril Belica
- Dr. Marc Kupietz
- Dr. Harald Lüngen
- Dipl.-Inf. Rainer Perkuhn
- Heike Stadler
- Dr. Andreas Witt

Kontakt

Marc Kupietz
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

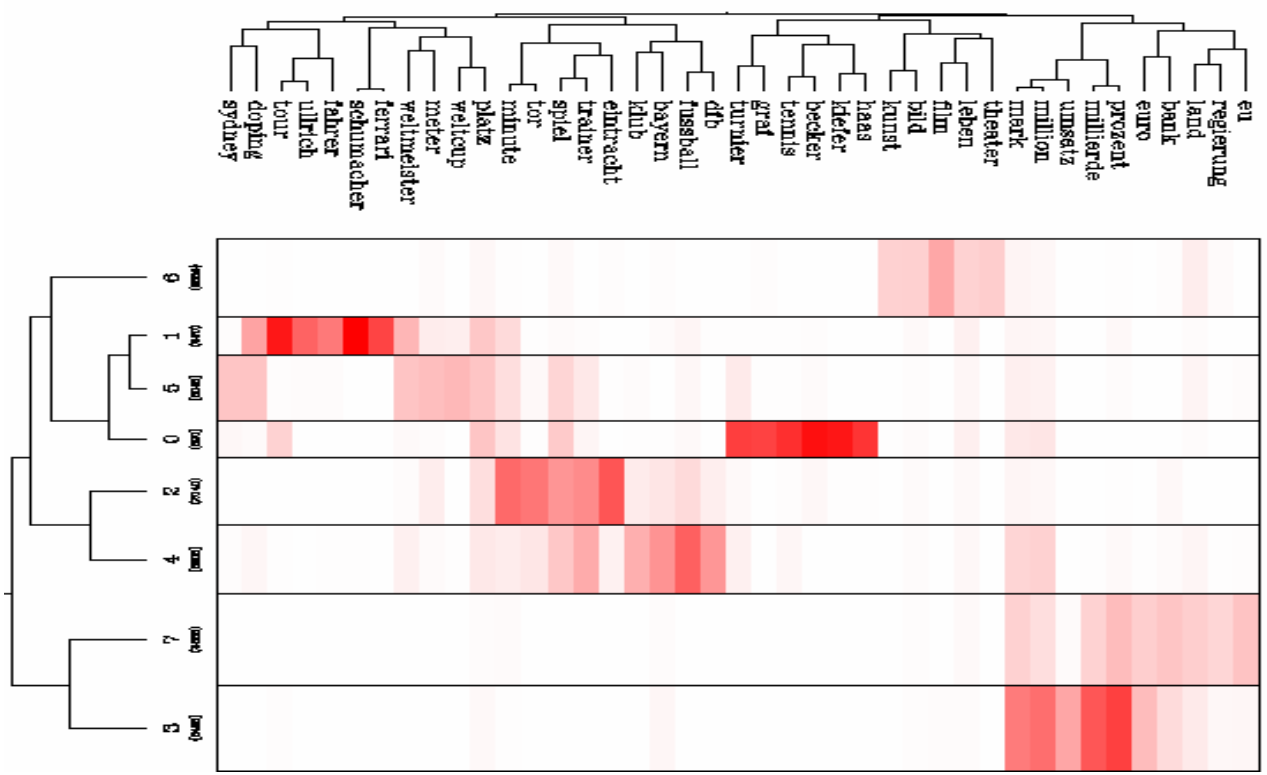
Programmbereich Korpuslinguistik

Thematische Erschließung von Korpora

Im Fokus stehen mathematisch-statistische Methoden zur thematischen Erschließung von Korpora, um sowohl themenspezifische virtuelle Subkorpora zusammenstellen als auch aufgrund der Analyse sachgebietsbezogener Häufigkeitsverteilungen z.B. Lesarten disambiguieren zu können.

Wichtige Komponenten

- eine formale und extern verankerte Themenontologie (Open Directory)
- eine mathematisch-statistische Methode zum automatischen Auffinden von Themen und Belegtexten (Dokumentclustering)
- eine mathematisch-statistische Methode zur thematischen Bestimmung eines (zuvor nicht klassifizierten) Textes (automatische Textklassifikation)
- weitere verwandte Methoden wie Spam-Checking und Schlüsselwortextraktion
- Anwendung des Verfahrens auf das gesamte DEUTSCHE REFERENZKORPUS
- Abschlussbericht (OPAL, 2005/1)
  http://www.ids-mannheim.de/pub/laufend/opal/privat/opal05-1.html



Hierarchische Cluster als Rubriken aus dem Bereich des Zeitungswesens (Feuilleton, Sport, Wirtschaft) erstellt mit Hilfe des Clusteringprogramms CLUTO

: aids_85	Gesundheit_Ernaehrung: Gesundheit
: affaere_barschel_87	Politik: Inland
: bse_01	Gesundheit_Ernaehrung: Gesundheit
: chemieunfall_sandoz_86	Technik_Industrie: Unfaelle
: eherecht_scheidungsrecht	Staat_Gesellschaft: Recht
: elfter_september_01	Politik: Ausland
: filmkritik_02	Kultur: Film
: fussball_uefa_94	Sport: Fussball
: polizeieinsaetze_02	Staat_Gesellschaft: Verbrechen

Zuordnung von Clustern zu Themen der Ontologie

Stand: 04/12