Automatische Sprachverarbeitung

Processing Extensive Text Data at the Leipzig Corpora Collection

Dirk Goldhahn
MLODE 2014, Leipzig

Natural Language Processing Group
Institute of Computer Science
University of Leipzig

UNIVERSITÄT LEIPZIG

Automatische Sprachverarbeitung

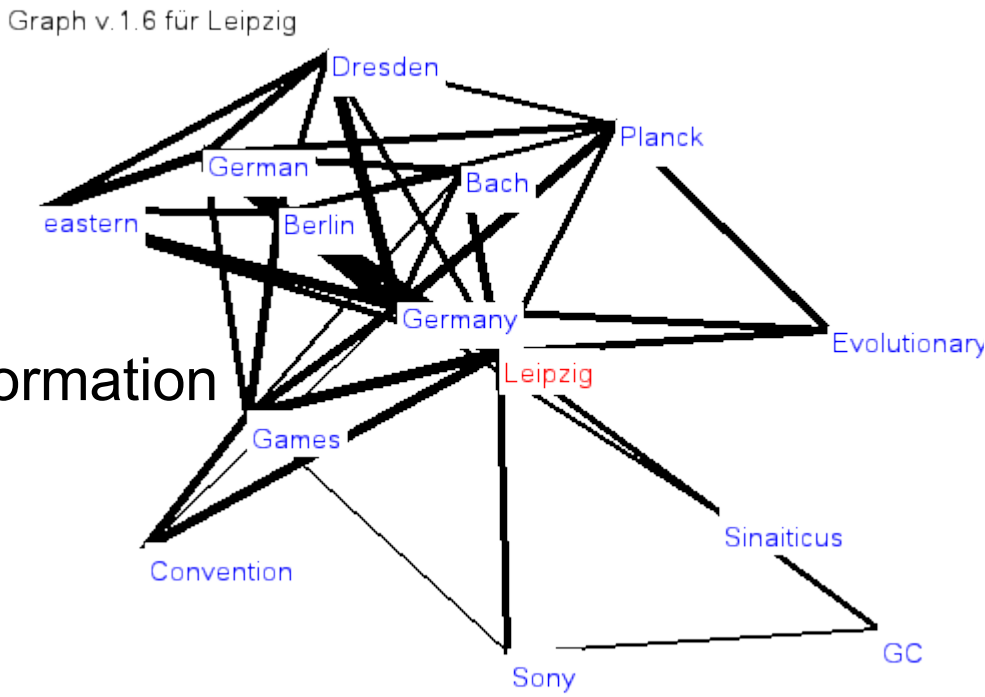
WORTSCHATZ
UNIVERSITÄT LEIPZIG

Leipzig Corpora Collection

- Project that offers
  - Corpus-based monolingual full form dictionaries
  - Started as project „Deutscher Wortschatz“
  - Accessible at http://corpora.uni-leipzig.de
    - Web interface
    - Web services
    - Integration into CLARIN
    - Export as rdf (Sebastian Hellmann)

Available Resources (Online)

- Corpora in >220 languages
  - Text material
  - Word frequencies
  - Word co-occurrences
  - Semantic maps (strongest word co-occurrences)
  - Topics
  - Subject areas (partially)
  - Morphological information (partially)
  - POS-tagged sentences (partially)
  - Further semantic and grammatical information



Example

Automatische Sprachverarbeitung

WORTSCHATZ UNIVERSITÄT LEIPZIG

Wort: Suche! ? Beachte Groß-/Kleinschreibung

Wort: Leipzig
Anzahl: 13571
Häufigkeitsklasse: 10 (d.h. *der* ist ca. 2^10 mal häufiger als das gesuchte Wort)
Beschreibung: Messestadt des Buchhandels
Stadt in Deutschland (über 250000 E)
Stadt in Sachsen
Stadt an der Weißen Elster
Universitätsstadt in Deutschland
Sachgebiet: Nachname
Militär
-
Grammatikangaben: Wortart: Eigenname
Links zu anderen Wörtern:

- Grundform: Leipzig
- Partitionsteil von: Leipzig
- ist ein(e) Stadt
- -er-Adj. / Einwohner zu Stadt: Leipziger, Leipziger, Leipziger
- Teilwort von: Universität Leipzig, RB Leipzig, Lok Leipzig, Zoo Leipzig, Sachsen Leipzig, Taxi nach Leipzig, Landgericht Leipzig, DHfK Leipzig, SOKO Leipzig, Arena Leipzig, SC L Handelshochschule Leipzig, BBV Leipzig, VfB Leipzig, FC Lok Leipzig, FC Sachsen Leipzig, Flughafen Leipzig, Roter Stern Leipzig, Stadtwerke Leipzig, Völkerschlacht bei Leipzig Landkreis Leipzig, Gewandhaus zu Leipzig, Bezirk Leipzig, Universitätsbibliothek Leipzig, JC Leipzig, Sparkasse Leipzig, Bachfest Leipzig, Gewandhausorchester Leipzig, Stadtwerk Theaterhochschule Leipzig, RasenBallsport Leipzig, Schauspiel Leipzig, Amtsgericht Leipzig, Arbeitsgericht Leipzig, SC Leipzig, Bach-Archiv Leipzig, Media City Leipzig, Verwaltu Radio Leipzig, Rasenball Leipzig, Leipzig Fernsehen, Staatsarchiv Leipzig, VC Leipzig, Hochschule für Grafik und Buchkunst Leipzig, VV Leipzig, Porsche Leipzig, Blue Lions Leipz Leipzig, Universitätsklinikum Leipzig, Herzzentrum Leipzig
- Form(en): Leipzig, Leipzigs
- Partition: Leipzig, Holzhausen, Lindenthal, Borsdorf, Borna, Ramsdorf, Markkleeberg, Groitzsch, Markranstädt, Schkeuditz, Espenhain, Böhlen, Frohburg, Taucha, Großpösna, Miltitz Heuersdorf, Bad Lausick, Engelsdorf, Pegau, Kitzen, Kitzscher, Kleinpösna, Panitzsch, Böhlitz-Ehrenberg, Mölkau, Wiederau, Neukieritzsch, Dölzig, Deutzen, Großstolpen, Lützsch

Beispiel(e):
Dieses Werk wurde am 22. April 1714 in **Leipzig** uraufgeführt. (Quelle: *www.elgger.ch, 2010-12-28*)
Die Sonntagsredaktion lockte den in seiner sorbischen Heimat berühmten Hochzeitsbitter mit einem Foto-Shooting nach **Leipzig** und überraschte ihn auf der Bühne mit einem ganz besond *www.mdr.de, 2011-01-11*)

Automatische Sprachverarbeitung

Example

Signifikante Kookkurrenzen für Leipzig:

in (6748.31), Dresden (4588.59), Bundesverwaltungsgericht (2951.17), RB (2183.9), HC (1692.71), Sachsen (1627.11), Leipziger (1592.44), Berlin (1467.9), Chemnitz (1232.57), Universität
Convention (990.57), Magdeburg (924.9), Halle (915.54), Erfurt (896.33), nach (814.68), Hamburg (793.3), München (779.32), Convention (744.11), AMI (708.29), Lok (699.62), studierte
(601.29), Jena (588.93), Nürnberg (572.19), Gewandhaus (568.16), 1989 (555.12), und (550.4), Handelshochschule (549.2), Buchmesse (538.06), Messe (512.71), Frankfurt (509.98), Litera
Games (486.69), DDR (473.74), Fußball-Einheit (471.21), DOK (463.7), Montagsdemonstrationen (462.63), SC DHfK (459.25), Köln (454.08), MDR (449.48), Anthropologie (447.1), Strom

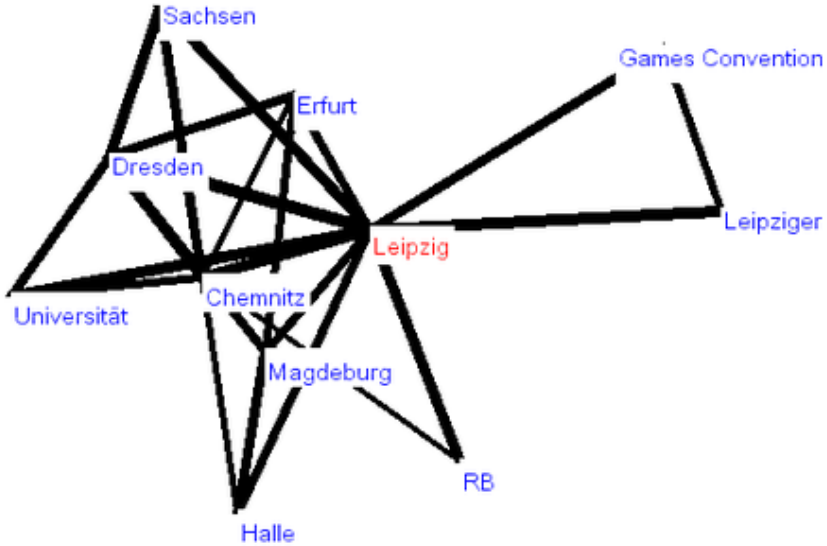
Signifikante linke Nachbarn von Leipzig:

in (23086.5), RB (2978.51), HC (2775.99), nach (2298.9), Universität (2251.3), aus (2025.66), In (1419.26), Lok (1020.48), DHfK (787.27), Zoo (682.2), DOK (675.42), SOKO (637.63), H
(611.92), an der Universität (606.66), Sachsen (540.96), FC Lok (486.02), BBV (477.5), Uni (458.95), FC Sachsen (457.66), Soko (444.89), von (402.31), Landgericht (389.21), Literaturinst
Oper (290.02), bei (285.84), Arena (276.32), Wasserwerke (216.19), Museum der bildenden Künste (207.08), Lokomotive (205.27), Staatsanwaltschaft (192.61), Messestadt (191.65), und (1
(178.39), Musikstadt (157.12), LVB (156.93), an der Uni (154.01), Richtung (148.91), Karl-Marx-Universität (148.39), Bachfest (147.76), JC (141.82), Gewandhausorchester (141.59), Stern
(133.02), Stadtwerke (130.04), Raum (129.73), Künste (128.22), in der Nähe von (118.62), Theaterhochschule (110.46), Calmus Ensemble (106.76), Standort (103.66), Panometer (102.63),
(99.92), Flughafen (99.74), Die Stadt (98.85), Messe (96.37), Marktforschung (95.11), euro-scene (93.34), Werk (92.78), Literaturinstituts (90.06), Bach-Archiv (88.76), Congress Center (8
Bundesverwaltungsgericht (85.23), Consort (83.76), DhfK (83.14), Großraum (77.67), Gewandhausorchesters (77.65), Rasenballsport (75.5), mit Sitz in (74.35), Buchstadt (73.58)

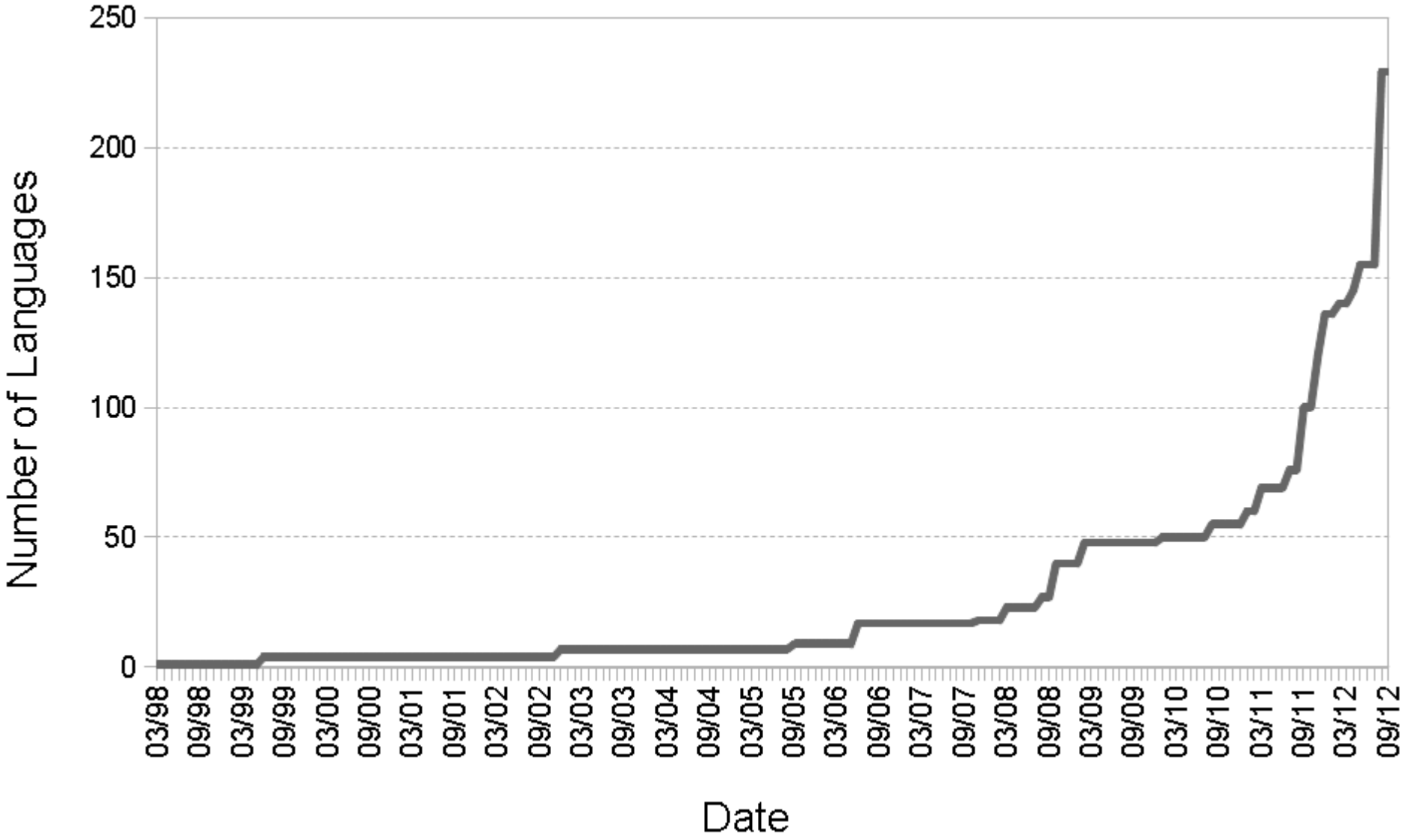
Signifikante rechte Nachbarn von Leipzig:

((2092.07), und (982.57), . (972.77), - (972.45), , (522.13), geboren (391.25), hat (356.58), liest (294.48), Eagles (264.7), lob (215.3), – (195.84),) (184.74), Tourist Service (168.1), ist (1
(118.42), uraufgeführt (104.1), noch zu retten (93.55), GmbH (91.9), Journalistik (87.22), —Sächsisch (83.14), mitteilte (83), sowie (81.36), apn (77.14), vorgestellt (75.42), Tourist (75.31)
Serie bauen (69.68), entschieden (69.22), war (64.31), haben (62.89), zu (61.51), gebaut (59.64), Information (58.94), statt (57.52), KWL (57.35), gekommen (55.67), Germanistik (51.68),
(47.08), mit (44.23), wird (44.1), gezeigt (43.43), nach (43), 1813 (42.5), 20 Jahre (42.01), in die Welt (41.73), kam (40.29), e.V (38), gebracht (37.81), am (36.85), aufgewachsene (35.32),
(33.61), umgeleitet (33.47), stattfindet (32.48), stammenden (32.12), rüstet (30.69), bleiben (30.41), stattfinden (30.22), oder sonstwo (29.91), Physik (29.04), angereist (28.59), / (28), läutet
(26.65), teil (25.37), gab (25.12), Theologie (24.81), widmet (24.46), lernte (24.33)

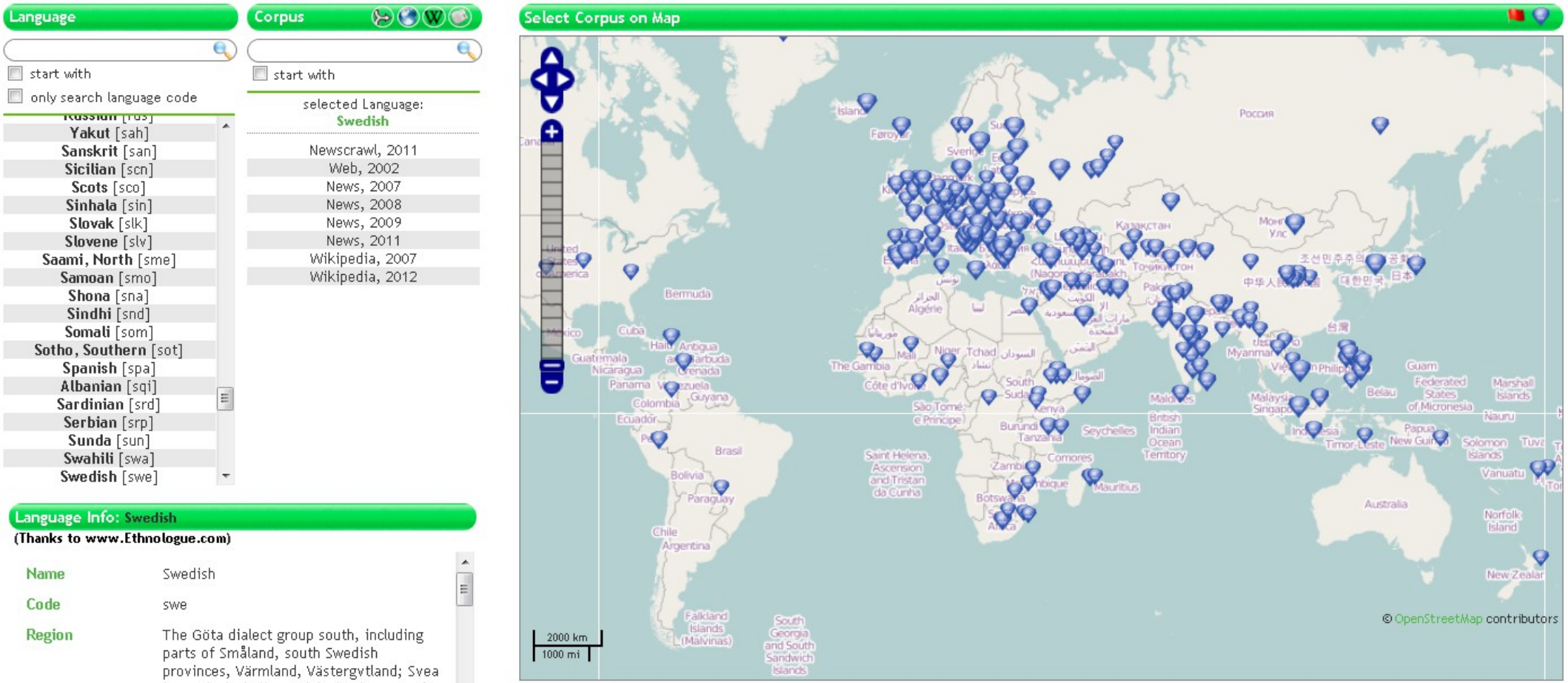
Graph v.1.6 für Leipzig



Number of Languages over Time

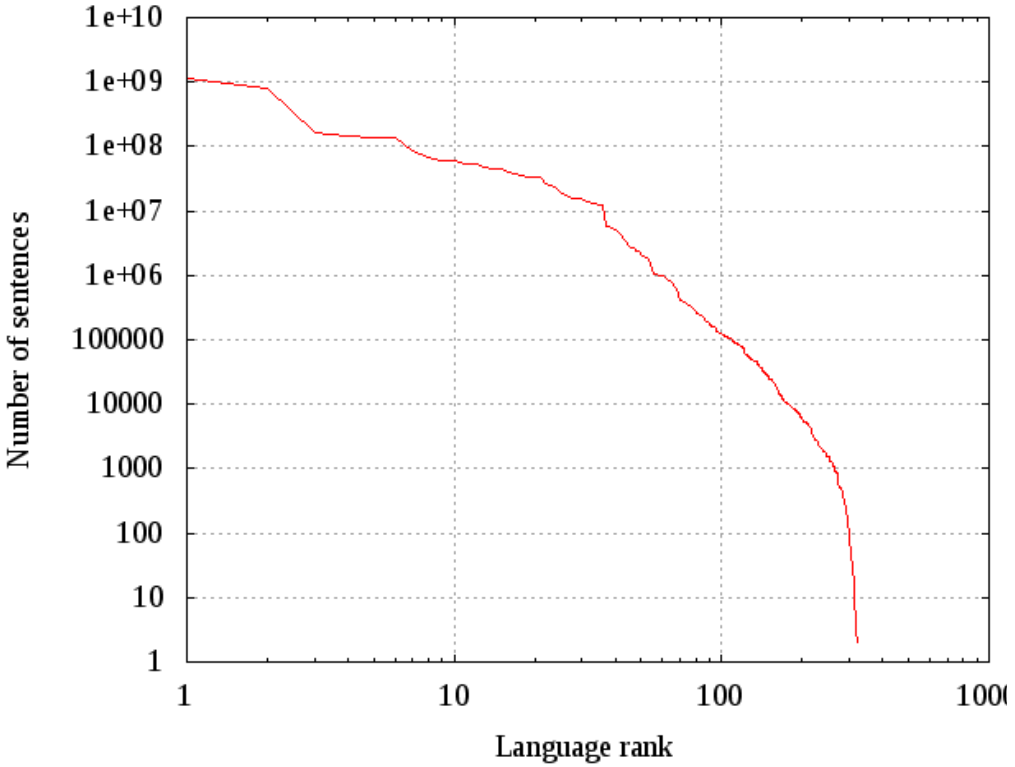


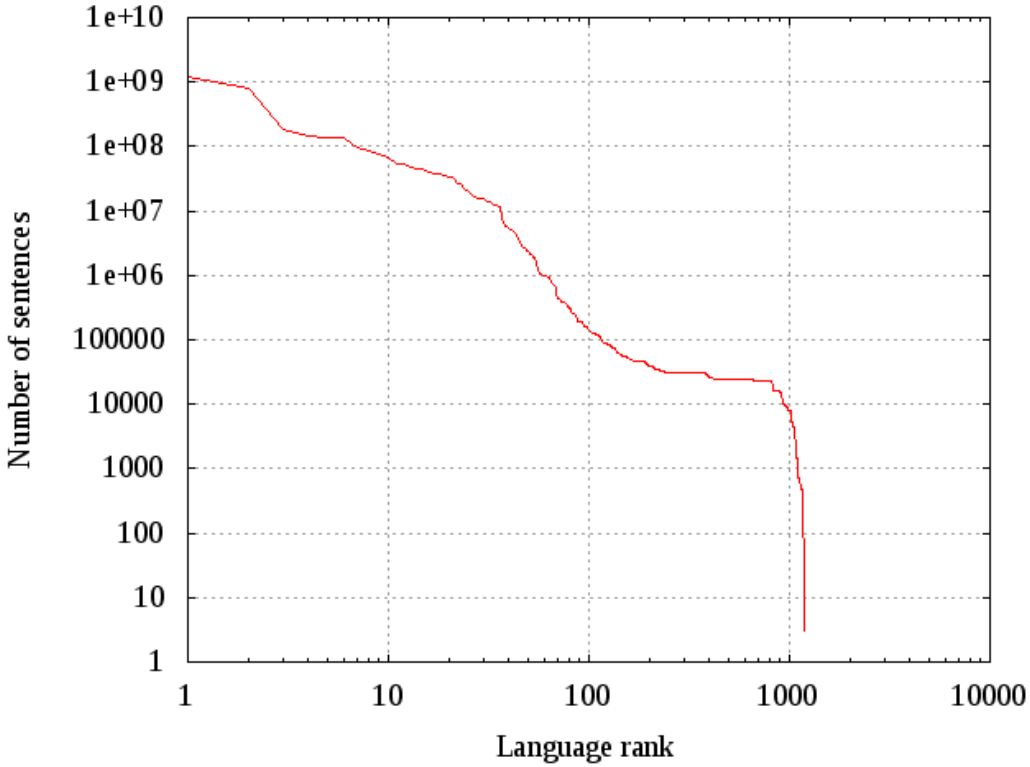
World Map Interface



Data Stock

- Number of languages: >300, >1000 including religious texts
- Number of sentences (overall): ~4 billion
- Biggest language: English (1.1 billion)





Text Acquisition Methods

- Generic web crawler (Heritrix)
  - Whole TLDs
  - News (abyzNewsLinks)
- RSS feeds (daily)

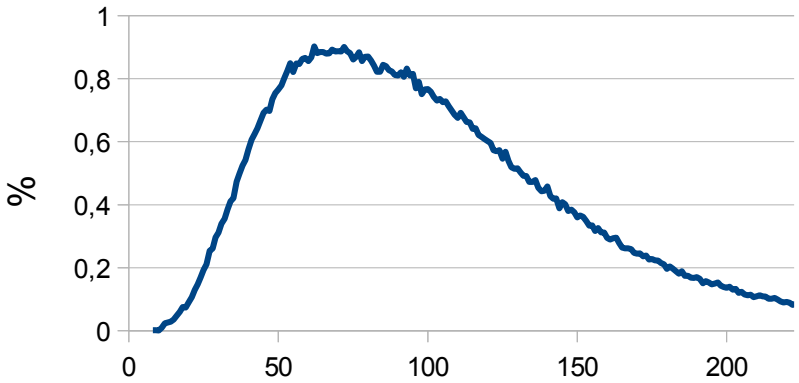
  → http://wortschatz.uni-leipzig.de/wort-des-tages/
- Distributed text crawling (FindLinks)
- Bootstrapping web corpora
- Text dumps (Wikipedia, Watchtower)
- Parallel corpora (Bible)

Corpus Creation

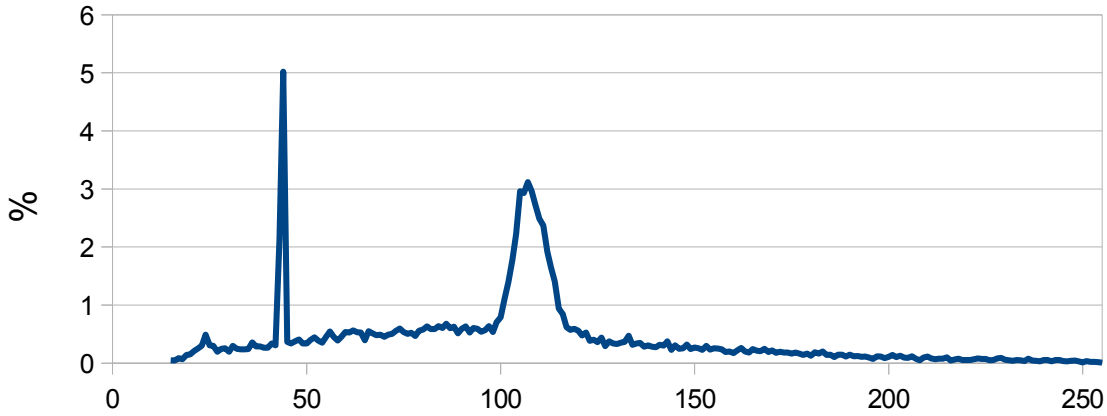
- Processing steps:
  - Stripping
  - Language Separation (about 500 languages)
  - Sentence Segmentation
  - Cleaning
  - Sentence Scrambling (copyright reasons)
  - Tokenization
  - Database Creation
  - Statistical Evaluation

Statistical Evaluation

- Enrichment of corpora with statistical annotations
  - Word frequencies
  - Co-occurrence frequencies and significance (based on left or right neighbours or whole sentences)
  - Topics, POS
- Creation of further corpora statistics
  - >200 statistical properties based on letter, word, sentence, sources level etc.
  - e.g.: For automatic quality assurance using typical distributions



Hindi newspaper 2010



Sundanese Wikipedia 2007

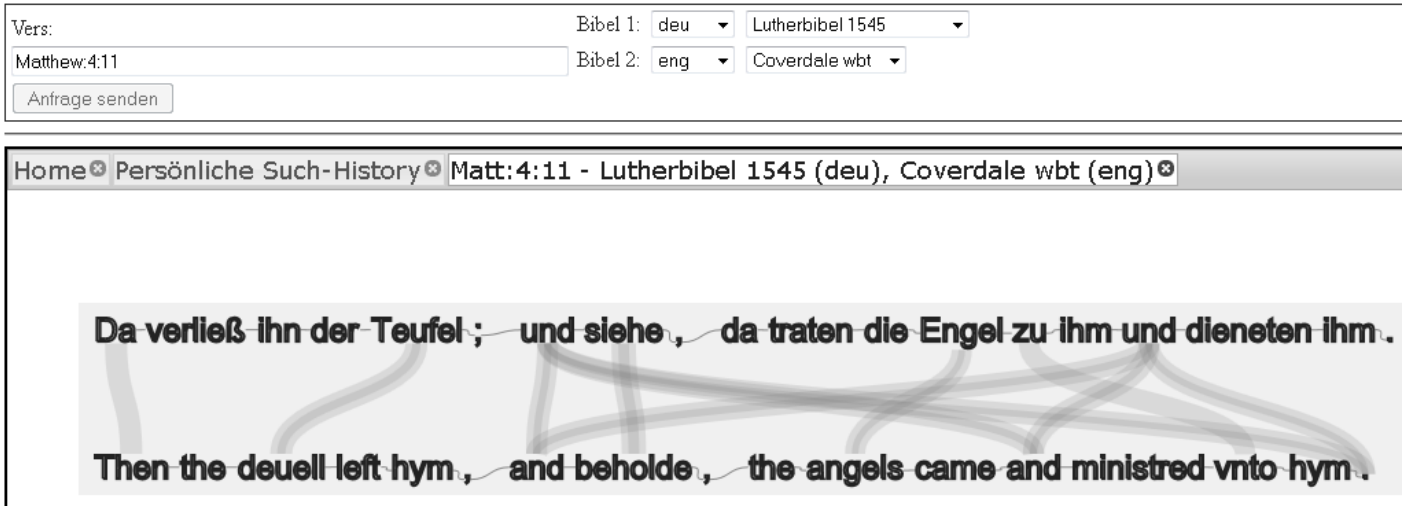
LCC as Linked Data

- LD starts with data and metadata
  - Data + metadata for hundreds of languages
  - All data processed and stored identically, including metadata
  - Still at the beginning of turning into LD

LCC as Linked Data

- Monolingual Metadata:
  - Frequencies, word co-occurrences, NER, POS-tags, topics
  - Integration of external sources (DBPedia, FreeBase, Wordnet)
    - Still problems to solve (linking, disambiguation, ...)

  → New Web Interface
- Multilingual Metadata:
  - Crosslingual word co-occurrences (parallel text)



  - External sources (DBPedia...)

- Data available via:
  - Web interface (http://corpora.informatik.uni-leipzig.de)
  - Download (http://corpora.informatik.uni-leipzig.de/download.html)
  - Webservices (http://wortschatz.uni-leipzig.de/Webservices/)

WORTSCHATZ UNIVERSITÄT LEIPZIG

Thank you!