

Textdatenbanken

Sommersemester 2020

Uwe Quasthoff

Universität Leipzig
Institut für Informatik

quasthoff@informatik.uni-leipzig.de

Organisatorisches

Die Präsenz-Vorlesungen beginnen am 8.5.2020 mit einer Hörsaalübung zum Inhalt der ersten 4 Veranstaltungen. Bis dahin werden wöchentlich Folien und weitere Materialien zum Selbststudium angeboten.

10.04.2020 Vorlesung 1: Material für E-Learning

17.04.2020 Vorlesung 2: Material für E-Learning

24.04.2020 Vorlesung 3: Material für E-Learning

01.05.2020 Vorlesung 4: Material für E-Learning

08.05.2020 Hörsaalübung: Fragen und Antworten zu Vorlesungen 1 bis 4

- **Vorlesung ab 8.5.:** wöchentlich freitags 9:15 Uhr in S 3-14
- **Übungen ab 8.5.:** freitags 11:15 Uhr in S 3-14
gehalten von: Felix Helfer <helfer@informatik.uni-leipzig.de>)
- **Vorlesungsfolien** und weiteres Material:
<http://asv.informatik.uni-leipzig.de/de/courses/281>

Wünsche beim Arbeiten mit viel Text

- Sehr viel Text, Größenordnung 10^9 laufende Wörter, 10 GB. Gern mehr.
- Ständige Erweiterung der Ressourcen
- Auswahl nach verschiedenen Kriterien, z.B.
 - Sprache
 - Sachgebiet
 - Entstehungszeit
- Gute Qualität (d.h. wenig Datenmüll, siehe später).
- Schneller Zugriff (kein sequenzielles Suchen auf der Datei)
- Intelligente Suchmöglichkeiten (d.h. wie Suchmaschine oder besser)
- Vorverarbeitung, Bereitstellung von vorberechneten Daten
- Einheitliches Format der Texte
- Viele nützliche Tools, die auf diesem Format arbeiten

Nutzerkreis und Anwendungsmöglichkeiten

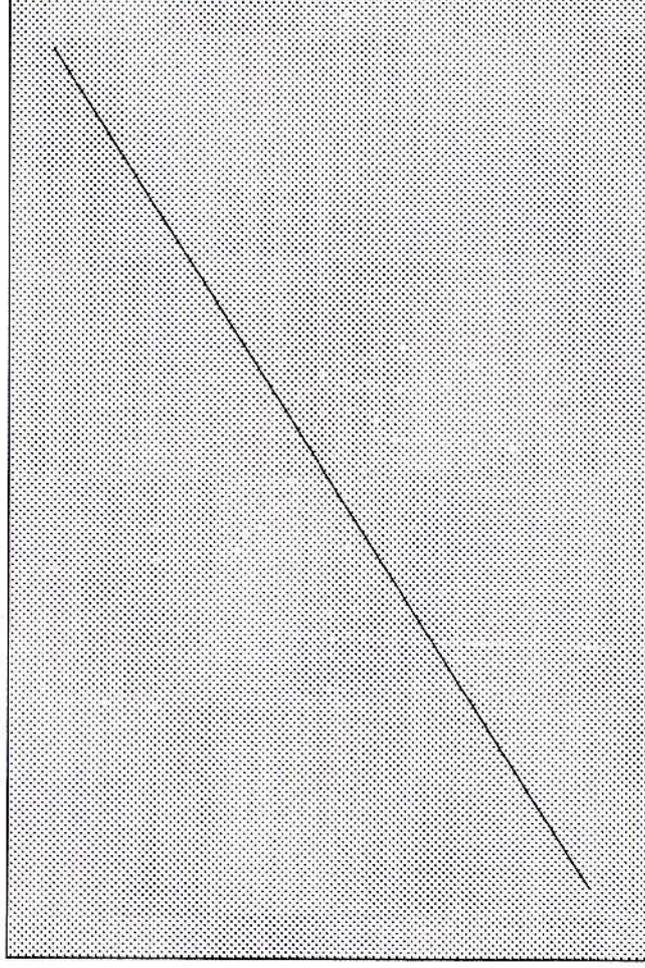
Der Interessentenkreis ist weit gefächert und die Anwendungsmöglichkeiten sind vielfältig. Beispielsweise:

- Unterstützung bei der korpusbasierten Erstellung klassischer Printwörterbücher
- Arbeitsmaterial für Linguisten mit Interesse an der entsprechenden Sprache
- Erforschung von Fragestellungen der Linguistik mit korpusbasierten Methoden
- Sprachvergleich mittels verschiedener Korpora gleicher Größe
- die Berechnung statistischer Sprachmodelle, wie sie beispielsweise bei der Erkennung gesprochener Sprache benötigt werden
- Unterstützung von Suchmaschinen durch Auswahl bedeutungsähnlicher Wörter
- Wissensmanagement: Erkennen inhaltlich wichtiger Begriffe durch Auswertung statistischer Auffälligkeiten
- Wortauswahl für psycholinguistische Experimente

Corpus size since the 1960s

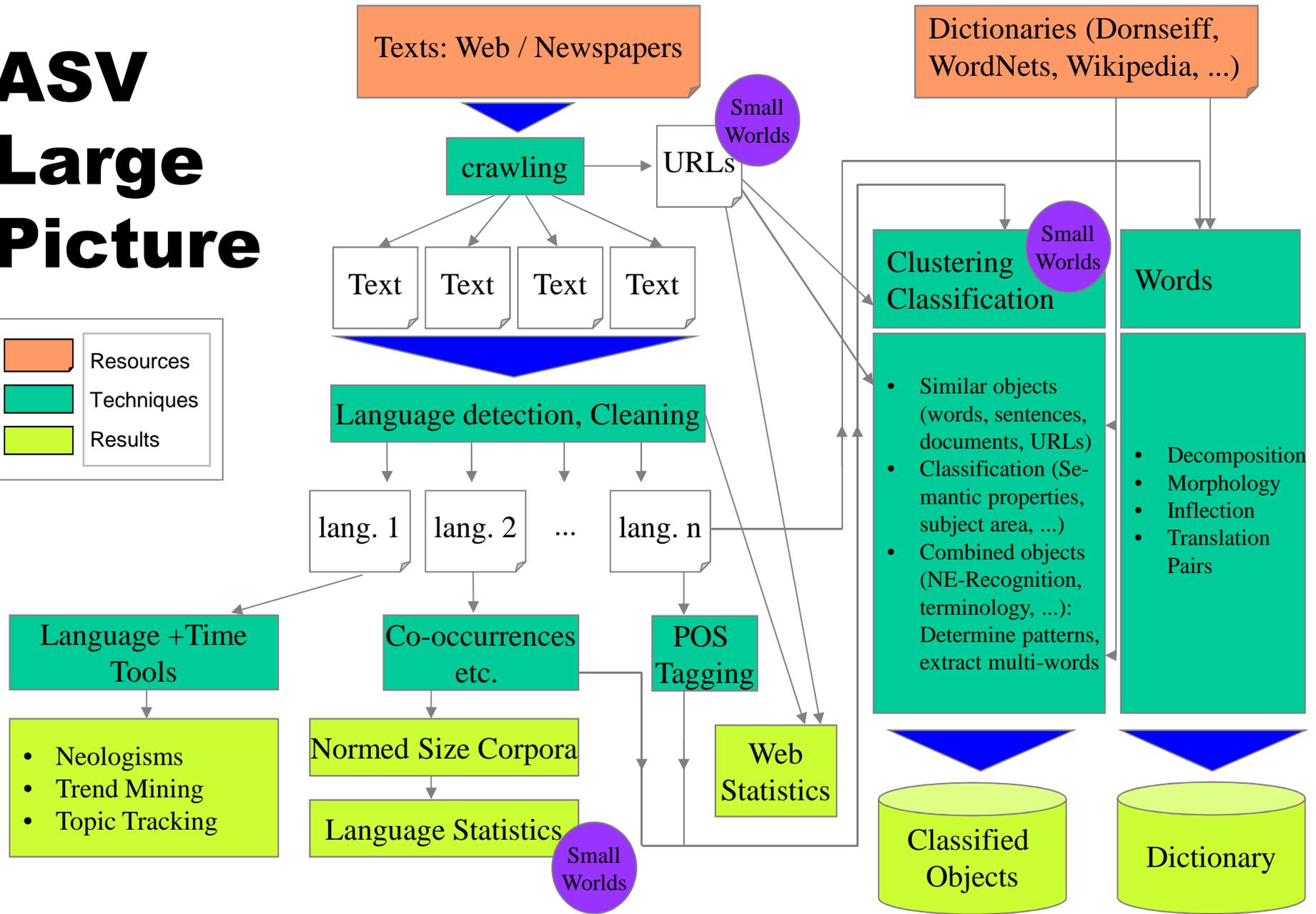
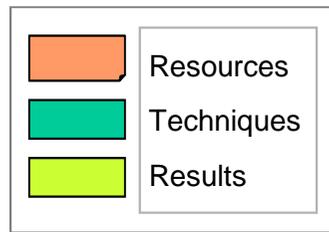
Size of
Corpora
(in
words)

10^9
 10^8
 10^7
 10^6



1960s 1970s 1980s 1990s 2000s 2008
Brown/LOB COBUILD BNC OEC no limits

ASV Large Picture



Datenbeschaffung: Fremde oder eigene Ressourcen?

1. Eigene Beschaffung

Web-Crawling und Weiterverarbeitung (siehe später)

2. Andere Korpusprojekte in Deutschland

IDS, Mannheim

Ausschließlich für die deutsche Sprache. Datengrundlage bildet ein Korpus mit geschätzten knapp 100 Millionen Sätzen. Erhältlich sind drei kleine Teile von maximal $\frac{1}{4}$ Million Sätzen für ca. 2000 €.

DWDS (BBAW, Potsdam)

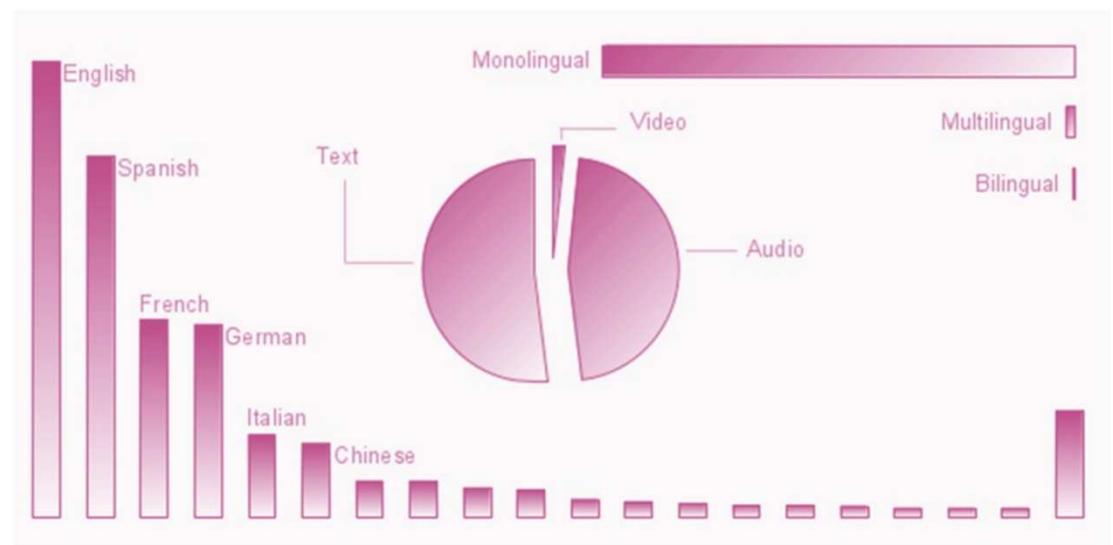
Ausschließlich für die deutsche Sprache. Datengrundlage bildet ein Korpus mit geschätzten 70 Millionen Sätzen. Diese Daten sind nicht erhältlich.

Internationale Korpusanbieter I

ELDA: Evaluations and Language resources Distribution Agency

<http://www.elda.org/>

- Mitgliedschaft für jährlich €750 - €5.000
- Preise: Unterschiedlich für Mitglieder / Nichtmitglieder und nichtkommerzielle / kommerzielle Nutzung. Die Preise sind häufig recht hoch.
- Daten: Unterschiedliche Formate, teilweise mit grammatischen oder anderen Informationen angereichert.
- Sprachen: Größere Korpora (>1 Mill. Sätze) in Englisch, Französisch, Deutsch, Italienisch, Spanisch, Arabisch



Internationale Korpusanbieter II

LDC: Linguistic Data Consortium

<http://www ldc upenn edu/>

- Mitgliedschaft für jährlich \$2.400-\$27.500 je nach Status.
- Preise: Für Mitglieder sind neue Korpora (in vielen Fällen) kostenlos. Preise für Nichtmitglieder. Beispiel für großes Textkorpus:
<http://www ldc upenn edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05>
\$3000 für ca. 100 Millionen Sätze in Form ganzer Dokumente in XML (mit Quelle), keinerlei zusätzliche Daten.
- Daten: Unterschiedliche Formate, teilweise mit grammatischen oder anderen Informationen angereichert.
- Sprachen: Größere Korpora (>1 Mill. Sätze) in Englisch, Französisch, Deutsch, Italienisch, Spanisch, Arabisch, Chinesisch, Japanisch, Koreanisch

Rohtext

Unstrukturierter **ASCII**-Text wird als Grundlage für die statistischen und clusterbasierten Verfahren des *Text Mining* benötigt.

Quellen:

- Meist: HTML-Text aus dem Web (unproblematisch wegen einheitlicher und bekannter Struktur).
- Seltener: pdf-Dokumente, XML aus Satzsystemen (problematisch, da pdf-Konverter unzuverlässig und XML-Struktur uneinheitlich).

Korrekte Konvertierung

```
<source><location>http://www.rohschnitt.de/memory/index.htm</location><date>2011-00-15</date>  
<user>ich</user><original_encoding>iso-8859-1</original_encoding><language>deu</language>  
</source>
```

Rohschnitt - Spiele für alle

Hinter jedem Button ("NR. 01" - "NR. 16" bzw. "NR. 01" - "Nr. 24") verbirgt sich ein Begriff, der bei Betätigung des Buttons sichtbar wird. Insgesamt kommt jeder Begriff zweimal vor, sodaß sich hinter den 16 (24) Buttons die Namen von 8 (12) Begriffen verbergen.

Der Zweck des Spiels ist es, jeweils den gleichen Begriff direkt hintereinander anzuwählen. In diesem Falle gilt der entsprechende Begriff als "gelöst" und die Beschriftung des Buttons ändert sich dementsprechend.

Seitenanfang

Bei Navigieren zwischen verschiedenen Seiten (z.B. zurück vom gewählten Begriff zum Spiel) immer die Buttons benutzen und nicht etwa die Vor- oder Zurückfunktion des Browsers ! Sonst funktioniert das Spiel nicht !

Seitenanfang

Das Spiel ist teilweise, in der Variante mit 16 Feldern komplett, mit Accesskeys bedienbar, sofern der Browser diese interpretiert (Internet Explorer ab 5.0 und Netscape ab 6.0).

Eigenschaften des Textformates

- Alles in utf-8
- XML-Header mit URL, Datum sowie User, Original-Encoding und ermittelter Sprache
- Datum ist Crawling-Datum, bei RSS-Newsfeeds auch (näherungsweise) Erscheinungsdatum.

Nach dem Header folgt der Text:

- Doppelter Zeilenumbruch ist Absatzende (und damit immer Satzende)
- Einfacher Zeilenumbruch kann konfiguriert werden als Absatzende ja/nein. Diese globale Entscheidung für eine große Menge von Texten versagt immer bei einigen Ausnahmen.

Schlechtere Quelle: Harte Umbrüche

<source><location><http://www.dein-garten.at/gruen/></location><date>2011-00-15</date><user>ich</user><original_encoding>iso-8859-1</original_encoding><language>deu</language> </source>

Als Gartengestaltungsbetrieb ist eine unserer Hauptaufgaben die Begrünung der gewünschten Objekte. Die richtige Pflanzenauswahl setzt beim zu begrünenden Objekt neue architektonische Aspekte. Entscheidend ist aber auch der pflanzenphysiologisch richtige Standort um auch nach vielen Jahren Freude an deren Anblick zu haben. Eine oft vernachlässigte Grundvoraussetzung ist der Boden, speziell bei Rasenflächen, Dachgärten und Trögen.

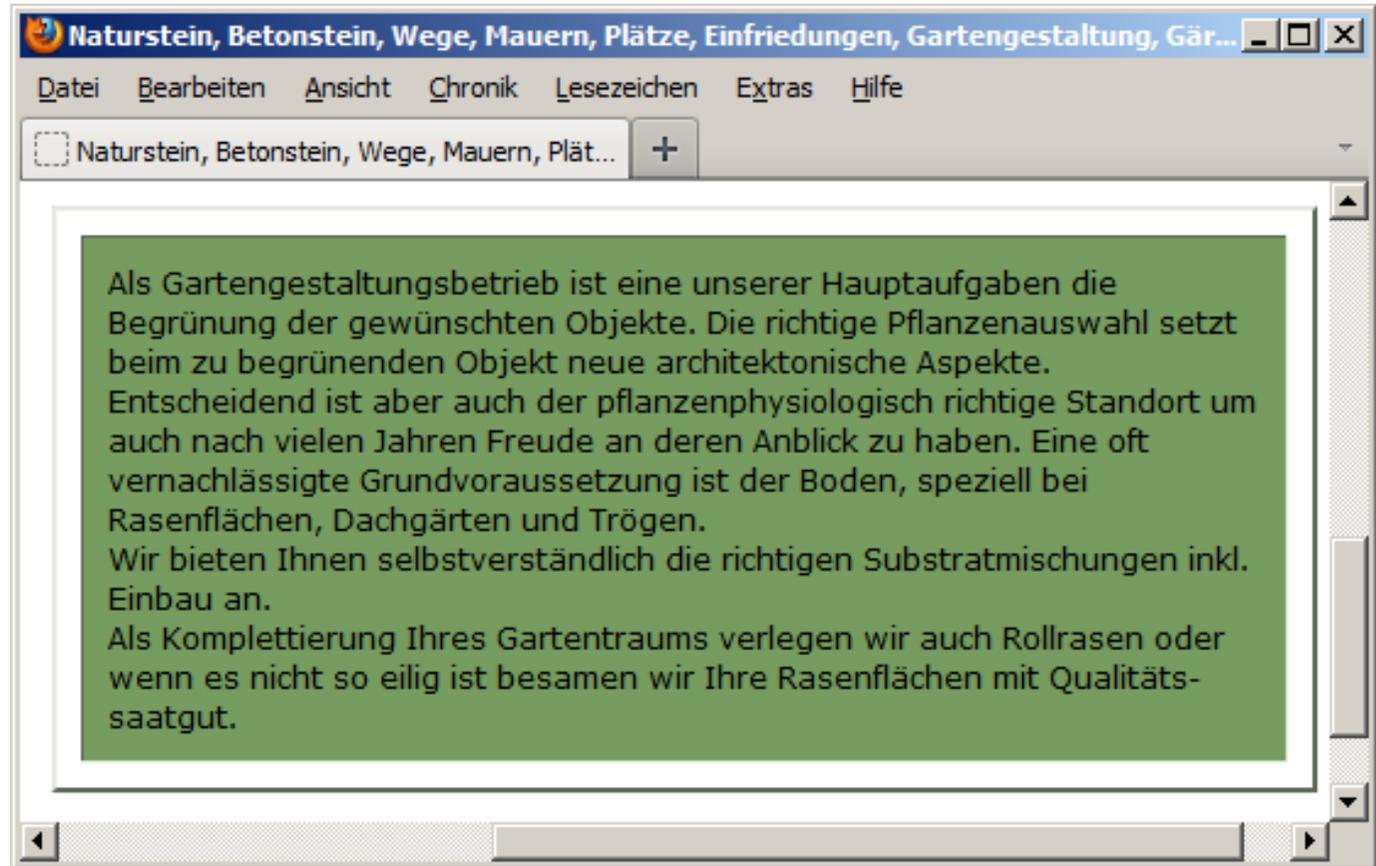
Wir bieten Ihnen selbstverständlich die richtigen Substratmischungen inkl. Einbau an.

Als Komplettierung Ihres Gartentraums verlegen wir auch Rollrasen oder wenn es nicht so eilig ist besamen

wir Ihre Rasenflächen mit Qualitäts-saatgut.
U. Quasthoff

Textdatenbanken

Das Original



HTML

...

Wir bieten Ihnen selbstverständlich die richtigen
Substratmischungen inkl. Einbau an.

Als Komplettierung Ihres Gartentraums verlegen wir
auch Rollrasen oder wenn es nicht so eilig ist besamen
wir Ihre Rasenflächen mit Qualitäts-saatgut.</td>

Verarbeitungsfehler (weiche Trennungen als hart gekennzeichnet)

```
<source><location>http://mainz-online.de/on/07/04/10/news/r/pumuckl.html?a</location>  
<date>2011-01-15</date> <user>ich</user><original_encoding>iso-8859-1  
</original_encoding><language>deu</language> </source>
```

Ber-**lin** - Bei der Finan-**zie**-**run**g der Kin-**der**krip-**pen**-**Pl**<E4>**ne** von Bun-**des**-**fami**-**lien**-**minis**-**terin** Ursula von der Leyen (CDU) steht der großen Koali-**tion** ein hef-**tiger** Streit ins Haus.

Mün-**chen** - Die Frage, ob der Kobold Pumuckl eine Freun-**din** haben oder gar hei-**raten** darf, beschäf-**tigt** dam 26. April das Land-**gericht** München I. Pumuckl-**Erfin**-**derin** Ellis Kaut (Foto) hat die Pumuckl-**Zeich**-**nerin** Barbara von Johnson ver-**klagt**. Diese unter-**stützt** einen Kin-**der**-**mal**-**wett**-**bewerb**, bei dem eine zeich-**neri**-**sche** Dar-**stel**-**lung** einer Freun-**din** des Pumuckl gesucht wird, wie das Gericht am Diens-**tag** mit-**teilte**. Dem Gewin-**ner** winkt die Teil-**nahme** an einer "Hoch-**zeit**" des Pumuck-**l**. Das will Kaut aber nicht erlau-**ben** und hat eine einst-**wei**-**lige** Ver-**fü**gung bean-**tragt**. "Der Pumuckl ist und bleibt ein Nach-**fahre** der Kla-**bau**-**ter**, also ein Geist-**wesen**. Grundsätz-**lich** haben Geist-**wesen** kein aus-**gepräg**-**tes** Geschlecht", begrün-**dete** die Autorin ihre Zivil-**klage**. Ein Heirat vom Pumuckl wider-**spre**-**che** seinem lite-**rari**-**schen** Cha-**rak**-**ter**.

Verarbeitungsfehler: Zeichensatzprobleme

<source> ... </source>

Hufenstuhl war bei der politischen Abteilung der Kriminalpolizei Wuppertal im Rang eines Kriminalkommissars beschäftigt. Er galt überdies als willfähriger und opportunistischer **NS-Roboter**, der als Leiter der Exekutivabteilung eine wütende Feindschaft zu seinem Vorgesetzten und Dienststellenleiter Kriminalrat Wilhelm Müller kultivierte.

Annotierter Text

Annotierter Text unter Verwendung einer standardisierten Menge von *Tags* ist für die Analyse syntaktischer Muster, die Klassifizierung von Wortformen nach syntaktischen Kategorien und für Sachgebietszuordnungen von Wortformen und Texten erforderlich.

Arten von Annotation:

- Dokumentenstruktur
- POS-Tags (Wortarten zu Wörtern)
- Semantische Annotation von Eigennamen

Text mit POS-Tags (TNT)

Input	Basic Output	Optional Extended Output
Der	ART	ART 1.000000e+00
Mandolinen-Club	NN *	NN 1.000000e+00 *
Falkenstein	NE *	NE 8.001280e-01 NN 1.998720e-01 *
und	KON	KON 1.000000e+00
der	ART	ART 1.000000e+00
Frauenchor	NN *	NN 9.828203e-01 NE 1.717975e-02 *
aus	APPR	APPR 1.000000e+00
dem	ART	ART 1.000000e+00
sächsischen	ADJA	ADJA 1.000000e+00
Königstein	NN	NN 7.762892e-01 NE 2.237108e-01
gestalten	VVINF	VVINF 1.000000e+00
die	ART	ART 9.796126e-01 PRELS 1.443545e-02 ...
Feier	NN	NN 1.000000e+00
gemeinsam	ADJD	ADJD 1.000000e+00
.	\$.	\$. 1.000000e+00

Annotation von Eigennamen

< TIMEX TYPE='DATE' > all of 1987 < /TIMEX >

< TIMEX TYPE='TIME' > 8:24 a.m. Chicago time < /TIMEX >

< NUMEX TYPE='MONEY' > several million New Pesos < /NUMEX >

more than < NUMEX TYPE='PERCENT' > 95% < /NUMEX >

in < ENAMEX TYPE='LOCATION' > North and South America < /ENAMEX >

the < ENAMEX TYPE='ORGANIZATION' > U.S. Fish and Wildlife Service < /ENAMEX >

the < ENAMEX TYPE='PERSON' > Clinton < /ENAMEX > government

< ENAMEX TYPE='ORGANISATION' > Microsoft < /ENAMEX >

chairman < ENAMEX TYPE='PERSON' > Bill Gates < /ENAMEX > said
yesterday

Normgrößenkorpora

Korpora

- Verfügbarmachen der Korpora (in verschiedenen Normgrößen) für Nutzer weltweit.
- Wir können mit Daten und Austauschformaten Standards setzen.
- Es besteht dringender Bedarf an solchen Korpora.

Größen

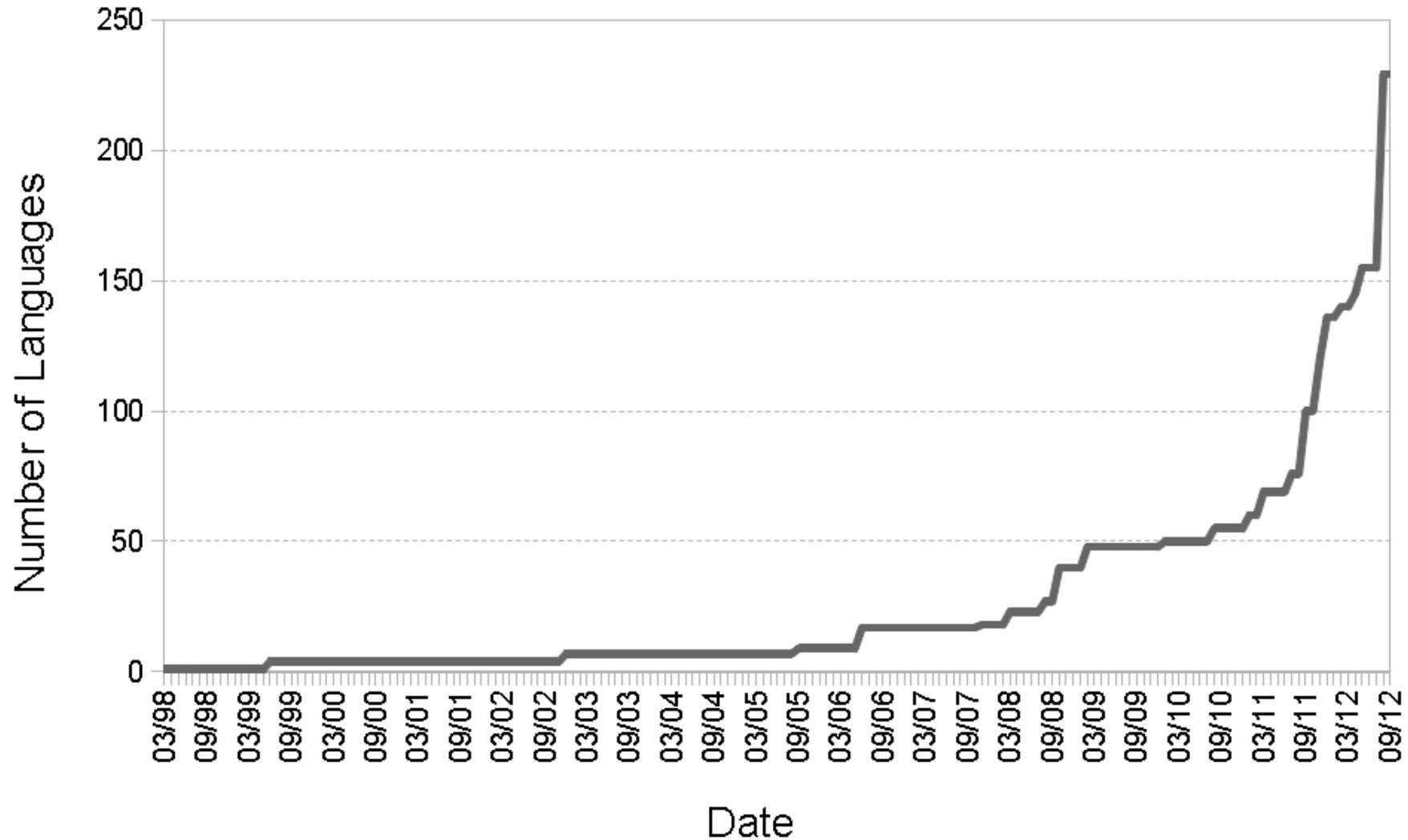
- Anzahl von Sätzen: 10.000, 30.000, 100.000, 300.000, 1, 3, 10, 30 Millionen
- Für jede Sprache bis zum jeweiligen Maximum
- „Reinigen“ von fremdsprachlichen und nicht wohlgeformten Sätzen.
- Sprachvergleich durch völlig gleiche Verarbeitung
- Bereitstellung von Kookkurrenzen zur Weiterverarbeitung

Normgrößenkorpora

Das geplante Vorhaben zeichnet sich aus

- durch die größere Anzahl von Sprachen und den großen Umfang pro Sprache
- durch das einheitliche Format der Daten für alle Sprachen und
- durch die Vergleichbarkeit der Korpora auf Grund der Normgrößen
- durch die zusätzliche Lieferung statistischer Daten. Kookkurrenzdaten werden nirgendwo angeboten und sind für viele Anwender nützlich.

Anzahl der verschiedenen Sprachen



Stand April 2012

Search in 143 Corpus-Based Monolingual Dictionaries

Newest Dictionaries

Word:

Active dictionary: English case sensitive search

Random words:

lights two-run obtained 2.5 ABC

Change Dictionary:

Acholi	Afrikaans	Albanian	Amharic	Arabic	Aragonese
Armenian	Asturian	Azerbaijani	Bashkir	Basque	Belarusian
Bengali	Bicolano	Bishnupriya	Bosnian	Bretonian	Bulgarian
Catalan	Cebuano	Chinese (simplified)	Chuvash	Corsican	Croatian
Czech	Danish	Dimli	Dutch	Egyptian Arabic	English
English (AU)	English (CA)	English (NZ)	English (UK)	Estonian	Faroese
Fijian	Finnish	French	Ganda	Georgian	German
German (CH)	Gilaki	Goan Konkani	Greek	Greenlandic	Gujarati
Haitian	Hausa	Hebrew	Hindi	Hindi, Fiji	Hungarian
Icelandic	Ido	Indonesian	Interlingua	Italian	Japanese
Javanese	Kannada	Kazakh	Khmer, Central	Kiswahili	Korean
Kurdish	Kyrgyz	Latin	Latvian	Lithuanian	Lushai
Luxemburgian	Macedonian	Malay	Malayalam	Maldivian	Maltese
Maori	Marathi	Min Nan Chinese	Mongolian Cyrillic	Nahuatl	Nepali
Newari	Norwegian (Bokmål)	Norwegian (Nynorsk)	Occitan	Ossetian	Pampanga
Panjabi	Papiamentu	Pashto	Pennsylvanian Dutch	Persian	Piemontese
Polish	Portuguese (Brazil)	Portuguese (Macao)	Portuguese (Portugal)	Romanian	Romansch
Russian	Rusyn	Sami	Samogitian	Sanskrit	Scots
Scottish Gaelic	Serbian	Sicilian	Sinhala	Slovak	Slovenian
Somali	Sorbian (Upper)	Spanish	Spanish (Mexico)	Sundanese	Swahili
Swedish	Tagalog	Tajik	Tamil	Tatar	Telugu
Thai	Turkish	Ukrainian	Urdu	Uyghur	Uzbek
Uzbek, Latin	Venetian	Vietnamese	Waray	Welsh	Western Frisian
Western Mari	Western Panjabi	Yakut	Yiddish	Yoruba	

Stand April 2016



Search in 238 Corpus-Based Monolingual Dictionaries for 219 Languages.

🔍 ?

 Korpus: German (deu_newscrawl_2011)

German newspaper corpus based on material crawled in 2011.
Sentences: 26,142,898 · Types: 5,876,655 · Tokens: 425,703,278

German English French Arabic Russian  all...

 Zufällige Wörter:

wahrscheinlich längst täglich 13. Gerät

Stand April 2016

Korpus

A Abkhazian Acoli Afrikaans Akan Albanian Amharic **Arabic** Aragonese Armenian Assamese Assyrian Neo-Aramaic Asturian Avaric Azerbaijani

B Bambara Banjar Bashkir Basque Bavarian Belarusian Bengali Bishnupriya Breton Bulgarian Buriat

C Catalan Cebuano Central Bikol Central Khmer Chavacano Chechen Cherokee **Chinese**
▼ Chuvash Comish Corsican Crimean Tatar Croatian Czech

D Danish Dhivehi Dimli Dutch

E Eastern Mari Egyptian Arabic Emiliano-Romagnolo **English** ▼ Esperanto Estonian

F Faroese Fiji Hindi Fijian Finnish French Friulian Fulah

G Gagauz Galician Ganda ▼ Georgian German ▼ Gilaki Goan Konkani Guarani Gujarati

H Haitian Hausa Hebrew Hindi Hungarian

I Icelandic Ido Iloko Indonesian Interlingua Interlingue Irish Italian

J Japanese Javanese

K Kölsch Kabardian Kabyle Kalaallisut Kalmyk Kannada Kara-Kalpak Karachay-Balkar Kashubian Kazakh Kirghiz Klingon Komi Konkani **Korean** Kurdish

L Ladino Lak Lao Latgalian Latin Latvian Ligurian Limburgan Lingala Lithuanian Low German Lower Sorbian Lushai Luxembourgish

M Macedo-Romanian Macedonian Malagasy Malay Malayalam Maltese Manx Maori Marathi Min Dong Chinese Min Nan Chinese Mingrelian Mirandese Modern Greek Moksha Mongolian ▼

N Nahuatl Navajo Nepali Newari Northern Frisian Northern Sami Northern Uzbek ▼ Norwegian Bokmål Norwegian Nynorsk Novial

O Occitan Old English Old Norse Oriya Oromo Ossetian

P Pampanga Pangasinan Panjabi Papiamentu Pedi Pennsylvania German Persian Picard Piemontese **Polish** Portuguese ▼ Pushto

R Romanian Romansh Romany **Russian** ▼ Rusyn

S Sami Samoan Samogitian Sanskrit Sardinian Scots Scottish Gaelic Serbian Serbo-Croatian Sicilian Silesian Sindhi Sinhala Slovak Slovenian Somali Southern Sotho **Spanish** ▼ Sundanese Swahili ▼ **Swedish** Swiss German

T Tagalog Tajik Tama (Colombia) Tamil Tatar ▼ Telugu Tetum Thai Tibetan Tigrinya Tok Pisin Tonga (Tonga Islands) Tswana **Turkish** Turkmen

U Udmurt Uighur **Ukrainian** Upper Sorbian ▼ Urdu

V Venda Venetian Vietnamese Vlaams

W Walloon Waray (Philippines) Welsh West Central Oromo Western Frisian Western Mari Western Panjabi Wolof

Y Yakut Yiddish Yoruba

Z Zeeuws Zhuang Zulu ▼