Textdatenbanken

Sommersemester 2020

2. Vorlesung

Uwe Quasthoff

Universität Leipzig
Institut für Informatik
quasthoff@informatik.uni-leipzig.de

Facts about the Leipzig Corpora Collection (- 2006)

History: Projekt Deutscher Wortschatz

- Collection of word lists since mid-90s
- 1996: Collection of sample sentences in a relational database
- 1998: Searchable via http://wortschatz.uni-leipzig.de, 3 Mio. sentences
- 2001: Starting the daily collection of newspapaer texts
- 2002: First collection of Web-text
- 2003: Larger German Corpus with 35 Mio. sentences
- 2004: Language Detection on sentence level
- 2005: WebServises
- 2006: Standard Size Corpora for 15 languages on DVD: and online at http://corpora.informatik.uni-leipzig.de/

Facts about the Leipzig Corpora Collection (2007 - 2015)

- 2007 April: Sammeln mit RSS-Feeds in ca. 35 Sprachen für die Wörter des Tages
- 2007 Juni: Wikipedia-Datenbanken in vielen Sprachen
- 2010 Domaincrawler: Crawling der Zeitungen von AbyZ mit HTTrack
- 2010 Korpusstatistiken auf **http://cls.informatik.uni-leipzig.de/**
- 2011 FindLinks: 10 Mrd. Seiten gecrawlt
- 2011 Stoppwort-Listen für 450 Sprachen aus UDHR und Watchtower
- 2011 Erster neuer DEU-Wortschatz seit 2006, ca. 250 Millionen Sätze
- 2012 Bibeln in mehr als 800 Sprachen
- 2013 Crawling mit Heritrix
- 2014 Prozesskette automatisiert von Archiv bis Normgrößenkorpora
- 2014 POS-Tagging für mehrere Sprachen
- 2014 NoSketch-Engine
- 2015 neues Portal basierend auf Webservices

Textgenres

- **Zeitungstext**
- **Wikipedia**
- **Web**
- Administrative Texte (Verwaltungssprache)
- Film-Untertitel (nahe an gesprochener Sprache)
- Religiöse Texte (z.B. Bibel)
- Literatur (z.B. Projekt Gutenberg)

Korpora in wie vielen Sprachen sind möglich? (1)

Wie viele Sprachen gibt es?

- Ca. 6 - 7.000
- Theoretische Obergrenze, da meist kein geschriebenes Material vorliegt.

Korpora lassen sich erstellen aus

- Zeitungen: ca. 100 - 120 Sprachen
- Wikipedias mit mehr als 10.000 Artikeln: 100 Sprachen
- Wikipedias mit mehr als 1.000 Artikeln: 200 Sprachen

Verzeichnis „aller“ Sprachen: http://www.ethnologue.org

“An encyclopedic reference work cataloging all of the world’s 6,909 known living languages.”

rank	name	code	spoken_by
1	Chinese, Mandarin	cmn	845.456.760
2	Spanish	spa	328.518.810
3	English	eng	328.008.138
100	Croatian	hrv	5.546.590
200	Tswa	tsc	1.180.000
300	Adyghe	ady	499.300
400	Nyemba	nba	231.540
500	Boko	bqc	110.000

Wieviel Text pro Sprecher (bisher)?

Alle Zahlen grob gerundet

Landessprachen:

- DEU: 300 Millionen Sätze / 100 Millionen Sprecher
- ENG: 650 Millionen Sätze / 330 Millionen Sprecher

Geförderte Minderheitensprachen:

- HSB (Sorbisch): 230.000 Sätze / 50.000 Sprecher
- MRI (Maori (Web), Neuseeland): 110.000 Sätze / 60.000 Sprecher

Nicht gefördert und/oder in Entwicklungsländern

- Fast nur Bibel (-50.000 Sätze) / mehrere Millionen Sprecher

Textsammlung 1

http://www.abyznewslinks.com/

- Ca. 12.000 Zeitungen
- In ca. 120 Sprachen
- Einmaliger kompletter Download
  Heritrix (s. später)
- Täglicher Download, falls RSS-Feeds angeboten werden

ABYZ News Links — Nepal
Newspapers and News Media Guide

Nepal Newspapers and News Media - Local
Bagmati

Kathmandu	Annapurna Post	NP GI NEP
Kathmandu	Budhabar	NP GI NEP
Kathmandu	Deshantar	NP GI NEP
Kathmandu	Dishanirdesh	NP GI NEP
Kathmandu	Drishti	NP GI NEP
Kathmandu	Gorkhapatra	NP GI NEP
Kathmandu	Himalayan Times	NP GI ENG
Kathmandu	Jana Aastha	NP GI NEP
Kathmandu	Janadharna	NP GI NEP
Kathmandu	Kantipur	NP GI NEP
Kathmandu	Kathmandu Post	NP GI ENG
Kathmandu	Mahanagar	NP GI NEP
Kathmandu	Nepal Samacharpatra	NP GI NEP
Kathmandu	Nepali Times	NP GI NEP
Kathmandu	Nispakshya	NP GI NEP
Kathmandu	People's Review	NP GI ENG
Kathmandu	Rajdhani	NP GI NEP
Kathmandu	Rising Nepal	NP GI NEP
Kathmandu	Sandhya Times	NP GI NEW
Kathmandu	Saptahik	NP GI NEP
Kathmandu	Tasapaw	NP GI NEW
Kathmandu	Wave	NP AL ENG
Kathmandu	Weekly Telegraph	NP GI ENG

Textsammlung 2: Wikipedias

1 Wikipedia mit mehr als 3.500.000 Artikeln: Englisch (English)

2 Wikipedias mit mehr als 1.500.000 Artikeln: Deutsch – Französisch (Français)

7 Wikipedias mit mehr als 500.000 Artikeln: Italienisch (Italiano) – Japanisch (日本語) – Niederländisch (Nederlands) – Polnisch (Polski) – Portugiesisch (Português) – Russisch (Русский) – Spanisch (Español)

6 Wikipedias mit mehr als 250.000 Artikeln: Chinesisch (中文) – Finnisch (Suomi) – Katalanisch (Català) – Norwegisch (Norsk (bokmål)) – Schwedisch (Svenska) – Ukrainisch (Українська)

20 Wikipedias mit mehr als 100.000 Artikeln: Arabisch (العربية) – Bulgarisch (Български) – Dänisch (Dansk) – Esperanto (Esperanto) – Hebräisch (עברית) – Indonesisch (Bahasa Indonesia) – Koreanisch (한국어) – Litauisch (Lietuvių) – Malaiisch (Bahasa Melayu) – Persisch (فارسی) – Rumänisch (Română) – Serbisch (Српски / Srpski) – Slowakisch (Slovenčina) – Slowenisch (Slovenščina) – Tschechisch (Česky) – Türkisch (Türkçe) – Ungarisch (Magyar) – Vietnamesisch (Tiếng Việt) – Volapük (Volapük) – Wáray-Wáray (Winaray)

14 Wikipedias mit mehr als 50.000 Artikeln: Aserbaidschanisch (Azərbaycanca) – Baskisch (Euskara) – Einfaches Englisch (Simple English) – Estnisch (Eesti) – Galicisch (Galego) – Griechisch (Ελληνικά) – Haitianisch (Kreyòl ayisyen) – Hindi (हिन्दी) – Kroatisch (Hrvatski) – Latein (Latina) – Newari (नेपाल भाषा) – Norwegisch (Nynorsk) – Tagalog (Tagalog) – Thailändisch (ไทย)

Text Collection 3: *Google + Stopwords*

Für eine Sprache wird wenige Seiten Text benötigt, um daraus die häufigsten Wörter zu extrahieren.

Verfahren: Aus den häufigsten 50 Wörtern werden jeweils 3 bis 5 zufällig ausgewählt und an die Suchmaschine geschickt. Die zurückgelieferten Texte werden gesammelt.

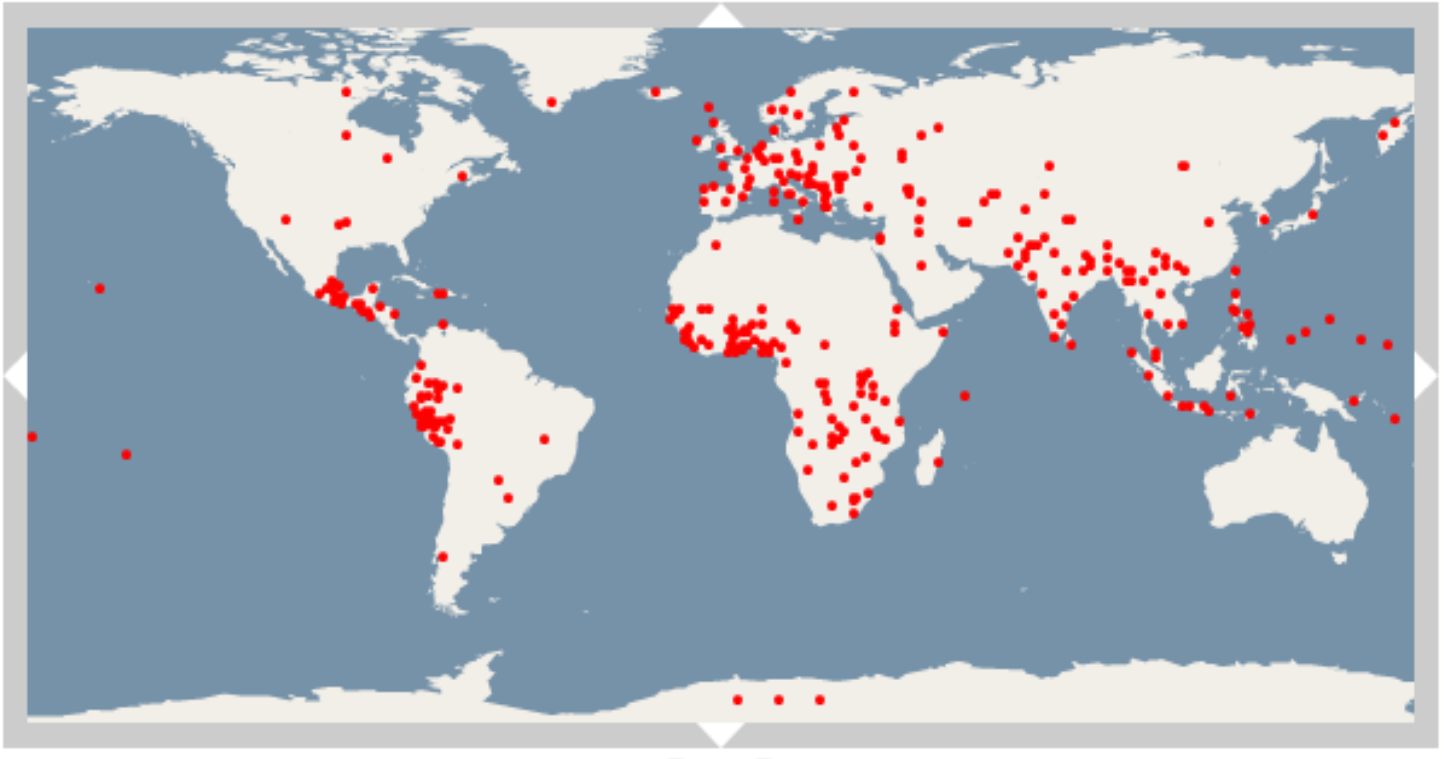
Beispiel für Deutsch: *der es für in durch*

Obwohl *in* allein nicht nur in deutschen Texten vorkommt, sorgen die weiteren Wörter für Auswahl der korrekten Sprache.

Quelle für Suchwörter: UDHR in Unicode

Die *Allgemeine Erklärung der Menschenrechte* der Vereinten Nationen liegt in über 370 Sprachen in Unicode vor.

Umfang (deutsch): 1700 Wörter



Textsammlung 4: Domain Crawling mit *Heritrix* (ab 2013)

- Entwickelt und benutzt von Archive.org
- Zusammen mit nationalen Bibliotheken aus: USA, Canada, UK, France, Iceland, Sweden, Norway, Finland, Denmark, Italy, Australia

Unser Vorgehen

Crawling von Zeitungen: Einmal jährlich mit Liste der Domains von www.abyznewslinks.com

Crawling des Web:

- Einmal jährlich jede TLD
- Funktioniert für alle TLDs außer den zwei größten (.com und .cn)
- Problematisch ist hier die Größe der Warteschlange
- Laufzeit 1-60 Tage pro TLD je nach Größe, bis zu 5 parallel pro Rechner.

Korpora in wie vielen Sprachen sind möglich? (2)

Erschlossene Quellen zur Textsammlung durch Crawling (alles in utf-8)

- Wikipedia: insgesamt: 270 Sprachen
- UDHR: 360 Sprachen
- http://www.watchtower.org/: Texte in 418 Sprachen, davon ca. 320 in utf8
- Bibel (zumindest große Teile, mehr als 7.000 Verse): >800 Sprachen

Weitere Quellen zur Textsammlung durch Crawling

Unterschiedlichste Formate, teilweise Bitmaps

- http://www.christusrex.org/www1/pater/: Vaterunser in 1697 Sprachen
- http://www.unitedbiblesocieties.org/: Bibelteile in 2508 Sprachen (Stand 2010)

Sprache ermitteln I

- Die Sprache eines Textes muss ermittelt bzw. verifiziert werden.
- Die Aufgabe ist um so einfacher, je länger ein Text ist.
- Sehr leicht für Texte mit N=200 Wörtern oder mehr:
  - Benutze für jede der zu untersuchenden Sprachen die Liste der häufigsten L=50 Wörter.
  - Stelle fest, welche Liste die meisten Vertreter im Text hat.
  - Falls es einen deutlichen Sieger gibt, ist dies die Sprache des Textes.
- Der Algorithmus funktioniert auch für kleinere N (z.B. N=10, ein Satz), wenn L entsprechend größer gewählt wird. Faustregel: N*L=10.000

Voraussetzung für diesen Algorithmus: Sprache bekannt, denn Stoppwörter müssen bereitgestellt werden.

Sprache ermitteln II

Zweistufiges Verfahren:

1. Schritt: Identifikation der Dokumentsprache (d.h. auf Dokumentenbasis) mit ca. 50 Stoppwörtern. Danach Zusammenfassung aller Dokumente einer Sprache. Zerlegung in Sätze.

2. Schritt: Sprachidentifikation auf Satzbasis mit den 5.000 häufigsten Wörtern pro Sprache. Sätze mit falscher Sprache werden verworfen.

Sprache ermitteln 3: Buchstabentrigramme

- Crawler language trigram vectors of the 30 most frequent trigrams for known languages
- For each crawled page, the 30 most frequent trigrams are determined
- This vector is compared to the language trigram vectors. If they agree in at least 12 trigrams, the document is assumed to be in this language.

Sample language trigram vectors:

```
de 12 en_170 er_132 _de86 der68 ie_68 ich61 sch59 ein58 ch_53 die49 _di48 che47 den44 nd_43
    in_42 ten42 und39 _ei36 n_d36 gen36 ine36 _un35 cht35 ung34 nde33 n._32 ter32 te_30 _au28
    es_28
dk 12 er_149 en_100 et_85 _de79 for55 der55 _og54 de_54 og_53 _fo47 ing46 nde45 _i_41 til40
    _ti39 _me39 ere39 den38 at_37 ter36 _at35 _af34 il_34 _er34 re_33 ed_32 _en32 or_30 det30
    lle29
ee 12 _,_96 _._84 st_51 se_50 le_48 ud_44 ja_42 mis42 on_41 _se41 ise40 use38 _on38 est37 ast36
    _ko35 sel35 ist34 ks_34 _ka34 da_33 _ja33 sta32 es_32 _te32 id_31 ga_31 _va30 ust29 te_28
en 12 _th133 he_116 the113 ed_64 _in55 ing52 _of52 _to50 ng_46 _an46 to_46 of_46 nd_43 ion42
    and42 er_39 on_39 in_38 _a_36 ent36 _co35 es_32 _re29 s_a29 as_28 tio28 re_28 d_t28 at_26
    or_26
fi 12 en_113 an_58 in_56 ist56 ja_55 sta50 _ja50 ta_48 on_43 n_k40 aan37 ise36 ssa36 n_t34
    tta33 a._31 itt30 _va30 sen30 _on29 sa_29 lla28 tä_28 ksi27 taa27 ett26 lis26 _ta26 een26
    ais24
fr 12 _de131 es_123 de_96 nt_68 _le68 ent67 e_d62 le_58 s_d50 _la49 la_47 ion47 e_l46 re_46
    on_43 les40 _qu39 ne_38 _co38 ur_37 que37 ns_36 et_35 _pa35 tio34 _à_34 _l'33 e_p33 our33
    t_d32
```

Neue Sprachen entdecken

If the trigram vector of a document is not similar to one of the predefined language trigram vectors:

- Send trigram vector to server
- Cluster trigram vectors
- Some clusters may correspond to “new” languages.

Sample clusters:

Icelandic:

```
http://www.vlfs.is/, http://www.gransking.fo/Default.asp?sida=6,
http://xd.is/skipulag/stjornmalaskolinn/, http://www.bladid.is/index.php?
id=1&tx_ttnews[pointer]=6&cHash=36436b1024, http://www.ttfi.is/i_deiglunni/verkefni.htm,
http://www.computer.is/umokkur/, http://www.melavollur.is/fullstory.php?idStory=20,
http://www.veislan.is/press/fyrirt.asp?strAction=getPublication&intPublId=124,
http://veldi.is/, http://myndir.grundaskoli.is/sjo/undirsidur/menning/tonlist/kvoldsigl.htm
```

Also link farms etc.:

```
http://5515.n7ky2n.info/, http://122.kfupkj.info/, http://msserversql.muonsql.com
/mssqlserverstoredprocedures/, http://4924.ck45ve.info/
```

Die indo-europäische Sprachfamilie in Europa

heutige Verbreitung der indo-europäischen Sprachen und ihre Nachbar-sprachen in Europa

Die indo-europäische Sprachfamilie in Asien

Austro-Asiatic

Dokumente

Text liegt in folgender Form vor:

<source><location>http://…</location><date>…</date>…<language>deu</language> > </source>

Die beiden englischen Vereine FC Chelsea ...

Dies ist nicht das ganze Originaldokument, sollte aber den Text (nicht aber die Bilder, Tabellen, Links, ...) des Originaldokuments im wesentlichen vollständig enthalten.

Nächster Schritt: Wir zerlegen den Text in die nächstkleineren Teile: Sätze.

Dabei verlieren eventuelle Unvollständigkeiten auf der Dokumentebene an Bedeutung.

Regeln zur Satzsegmentierung I

Zunächst einige einfache Regeln für den Satzanfang:

- Sätze beginnen niemals mit Kleinbuchstaben.
- Nach einer Überschrift beginnt ein neuer Satz.
- Am Anfang eines Absatzes beginnt ein neuer Satz.
- Groß geschriebene Artikel (wie *Der*, *Die*, *Den*, ...) sprechen für einen Satzanfang.
- Beginnt kein neuer Absatz, so steht vor dem neuen Satz ein Satzendezeichen.

Regeln zur Satzsegmentierung II

Analog gibt es einige einfache Regeln für das Satzende:

- Sätze enden mit einem Satzendezeichen. Solche Satzendezeichen sind Punkt, Fragezeichen und Ausrufezeichen. Nach dem Satzendezeichen muss zusätzlich ein *white space* (meist ein Leerzeichen, s.u.) stehen. Achtung, Punkte können auch an anderer Stelle stehen, z.B. in URLs, nach Abkürzungen oder Zahlen.
- Vor einer Überschrift endet ein Satz.
- Am Ende eines Absatzes endet ein Satz.

Schwierige Fälle

- Er trägt den Titel Dr. rer. nat.
- Seit einem halben Jahr gehört Dr. rer nat. Stefan Schlatt dazu.
- Sein Glückstag ist Freitag der 13.
- Gestern war es wieder soweit: Freitag der 13. März.

Ein Satz oder mehrere?

- „Ich kann es hören! Es kommt immer näher“, rief er entsetzt.

„Natürliche“ Zerlegung:

- „Ich kann es hören!
- Es kommt immer näher“, rief er entsetzt.

Satzenden in verschiedenen Sprachen I

Standard Latin

There is usually no blank before puntuation (with exception French) and one blank (or two in English) following punctuation.

```
. U+002E FULL STOP
! U+0021 EXCLAMATION MARK
? U+003F QUESTION
```

Exceptions in Greek

In Greek, the semicolon is used as question mark in the following forms:

```
; U+003B SEMICOLON
; U+037E GREEK QUESTION MARK
```

Exceptions in Spanish, Galician and Leonese

In Spanisch, Galician and Leonese there are an additional opening exclamation mark and an opening question mark. Both do not influence sentence segmentation.

```
¡ U+00A1 INVERTED EXCLAMATION MARK
¿ U+00BF INVERTED QUESTION MARK
```

Satzenden in verschiedenen Sprachen II

Arabic

- U+06D4 ARABIC FULL STOP
؟ U+061F ARABIC QUESTION MARK

Chinese, Japanese and Korean

Here we have full width forms of the punktuation marks. But, the usual standard latin puntuation marks might be used, too.

。 U+3002 Ideographic Full Stop
！ U+FF01 FULLWIDTH EXCLAMATION MARK
？ U+FF1F FULLWIDTH QUESTION MARK

Devanagari script

For Hindii, Sanskrit, Nepal, and some other Indian languages.

। U+0964 Devanagari Danda

Remark: The standard full stop is used for abbreviations and may appear within sentences.

Armenian

՝ U+055C armenian exclamation point
՞ U+055E armenian question mark

Sprachen ohne Satzzeichen am Satzende

Khmer benutzt Frage- und Ausrufezeichen, aber kein Zeichen am Ende eines Aussagesatzes. Jedes der (wenigen) Leerzeichen kann Satzende sein. Analog ***Laotisch*** und ***Thailändisch***. Es gibt momentan nur für Thai Programme, die solche Texte weiter zerlegen. Das Ergebnis sind nicht immer Sätze in unserem Sinne.

ក្រោយបញ្ចប់សវនាការកាត់ទោសឌុច រយៈពេលប្រាំថ្ងៃចុងក្រោយនេះ គឺចាប់ពីថ្ងៃទី២០ ដល់ថ្ងៃទី២៨ ខែកក្កដា ការមន្ទិលសង្ស័យនៅតែបន្តកើតមាន

ដដែលៗ ចំណោទនៅតែមាន ពោលគឺ តើការកាត់ទោសឌុច នេះមិនមែនកំពុងតែដើរលើផ្លូវខុសទេឬ? សំណួររបស់ភាគីដើមចោទមានភាពស្រពិចស្រពិល

និងមិនមានលំដាប់លំដោយៗ ដោយឡែក សាក្សីវិញ ពួកគេហាក់ដូចជាមានការស្ទាក់ស្ទើរនៅក្នុងការឆ្លើយចំឡើការពិតនៅចំពោះមុខតុលាការ ដោយសារពួក

គេមានការភិតភ័យក្រែងចម្លើយទាំងប៉ុន្មានរបស់ពួកគេបែរត្រឡប់មកដាក់បន្ទុកលើរូបខ្លួនវិញ ព្រោះពួកគេទាំងអស់នោះសុទ្ធតែជាអតីតខ្មែរក្រហមបម្រើការ

នៅក្រោមបង្គាប់របស់ឌុច នៅមន្ទីរសន្តិសុខ ស-២១ៗ មិនត្រឹមតែប៉ុណ្ណោះ ជានិច្ចកាល ការកាត់ទោសនេះហាក់បីដូចជាឃ្លាតចាកឆ្ងាយពីអង្គហេតុស្តីពីតួនាទី

របស់ឌុចនៅក្នុងមន្ទីរស-២១ បើទោះបីជាជនជាប់ចោទបានទទួលសារភាពកំហុសនិងព្រមទទួលខុសត្រូវផ្នែកសតិអារម្មណ៍ក៏ដោយៗ