Textdatenbanken

Sommersemester 2020

9. Vorlesung

- Klassische Korpuslinguistik -

Uwe Quasthoff

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

Zwei Wege der Lehrmaterialien über Korpuslinguistik

Def.: Die Korpuslinguistik ist ein Bereich der Linguistik, in dem Theorien über Sprache anhand von Belegen oder statistischen Daten aus Textkorpora aufgestellt oder überprüft werden. (*wikipedia*)

1. Möglichkeit: Ein spezieller Kurs für Linguisten
   - mit der Vorstellung bekannter Korpora sowie
   - einfachen Rezepten zur Programmierung „rund um Korpora“
   - Grundkurs grep und reguläre Ausdrücke
2. Möglichkeit: Bearbeitung spezieller linguistischer Fragestellungen mit Methoden der Korpuslinguistik
   - alte Fragestellungen mit neuen Methoden
   - isolierter statt systematischer Einsatz

Beispielkurs: Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge

Von *Noah Bubenhofer*, Universität Zürich

1. Einführung in die Korpuslinguistik: Korpustypen, Erstellung, Annotationen, Anfragesysteme
2. Web als Korpus: Wo liegen die Chancen und Risiken der Nutzung des Internets als linguistisches Korpus?
3. COSMAS II: Eines der wichtigsten Korpora deutscher Sprache des Instituts für Deutsche Sprache (IDS) in Mannheim. Einführung in die Bedienung und die Abfragesprache.
4. Weitere Korpora: Kurze Einführungen in weitere wichtige deutschsprachige Korpora.
5. Eigenes Korpus: Hilfe und Tipps zur Erstellung eines eigenen Korpus'.
6. Datenbank Filemaker: Dieses Datenbankprogramm bietet sich an zur einfachen Verwaltung des eigenen Korpus'.
7. Anwendungen: Beispiele für die Arbeit mit Korpora
8. Visualisierung: Einführung in die Möglichkeiten der Visualisierung von Sprachdaten.
9. Anhang: Informationen zu korpuslinguistischer Software, kleine Einführungen in grundlegende Unix-Befehle und in Reguläre Ausdrücke, sowie Literaturhinweise und ein Lexikon.

Daraus ein Beispiel für eine spezielle Fragestellung

Collostructions

Ausgehend von der "construction grammar"/"Konstruktionsgrammatik" experimentieren Stefanowitsch und Gries (2005) mit einer Kollokationsanalyse, die neben lexikalischen Elementen auch grammatische Konstruktionen berücksichtigt. So untersuchen Sie verschiedene Korpora z.B. auf englischen s-Genitiv-Konstruktionen, die prototypischerweise so beschrieben werden:

$NP_{possessors}$ *'s* $N_{possessee}$

Wie z.B.: *John's book, Mary's sister*

Dabei zeigen Sie (u.a.), dass die Semantik dieser Konstruktionen keinesfalls so klar ist, wie die prototypischen Fälle zeigen. Man findet z.B. in einem Korpus erstaunlich wenige Fälle, wo der s-Genitiv tatsächlich Besitz ausdrückt, dafür Konstruktionen wie:

Produzent-Produkt: *Roland's synth, [Pers. Name]'s [Work of Art]*

Partizipant-Ereignis: *earth's rotation, farmer's workshop, partner's earning*

Zeitpunkt-Ereignis: *tomorrow's final, moment's notice*

etc.

Die Semantik der s-Genitiv-Konstruktion im Englischen muss also differenzierter beschrieben werden.

Übersicht über Anwendungsbereiche für Korpora

- Computational Linguistics
- Cultural Studies
- Discourse Analysis and Pragmatics
- Grammar/Syntax
- Historical Linguistics
- Language Acquisition
- Language Teaching
- Language Variation
- Lexicography
- Linguistics
- Machine Translation
- Natural Language Processing (NLP)
- Psycholinguistics
- Semantics
- Social Psychology
- Sociolinguistics
- Speech
- Stylistics

Aus der Sicht des Linguisten

Corpus linguistics

an introduction

What is a corpus?

- A collection of **naturally occurring** language text, chosen to characterise a **state** or **variety** of language (Sinclair)

- A collection of linguistic data, either **written text** or a **transcription of recorded data**, which can be used as **starting-point of linguistic description** or as **a means of verifying hypotheses** about a language (Dictionary of linguistics and phonetics)

What is a corpus? (II)

- **Large body** of evidence typically composed of **attested language** use (McEnery)
- Usually a corpus is in **machine-readable format** and is ideally **viewable and analysable** through (a single) software package
- The word *corpus* comes from Latin *body* and the plural is *corpora*

What is not a corpus

- Lists of words
- Lists of sentences produced with the purpose of creating a corpus
- Archive = “a repository of readable electronic texts not linked in any coordinated way” (http://www.archive.org)

  “The Internet Archive is building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public.“

Corpus vs. archive

Text archive

- Collection of texts in their original format

  (Oxford Text Archive: http://ota.ox.ac.uk/)

Corpus

- texts collected and processed in a unified, systematic manner

  British National Corpus: http://www.natcorp.ox.ac.uk/

Why bother with corpora?

- Even “expert speakers” have only a partial knowledge of a language
  A corpus can be more comprehensive and balanced
- Even expert speakers tend to notice the unusual and think of what is possible
  A corpus can show us what is common and typical
- Even expert speakers cannot quantify their knowledge of language
  A corpus can give us accurate statistics

Why bother with corpora? (II)

- Even expert speakers cannot remember everything they know
  A corpus can store and recall all the information that has been input
- Even experts speakers cannot make up natural examples
  A corpus can provide us with a vast number of real examples
- Even expert speakers have prejudices and preferences and every language has cultural connotations and underlying ideology
  A corpus can give you more objective evidence

Why bother with corpora? (III)

- Even expert speakers are not always available to be consulted
  A corpus can be made permanently accessible to all
- Even expert speakers cannot keep up with language change
  A constantly updated corpus can reflect even recent changes in the language
- Even expert speakers lack authority: they can be challenged by other expert speakers
  A corpus can encompass the actual language use of many expert speakers

Sozialwissenschaftliche Anwendung: Sprachliche Stereotype

-

Findige Dänen und deutsche Gründlichkeit

Sprachliche Stereotype

Was denken Deutsche über Dänen und umgekehrt?

In Korpora soll gesucht werden

- nach Attributen für Dänen (in deutschsprachigen Texten) und
- nach Attributen für Deutsche (in dänischsprachigen Texten) und

Kandidaten sind

- Adjektive als linke Nachbarn zu „Deutsche“ und „Dänen“
- Komposita mit „Deutsche“ und „Dänen“
- Substantive als rechte Nachbarn zu „deutsche“ und „dänische“

Linke Nachbarn von Dänen: Gesamt

```
Use deu_mixed_2012
select w1.*, w2.word, c.freq, sig from words w1, words w2 , co_n c where w1.w_id=w1_id and w2.w_id=w2_id
    and w2.word="Dänen" and w1_id>1000 and w1.word regexp "^[a-z]" order by sig desc limit 30;
```

w_id	word	freq	word	freq	sig
24107	gebürtigen	12014	Dänen	25	173.48
47861	angetretenen	4941	Dänen	20	165.21
51877	verbündeten	4443	Dänen	15	118.48
5594	lebenden	70664	Dänen	30	115.48
30089	**sympathischen**	**9074**	**Dänen**	**15**	**97.26**
29360	siegreichen	9375	Dänen	13	79.79
1397	**starken**	**306312**	**Dänen**	**40**	**69.88**
7186	verletzten	52669	Dänen	19	67.32
24367	spielenden	11846	Dänen	12	66.27
13624	gesetzten	24381	Dänen	14	61.92
70970	**biederen**	**2911**	**Dänen**	**8**	**59.92**
475505	spindeldürren	197	Dänen	5	59.71
68942	aufspielenden	3028	Dänen	8	59.29
45718	eingestuften	5237	Dänen	9	59.04
109545	ausgeschlossenen	1612	Dänen	7	58.8
65818	zweitplatzierten	3225	Dänen	8	58.29
23129	**harmlosen**	**12610**	**Dänen**	**11**	**57.54**

Linke Nachbarn von Dänen: 2011

```
mysql> use deu_newscrawl_2011
mysql> select w1.*, w2.word, c.freq, sig from words w1, words w2 , co_n c where w1.w_id=w1_id and
    w2.w_id=w2_id and w2.word="Dänen" and w1_id>1000 and w1.word regexp "^[a-z]" order by sig desc limit
    30;
```

w_id	word	freq	word	freq	sig
62583	**findige**	**334**	**Dänen**	**6**	**90.94**
21709	**harmlosen**	**1346**	**Dänen**	**4**	**46.17**
93186	zweitklassigen	193	Dänen	3	44.59
56079	zweitplatzierten	388	Dänen	3	40.38
6650	stehende	5596	Dänen	4	34.79
20836	hunderten	1416	Dänen	3	32.6
5263	lebenden	7384	Dänen	4	32.59
4829	jüngeren	8200	Dänen	4	31.75
10668	qualifizierten	3217	Dänen	3	27.68
7234	schwache	5086	Dänen	3	24.95

Linke Nachbarn von Dänen: 2009

```
mysql> use deu_news_2009
mysql> select w1.*, w2.word, c.freq, sig from words w1, words w2 , co_n c where w1.w_id=w1_id and w2.w_id=w2_id and
    w2.word="Dänen" and w1_id>1000 and w1.word regexp "^[a-z]" order by sig desc limit 30;
```

w_id	word	freq	word	freq	sig
58476	**biederen**	**362**	**Dänen**	**7**	**106.22**
34183	eingestuften	752	Dänen	7	95.92
1089	starken	43802	Dänen	10	63.1
10622	qualifizierten	3427	Dänen	5	49.96
10263	engagierten	3568	Dänen	5	49.55
2931	führenden	15814	Dänen	6	43.87
16375	gebürtigen	1966	Dänen	4	42.62
28833	sympathischen	942	Dänen	3	34.65
26830	siegreichen	1037	Dänen	3	34.08
25832	trainierten	1092	Dänen	3	33.77
20260	robusten	1489	Dänen	3	31.91
17707	favorisierten	1780	Dänen	3	30.84
13716	gesperrten	2479	Dänen	3	28.85
9693	**freundlichen**	**3830**	**Dänen**	**3**	**26.25**

Beispiele für *findige Dänen*

```
mysql> select * from sentences where sentence like "% findige Dänen %";
```

Aus Discounter-Schnäppchen haben 2 **findige Dänen** eine Rakete gebaut.

Aus Supermarkt-Schnäppchen haben zwei **findige Dänen** eine Rakete gebaut.

Spannung vor einem höchst ungewöhnlichen Raketentest auf der idyllischen Ostseeinsel Bornholm: Zwei **findige Dänen** wollen am Donnerstag ihre selbst gebaute Billig-Rakete «Tycho Brahe» mit einer Astronautenpuppe in der Spitze Richtung Weltraum abschiessen.

Spannung vor einem höchst ungewöhnlichen Raketentest auf der idyllischen Ostseeinsel Bornholm: Zwei **findige Dänen** wollen am Donnerstag ihre selbst gebaute Billig-Rakete „Tycho Brahe“ mit einer Astronautenpuppe in der Spitze Richtung Weltraum abschießen.

Spannung vor einem höchst ungewöhnlichen Raketentest auf der idyllischen Ostseeinsel Bornholm: Zwei **findige Dänen** wollen heute ihre selbst gebaute Billig-Rakete „Tycho Brahe“ mit einer Astronautenpuppe in der Spitze Richtung Weltraum abschießen.

Zwei **findige Dänen** basteln aus Baumarktteilen eine eigene Rakete.

Beispiele für *findige Dänen*

Das war noch das Beste an einem Abend mit lausigem Sommerfußball, an dem sich die Hanseaten mit einer 0:1 (0: 1)-Niederlage gegen die **biederen Dänen** blamierten.

Das war noch das Beste an einem Abend mit lausigem Sommerfußball, an dem sich die Hanseaten mit einer 0:1 (0:1)-Niederlage gegen die **biederen Dänen** blamierten.

Niederlage gegen die **biederen Dänen** blamierten.

Trainer Bruno Labbadia nutzt das Rückspiel gegen die **biederen Dänen** heute Abend (20.30 Uhr im Live-Ticker) zum Debütantenball der Millionenstars.

Trainer Bruno Labbadia nutzt deshalb das Rückspiel gegen die **biederen Dänen** am Donnerstag (20.30 Uhr/live im ZDF) zum Debütantenball der Millionenstars.

Trainer Bruno Labbadia nutzt deshalb das Rückspiel gegen die **biederen Dänen** am heutigen Donnerstag (20.30 Uhr/live im ZDF) zum Debütantenball der Millionenstars. |

Beispiele für *freundliche Dänen*

Doch, man komme mit den **freundlichen Dänen** sehr gut über ins Gespräch über die hier den meisten komplett unbekannte Zweitliga-Elf, meinte entspannt ein Mitglied des Fanshop-Teams.

Doch, man komme mit den **freundlichen Dänen** sehr gut ins Gespräch über die hier den meisten komplett unbekannte Zweitliga-Elf, meinte entspannt ein Mitglied des "Fanshop-Teams".

Doch, man komme mit den **freundlichen Dänen** sehr gut ins Gespräch über die hier den meisten komplett unbekannte Zweitliga-Elf, meinte entspannt ein Mitglied des «Fanshop-Teams».

Linke Nachbarn von Deutschen: 2009

```
mysql> use deu_news_2009
mysql> select w1.*, w2.word, c.freq, sig from words w1, words w2 , co_n c where w1.w_id=w1_id and w2.w_id=w2_id and
    w2.word="Deutschen" and w1_id>1000 and w1.word regexp "^[a-z]" order by sig desc limit 30;
```

w_id	word	freq	word	freq	sig
12971	zweitgrößten	2661	deutschen	432	4488.59
1404	wichtigsten	33712	deutschen	659	3978.96
2931	führenden	15814	deutschen	534	3807.14
6966	erfolgreichsten	5756	deutschen	406	3503.39
2313	höchsten	20404	deutschen	403	2441.04
2632	einzigen	17784	deutschen	361	2206.18
2268	Internationalen	20802	Deutschen	326	2067.3
29452	gekaperten	915	deutschen	187	2037.66
21803	drittgrößten	1352	deutschen	189	1902.39
4059	Allgemeinen	11071	Deutschen	256	1821.77
1302	gesamten	37212	deutschen	351	1617.56
17738	reichsten	1776	Deutschen	149	1451.34
2789	angeschlagenen	16765	deutschen	257	1427.82
. . .					
37096	**hässlichen**	**673**	**Deutschen**	**60**	**592**
10181	getöteten	3606	deutschen	85	544.62

Beispiele für *hässlichen Deutschen*

60 Treffer, davon über 50 mit Peer Steinbrück:

"Peer Steinbrück definiert das Bild des hässlichen Deutschen neu", hatte Müller im Nationalrat gesagt.

"Peer Steinbrück, das darf man in aller Offenheit sagen, definiert das Bild des hässlichen Deutschen neu", wetterte Thomas Müller von der christdemokratischen Volkspartei CVP.

"Stattdessen sitzt die Kanzlerin still neben ihm auf der Regierungsbank und sieht zu, wie Herr Steinbrück das Bild des hässlichen Deutschen in der ganzen Welt verbreitet", sagte er der "Welt am Sonntag".

Der konservative Schweizer Parlamentarier Müller hatte gesagt, Peer Steinbrück definiere das Bild des hässlichen Deutschen neu.

Ein Schweizer Abgeordneter bezeichnete in dem Streit den deutschen Finanzminister Peer Steinbrück, der das Land besonders hart kritisiert hatte, als jemanden, der das Bild des hässlichen Deutschen neu definiere.

Im Zusammenhang mit dem deutschen Finanzminister Peer Steinbrück bemühte der St. Galler CVP-Vertreter Thomas Müller sogar einen Nazi-Vergleich: Steinbrück definiere das Bild des hässlichen Deutschen neu, sagte er.

Beispiele für *hässlichen Deutschen*

60 Treffer, davon 8 ohne Peer Steinbrück, trotzdem 3x implizit:

Am erfolgreichsten gelang dies, als er beißende politische Satire im warmen Gelsenkirchener Barock einer Familienserie versteckte: "Ein Herz und eine Seele" hielt dem hässlichen Deutschen einen Spiegel vor – und siehe da, er riss sich förmlich danach.

Die "Zonis" dürften naturgemäß ganz anderer Meinung sein und ihrerseits den Wessi für den Prototyp des hässlichen Deutschen halten, was ebenso berechtigt wäre.

Die 66-Jährige steht im Nachbarland für all das, was mit dem hässlichen Deutschen verbunden wird.

Die Kleingärtner, Hundetrainer und Fußballfans verhalten sich genau so, wie man es vom hässlichen Deutschen erwartet.

Die Schweizer wiederum, erschrocken über solche Töne aus der Nachbarschaft, entdecken wieder einmal den hässlichen Deutschen und fühlen sich an die Nazis erinnert.

Jünger schaffte es, den hässlichen Deutschen als Beau zu geben.

Man kann den Volkszorn in der Schweiz über den hässlichen Deutschen nicht aus Angst vor seinen Cowboys, die keiner fürchtet, verniedlichen.

Teenager mit dem Bild des hässlichen Deutschen – und das mit Erfolg!

DEU-Wörter in DAN: *ü*

```
use dan_mixed_2012
mysql> select * from words where word like "%ü%" and word regexp "^[a-z]" order by freq desc  limit 30;
```

w_id	word	freq
32857	für	681
84279	müsli	183
115505	nüvi	117
153103	prügelknabe	78
164870	kür	70
173230	flütes	65
202359	pürsch	52
270325	müesli	34
286767	glühwein	31
352693	lübeckerne	23
...		
386773	glücksborgske	20
388722	müslibarer	20
417939	müslibar	18
431601	fingerspitzgefühl	17
555729	prügelknaben	12

DEU-Wörter in DAN: *über*

```
select * from words where word like "%über%" order by freq desc  limit 30;
+---------+--------------------+------+
| w_id    | word               | freq |
+---------+--------------------+------+
|  117798 | über               |  114 |
...
|  529631 | übercool           |   13 |
|  937559 | überseksuelle      |    6 |
| 1026608 | gegenüber          |    5 |
|  967858 | Hotelübernachtung  |    5 |
| 1303397 | überklasse         |    4 |
| 1201593 | darüber            |    4 |
| 1680171 | überste            |    3 |
| 1680170 | übermensch         |    3 |
| 1680169 | übercoole          |    3 |
| 1680168 | überboss           |    3 |
| 1680167 | über-nørden        |    3 |
| 1409141 | Machtübernahme     |    3 |
| 2539653 | überhaubt          |    2 |
| 2539654 | überliga           |    2 |
| 2539655 | überläufere        |    2 |
+---------+--------------------+------+
```

DEU-Wörter in DAN: andere Wörter

Nur, wenn man sie kennt:

```
mysql> select w1.*, w2.*, w2.freq/w1.freq from words w1, deu_news_2009.words w2 where
    w1.word=lower(w2.word) and binary w2.word<>lower(w2.word) and w2.w_id>600 and
    char_length(w2.word)>5 and w1.word in
    ("eisbein","oktoberfest","apfelstrudel","sauerkraut");
+---------+--------------+------+--------+--------------+------+-----------------+
| w_id    | word         | freq | w_id   | word         | freq | w2.freq/w1.freq |
+---------+--------------+------+--------+--------------+------+-----------------+
|  118361 | oktoberfest  |  113 |  14125 | Oktoberfest  | 2386 |         21.1150 |
|  213875 | sauerkraut   |   48 |  62155 | Sauerkraut   |  332 |          6.9167 |
| 1512155 | eisbein      |    3 | 165635 | Eisbein      |   79 |         26.3333 |
| 1189113 | apfelstrudel |    4 | 202543 | Apfelstrudel |   58 |         14.5000 |
+---------+--------------+------+--------+--------------+------+-----------------+
4 rows in set (3 min 13.98 sec)
```

Beispielsätze lesen

Mange lande er tillagt en type mad, der skulle kendetegne landet: "**pølsetyskere**," "kartoffeldanskere," og "frøædere" er nogle klassiske.

Udvalget "De Gamles Jul" havde stor succes med arrangementet den 15. januar, hvor der blev budt på **eisbein** og sauerkraut eller wienerschnitzel.

Ich hätte mir gewünscht, dass die deutschen Ingenieure, Handwerker, Eisenbahner usw. weniger ordentlich, fleißig und diszipliniert gewesen wären.

Nach ihnen muss jeder Pfadfinder treu, entschlossen, aufrecht, gerecht, ordentlich, fleißig, hilfsbereit und naturliebend sein.

Die Begeisterung gipfelt in den bekannten Klischees: deutsche Pünktlichkeit, deutsche Wertarbeit, deutsche Disziplin.

Komposita: *dänen*

```
mysql> select * from words where lower(word) like "%dänen%" and binary word like "%änen%" and word not like "% %"
    order by w_id limit 150;
```

Dänen, mondänen, Dänenkönig, Dänenweg, Dänenprinzen, Dänen-Ampel, Dänenkönigs, Dänenprinz, Mondänen, Dänenkrone, Dänenherrschaft, Dänen-Partei, Dänenstraße, Dänen-Coach, Dänenpartei, Dänen-Prinz, Dänenkronen, Dänenzeit, Dänen-Trainer, Dänen-, Dänen-Keeper, Mundänen, Dänen-Krone, Dänenkönigin, Dänen-Duo, Dänentum, Dänen-Star, Dänen-Bomber, Dänen-Prinzessin, Deutsch-Dänen, Dänen-König, Dänenland, Eiderdänen, Deutschdänen, Dänen-Prinzen, Däneninsel, Süddänen, Dänen-Königin, Dänen-Stürmer, Dänenmark, Dänen-Votum, Dänenampel, Dänenprinzessin, Neudänen, Nicht-Dänen, Speckdänen, modänen, Dogma-Dänen, Dänen-Coup, Dänen-Import, Dänen-Königs, Dänen-Tor, Dänen-Torjäger, Dänen-Trip, Dänenaxt, Dänenheide, Dänenkamp, Dänenkönige, Neu-Dänen, Auslandsdänen, Durchschnittsdänen, "Dänen"-Ampel, Dänen-Schlussmann, Dänen-Spiel, Dänen-Turbo, Dänenberg, Dänenfürst, Dänenheer, Dänenhäuptling, Dänenkönigen, Dänenstadt, Dänenstr, Dänentums, Festlandsdänen, Nichtdänen, halbmondänen, ...,

Komposita: *tysker*

```
mysql>select * from words where lower(word) like "%tysker%" and freq>3 order by w_id limit 90;
```

tyskerne, tyskere, Tyskerne, tyskeren, tysker, tyskernes, Tyskeren, Tyskernes, Tyskere, tyskerens, østtyskere, sydtyskerne, østtyskerne, nordtyskerne, Tysker, tyskeres, vesttyskerne, tyskerpigerne, tyskerpiger, kartoffeltyskerne, tyskers, kartoffeltyskere, Østtyskerne, Nordtyskerne, østtysker, hjemmetyskerne, tyskertrækket, Sydtyskerne, Tyskerens, østtyskernes, tyskerhad, Kartoffeltyskerne, Tyskerhavnen, Tyskervejen, tyskerene, Tyskerpigerne, sydtysker, sydtyskernes, tyskertræk, tyskervenlige, hjemmetyskere, tyskertøs, vesttyskere, Tyskerpiger, kartoffeltysker, nordtyskere, sydtyskere, østtyskeren, tyskerpige, Tyskerhadet, Vesttyskerne, sudetertyskere, tyskerbørn, tyskertøser, vesttysker, østtyskeres, Tyskertårnet, nytyskerne, sudetertyskerne, tyskerhadet, tyskerkursen, tyskerladen, tyskertøserne, tyskerne, utysker, volgatyskere, volgatyskerne, Østtyskere, Kartoffeltyskere, hjemmetyskernes, ikke-tyskere, nordtyskernes, pølsetyskere, rigstysker, tyskerbarak, tyskerbarakker, tyskerbarn, tyskerkurs, tyskerpigernes, vesttyskernes, …

Rechte Nachbarn von *deutschen*

Völlig unspezifisch:

```
mysql> select w1.*, w2.word, c.freq, sig from words w1, words w2 , co_n c where w1.w_id=w1_id and w2.w_id=w2_id and
    w1.word="deutscher" and w2_id>1000 and w2.word regexp "^[A-Z]" order by sig desc limit 250,30;
```

w_id	word	freq	word	freq	sig
1773	deutscher	241778	Polizisten	254	334.8
1773	deutscher	241778	Militärausbilder	32	334.34
1773	deutscher	241778	Steuergelder	77	334.05
1773	deutscher	241778	Reisebus	71	333.63
1773	deutscher	241778	Torschütze	96	331.93
1773	deutscher	241778	Regierungsvertreter	73	328.2
1773	deutscher	241778	Härte	110	325.14
1773	deutscher	241778	Filmemacher	98	324.5
1773	deutscher	241778	Chemiker	80	321.55
1773	deutscher	241778	Popmusik	79	321.08
1773	deutscher	241778	Filmgeschichte	64	319.41
1773	deutscher	241778	Führung	375	319.06
1773	deutscher	241778	Kunstvereine	36	318.76
1773	deutscher	241778	Bundespräsident	145	315.26
1773	deutscher	241778	Eishockeymeister	31	314.35
1773	deutscher	241778	Aktien	365	310.63
1773	deutscher	241778	Hallenmeister	37	305.84

Wörterbucherweiterung

Aktualisierung eines Sachgruppenwörterbuchs

- Ziel: Einteilung des Wortschatzes der Alltagssprache in ca. 1000 Sachgruppen
- Gegeben: Letzte Auflage von 1959
- Aufgabe 1: Entfernen der veralteten Wörter und Wortgruppen
- Aufgabe 2: Finden der neu aufzunehmenden Wörter und Wortgruppen sowie deren Einordnung in vorhandene Gruppen (falls sinnvoll)
- Aufgabe 3: Einrichtung neuer Sachgruppen, Anpassung des Sachgruppensystems

Dornseiff Sachgruppen 10 + 21

Sachgruppe 10: **Fühlen, Affekt, Charaktereigenschaften**

U.a.: Charakter, Stolz, Eitelkeit, Menschenliebe, Härte, Hass, ...

Sachgruppe 21: **Recht, Ethik**

U.a.: Rechtschaffen, Selbstlos, Tugend, Frevel, Pflicht, ...

Sachgruppe 10

```
mysql> select distinct w1.word,w2.word, c.freq, c.sig, group_name from words w1, words w2, co_n c, r_word_group g, r_group_type t
     where w1.w_id=w1_id and w2.w_id=w2_id and w2.w_id=g.wort_nr and g.group_nr=t.group_nr and group_name like "%10.%" and
     wortart="N" and w1.word like "deutscher" and c.sig>9 order by group_name limit 300;
```

word	word	freq	sig	group_name
deutscher	Opfer	12	24	10.13 Unlust empfinden
deutscher	Philosoph	2	10	10.15 Zufriedenheit
deutscher	Designer	4	19	10.17 Geschmack, Kunstsinn
deutscher	Kultur	33	180	10.17 Geschmack, Kunstsinn
deutscher	Künstler	23	108	10.17 Geschmack, Kunstsinn
deutscher	Superstar	5	22	10.17 Geschmack, Kunstsinn
deutscher	**Humor**	**4**	**16**	**10.22 Witz**
deutscher	Satiriker	2	16	10.22 Witz
deutscher	Vertreter	44	238	10.23 Lächerlich
deutscher	**Spießer**	**2**	**17**	**10.25 Langeweile**
deutscher	Randalierer	4	23	10.28 Geschmacklos
deutscher	Musik	16	44	10.33 Tröstung
deutscher	Hilfe	59	287	10.33 Tröstung
deutscher	Sammler	4	21	10.35 Wunsch
...				
deutscher	**Innerlichkeit**	**2**	**21**	**10.7 Empfindlichkeit**
deutscher	Philosoph	2	10	10.8 Unempfindlichkeit, Seelenruhe

Sachgruppe 21

```
Use de;
mysql> select distinct w1.word,w2.word, c.freq, c.sig, group_name from words w1, words w2, co_n c, r_word_group g, r_group_type t
     where w1.w_id=w1_id and w2.w_id=w2_id and w2.w_id=g.wort_nr and g.group_nr=t.group_nr and group_name like "%21.%" and
     wortart="N" and w1.word like "deutscher" and c.sig>9 order by group_name limit 300;
```

word	word	freq	sig	group_name
deutscher	**Ehrlichkeit**	**2**	**10**	**21.1 Rechtschaffen**
deutscher	**Zucht**	**5**	**44**	**21.1 Rechtschaffen**
deutscher	**Zuverlässigkeit**	**2**	**10**	**21.1 Rechtschaffen**
deutscher	Schuld	10	41	21.11 Schuld, Vergehen
deutscher	Verbrechen	7	26	21.11 Schuld, Vergehen
deutscher	Täterschaft	2	18	21.11 Schuld, Vergehen
deutscher	Verteidiger	6	19	21.13 Rechtfertigung
deutscher	Anwalt	5	11	21.13 Rechtfertigung
deutscher	Agent	4	29	21.14 Vertrag
deutscher	Diplomat	30	281	21.14 Vertrag
deutscher	Diplomatie	3	15	21.14 Vertrag
deutscher	Unterhändler	11	90	21.14 Vertrag
deutscher	Kontrahent	2	13	21.14 Vertrag
deutscher	Rechtsauffassung	3	24	21.18 Recht, Gerechtigkeit
deutscher	Schuld	10	41	21.25 Pflichtverletzung
deutscher	Entführungsopfer	2	15	21.27 Kriminalität
deutscher	Verbrechen	7	26	21.27 Kriminalität
deutscher	Opfer	12	24	21.27 Kriminalität

...

38 rows in set (0.00 sec)

Allerletzte Folie:
Warum heißen Hunde „Corpus“?

Yukio Tono, Japan’s most prolific corpus linguist, gave a brilliant overview of recent and current developments in corpus-based language teaching. (…)

Though well-known in academic circles as a corpus linguist, Yukio Tono is something of a celebrity in Japan. He devised a long-running TV series in which each programme was devoted to one of the ‘100 English keywords’, with frequent references to ‘the corpus’ as the source of information on how these words behaved. The programme was watched by millions, and Tono himself appeared in it as the language guru. The result has been to popularize the term ‘corpus’, which is now so well known in Japan that there are, apparently, several dogs whose owners have called them ‘Corpus’. Whether this will catch on as a popular doggy name, only time will tell.

Quelle: http://www.macmillandictionaryblog.com/a-dog-named-corpus