

Tutorium der DGfS-Sektion Computerlinguistik

KORPUSLINGUISTIK

MIT ONLINE-

RESSOURCEN

EINE INTERAKTIVE

EINFÜHRUNG

FÜR LINGUISTEN

The screenshot shows a web-based corpus search interface. At the top, it says 'Home - CQP Mode - Simple Mode - Tools - Help Page' and 'Corpus: BUNDESTAG'. Below the navigation bar, there are search controls including a query field with the expression '<np_agr matches "Gen:*"> []+ [lemma="Gesetz"] </np_agr>', a sort dropdown set to 'ascending', and buttons for 'Run Query', 'Distribution', and 'Frequencies'. The main display area features a large, bold, black word 'LIPP' in the center. To the left and right of the word are smaller text snippets from the corpus, each with a line number and a context. For example, line 41.11 shows 'chim Großmann SPD)' and 'ag Beifall] [bei/SPD bei [No...'. Line 42.13 shows 'chim Großmann SPD)' and 'können [Nom / Pl / Wf] [Gen, Dat / S...'. Line 43.11 shows 'chim Großmann SPD)' and 'erh...'. Line 44.11 shows 'Exp' and 'hat, daß [Nom, Akk / Sg / ES] [L...'. Line 45.11 shows 'Auf [Nom, Akk / Sg, Pl] diese Weise]] wird [Nom, Akk / Sg die Grundintention] sowohl [Gen / Sg des Gesetzes] als auch [Gen / Sg des Einigungsvertrages] , nämlich [zu/Bildung zur [Nom, Gen, Dat, Akk / Sg Bildung [Gen / Sg selbstgenutzten Wohneigentums]]] [zu/Menschen für [Nom, Akk / Pl die Menschen]] [zu/Ostdeutschland in [Nom, Dat, Akk / Sg Ostdeutschland]] beizutragen , [zu/Eigentum] in [Nom, Akk / Sg das Gegenteil]] verkehrt .'. Line 46.12 shows 'Zentrale Änderung] ist der 1/2 3 [Gen / Sg des Gesetzes] .'. The interface also includes a 'Tokens' counter on the right side.

Linguistische Aufbereitung

Stefanie Dipper

Stefan Evert

Heike Zinsmeister

München, 28.01.2011

Linguistische Aufbereitung

Es war einmal eine kleine Hexe, die war erst einhundertsevenundzwanzig Jahre alt, und das ist ja für eine Hexe noch gar kein Alter. Sie wohnte in einem Hexenhaus, das stand einsam im tiefen Wald. Weil es nur einer kleinen Hexe gehörte, war auch das Hexenhaus nicht besonders groß. Der kleinen Hexe genügte es aber sie hätte sich gar kein schöneres Hexenhaus wünschen können ...

Ottfried Preussler, *Die kleine Hexe*

BNCweb Query result

/localhost/bncweb-cgi/main.pl?program=sort&textOrSpeaker=&theData=%5Bword%3D%22linguistics%22%5D

Search ▾ Info ▾ Urlaub ▾ TODO ▾ Apple ▾ Entertainment ▾ Comp ▾ FB WKW Acad PP ZX YouTube ▾ Vimeo ▾ TODO ▾ Yojimbo ▾ >>

" returned 784 hits in 100 different texts (98,313,429 words [4,048 texts]; frequency: 7.97 instances per million words), sorted by restriction any adjective (224 hits)

Show KWIC View Show in random order New Query Go!

Tag restriction: any adjective exclude Starting with letter: all Submit

Hits 201 to 224 Page 5 / 5

202	CGY 1200	generative grammar reflects the empirical nature of structural_AJO linguistics and instead uses linguistic intuitions of native speakers.
203	CGY 1183	The structuralism that Lévi-Strauss applies in cultural anthropology seems to be asking the question 'what does it mean?' — a very different problematic from that of structural_AJO linguistics .
204	H8V 441	But by integrating these theories with their version of structural_AJO linguistics , they developed them in a very important direction: they placed them within a semiotic or semiological framework, semiotics and semiology being to all intents and purposes interchangeable terms.
205	CGY 1271	But in his second phase the issues of systematicity and method become blurred and the connection with structural_AJO linguistics becomes
206	G1N 89	in structuralist_AJO linguistics is evident in her novels of the 1960s in the use of creative
207	CGY 14	work of both Lévi-Strauss and Barthes that emphasizes the influence of structuralist_AJO linguistics
208	CMR 1	JO linguistics view of language as a system of oppositions and differences.
209	CGY 1	regularity in synchronic_AJO linguistics is not a result of evolutionary continuity.
210	KAM 6	ere many in the U.K., by [gap:name] linguistic theory, then called 'scale and category grammar',
211	HH3 13	linguistics should help illuminate the human mind.
212	J7F 18	f language, and if the pragmatic, cognitive and procedural dimensions of language study which have
213	CM2 503	t_DT0-CJT linguistics (as an institution and a discipline) has to take these dimensions on board,

Go to "http://localhost/bncweb-cgi/fileinfo.pl?text=CM2&urlTest=yes"

Anforderungen / Wünsche

- Text durchsuchbar nach Wörtern, Phrasen, syntaktischen Mustern, ... ✓
- Wortarten (POS-Tagging) ✓
- Lemmatisierung ✓
- Morphosyntaktische Merkmale (✓)
- Syntaktische Analyse (Parsing) (✓)
- Textstruktur, Mehrwortausdrücke, ... ✗

Aufbereitungsschritte

1. Tokenisierung & Satzgrenzen
2. POS-Tagging (Wortarten)
 - regelbasiert (z.B. Brill 1995)
 - statistisch (Schmid 1995, Brants 2000)
3. Lemmatisierung (Zitierformen)
 - oft mit POS-Tagging integriert
4. Indexierung für effiziente Suche

I Tokenisierung

Um den linguistischen Reichtum zu beweisen, welchen ein inniger Kontakt mit der Natur und die Bedürfnisse des mühevollen Nomadenlebens haben hervorrufen können, erinnere ich an die Unzahl von charakteristischen Benennungen, durch die im Arabischen und Persischen Ebenen, Steppen und Wüsten unterschieden werden.

I Tokenisierung

<s> Um den linguistischen Reichtum zu **beweisen** , welchen ein inniger Kontakt mit der Natur und die Bedürfnisse des mühevollen Nomadenlebens haben hervorrufen **können** , erinnere ich an die Unzahl von charakteristischen **Benennungen** , durch die im Arabischen und Persischen **Ebenen** , Steppen . und Wüsten unterschieden **werden** . **</s>**



Wenn Tokenisierung
immer so einfach wäre ...

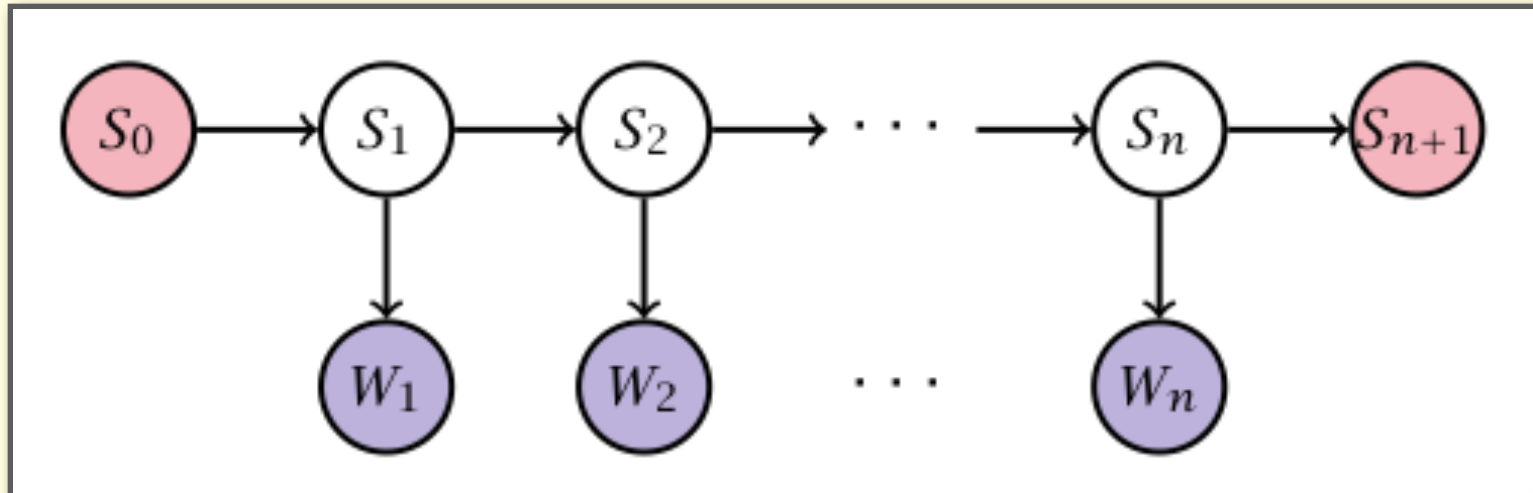
2 POS-Tagging

<s> Um den linguistischen Reichtum zu beweisen , welchen ein inniger Kontakt mit der Natur und die Bedürfnisse des mühevollen Nomadenlebens haben hervorrufen können , erinnere ich an die Unzahl von charakteristischen Benennungen , durch die im Arabischen und Persischen Ebenen , Steppen und Wüsten unterschieden werden . </s>

2 POS-Tagging

<s> Um/**KOUI** den/**ART** linguistischen/**ADJA** Reichtum/**NN** zu/**PTKZU** beweisen/**VVINF** ,/\$, welchen/**PRELS** ein/**ART** inniger/**ADJA** Kontakt/**NN** mit/**APPR** der/**ART** Natur/**NN** und/**KON** die/**ART** Bedürfnisse/**NN** des/**ART** mühevollen/**ADJA** Nomadenlebens/**NN** haben/**VAFIN** hervorrufen/**VVINF** können/**VMINF** ,/\$, erinnere/**VVFIN** ich/**PPER** an/**APPR** die/**ART** Unzahl/**NN** von/**APPR** charakteristischen/**ADJA** Benennungen/**NN** ,/\$, durch/**APPR** die/**PRELS** im/**APPRART** Arabischen/**NN** und/**KON** Persischen/**NN** Ebenen/**NN** ,/\$, Steppen/**NN** und/**KON** Wüsten/**NN** unterschieden/**VVPP** werden/**VAFIN** ./\$. </s>

POS-Tagging mit HMM



- HMM = Hidden Markov Model (generatives Modell)
- Church (1988), Schmid (1995), Brants (2000)
- Alternative: maschinelle Lernverfahren (SVM, MaxEnt, DT, NN, ...) → Tagging als Klassifikation

3 Lemmatisierung

<s> Um/**KOUI** den/**ART** linguistischen/**ADJA** Reichtum/**NN** zu/**PTKZU** beweisen/**VVINF** ,/\$, welchen/**PRELS** ein/**ART** inniger/**ADJA** Kontakt/**NN** mit/**APPR** der/**ART** Natur/**NN** und/**KON** die/**ART** Bedürfnisse/**NN** des/**ART** mühevollen/**ADJA** Nomadenlebens/**NN** haben/**VAFIN** hervorrufen/**VVINF** können/**VMINF** ,/\$, erinnere/**VVFIN** ich/**PPER** an/**APPR** die/**ART** Unzahl/**NN** von/**APPR** charakteristischen/**ADJA** Benennungen/**NN** ,/\$, durch/**APPR** die/**PRELS** im/**APPRART** Arabischen/**NN** und/**KON** Persischen/**NN** Ebenen/**NN** ,/\$, Steppen/**NN** und/**KON** Wüsten/**NN** unterschieden/**VVPP** werden/**VAFIN** ./\$. </s>

3 Lemmatisierung

<s> Um/**KOUI**/um den/**ART**/d linguistischen/**ADJA**/
linguistisch Reichtum/**NN**/Reichtum zu/**PTKZU**/zu
beweisen/**VVINF**/beweisen ,/\$,/, welchen/**PRELS**/welch
ein/**ART**/ein inniger/**ADJA**/innig Kontakt/**NN**/Kontakt mit/
APPR/mit der/**ART**/d Natur/**NN**/Natur und/**KON**/und die/
ART/d Bedürfnisse/**NN**/Bedürfnis des/**ART**/d mühevollen/
ADJA/mühevoll Nomadenlebens/**NN**/Nomadenleben
haben/**VAFIN**/haben hervorrufen/**VVINF**/hervorrufen
können/**VMINF**/können ,/\$,/, erinnere/**VVFIN**/erinnern
ich/**PPER**/ich an/**APPR**/an die/**ART**/d Unzahl/**NN**/Unzahl
von/**APPR**/von charakteristischen/**ADJA**/charakteristisch
Benennungen/**NN**/Benennung ,/\$,/, durch/**APPR**/durch
die/**PRELS**/d im/**APPRART**/im Arabischen/**NN**/Arabische

4 Indexierung

cpos	word	pos	lemma
0	Um	KOUI	um
1	den	ART	d
2	linguistischen	ADJA	linguistisch
3	Reichtum	NN	Reichtum
4	zu	PTKZU	zu
5	beweisen	VVINF	beweisen
6	,	\$,	,
7	welchen	PRELS	welch
...
46	.	\$.	.

