Tutorium der DGfS-Sektion Computerlinguistik

KORPUSLINGUISTIK MIT ONLINE-RESSOURCEN

EINE INTERAKTIVE EINFÜHRUNG FÜR LINGUISTEN

Linguistische Aufbereitung

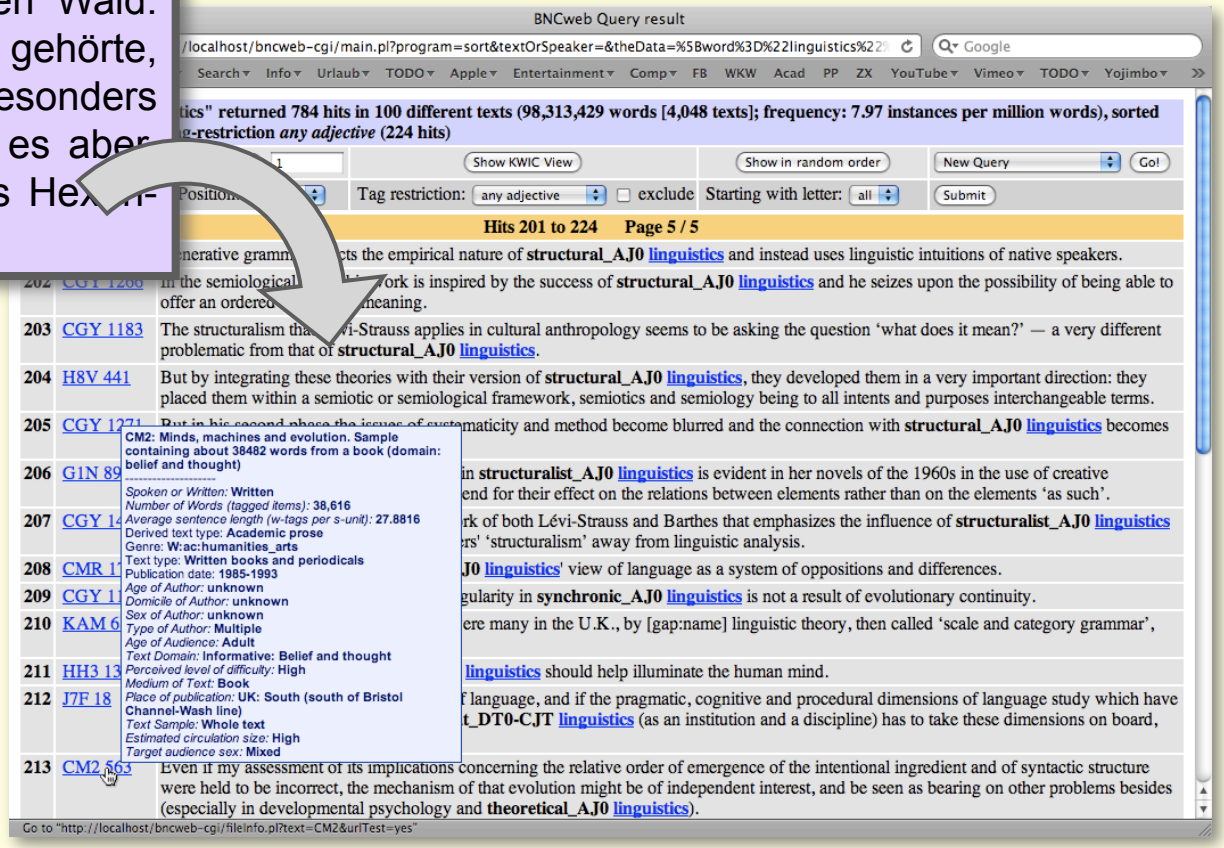
Stefanie Dipper
Stefan Evert
Heike Zinsmeister

München, 28.01.2011

Linguistische Aufbereitung

Es war einmal eine kleine Hexe, die war erst einhundertsiebenundzwanzig 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*



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 …

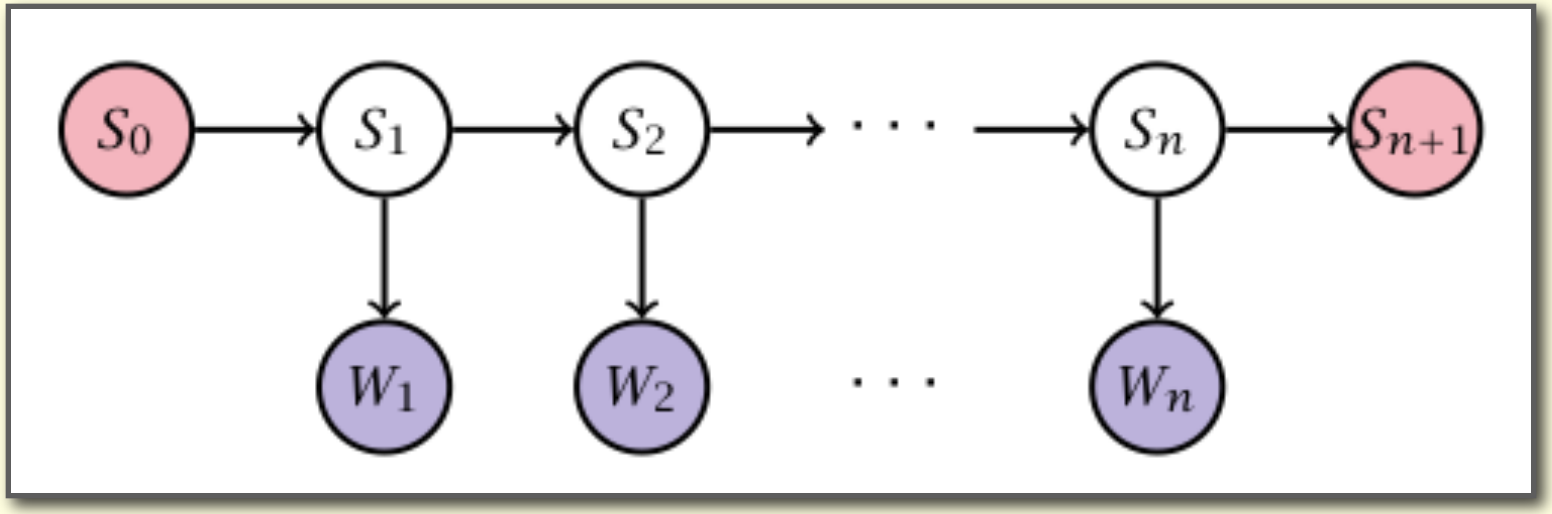
2 POS-Tagging

<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	.	$.	.