

What is a word?_{1/4}

- Morpheme: minimal linguistic unit with meaning. E.g. *nation*, *-al*, *-iz(e)*, *de-*, *-ation*.
- **Word** (difficult to characterize!): minimal **permutable** linguistic element with meaning. E.g. *denationalization*.
 - It can be moved in the sentence
 - Its position wrt other constituents can be altered by inserting new material.

What is a word?_{2/4}

- (1) The government is strongly opposed to denationalization.
- (2) a. What the government is strongly opposed to is denationalization.
b. It is denationalization that the government is opposed to.
- (3) The (present) government, (apparently), is (very) strongly (and implacably) opposed (not only) to (creeping) denationalization, but...

What is a word?_{3/4}

- **Word** (difficult to characterize!): minimal permutable linguistic element **with meaning**.
 - But a given phonological form may be ambiguous, in that it may express different meanings or senses.
 - If the two meanings are **unrelated**, we have two different words.
- (4) Engl: *bank*₁ (financial institution) vs. *bank*₂ (river)
- (5) Ge: *Schloss*₁ (‘castle’) vs. *Schloss*₂ (‘lock’)

⇒ Homonymy

What is a word?_{3/4}

- **Word** (difficult to characterize!): minimal permutable linguistic element **with meaning**.
 - But a given phonological form may be ambiguous, in that it may express different meanings or senses.
 - If the two meanings are **related**, we have one single word with a set of meanings or senses. E.g. *school*, *Schule*.
- (6) a. The school has a new director. **institution**
b. The school needs to be painted. **building**
c. The school of Plato had great consequences. **group of scholars**
d. School is fun. **type of activity**
⇒ **Polysemy**

QUESTION: Characterize the different senses of *school* in (6a-d).

Linguistische Annotationen - Beispiele

► Text

Er tritt in die GM-Verwaltung ein und wird Großaktionär des Autokonzerns .

Linguistische Annotationen - Beispiele

► Text + Lemmatisierung

Er tritt in die GM-Verwaltung ein und wird Großaktionär des Autokonzerns .
er treten in der GM-Verwaltung ein und werden Großaktionär der Autokonzern

Linguistische Annotationen - Beispiele

- ▶ Text + Lemmatisierung +
- ▶ Part-of-speech (POS) (Wortarten-Tagging)

Er	tritt	in	die	GM-Verwaltung	ein	und	wird	Großaktionär	des	Autokonzerns	.
er	treten	in	der	GM-Verwaltung	ein	und	werden	Großaktionär	der	Autokonzern	
PPER	VVFIN	APPR	ART	NN	PTKVZ	KON	VAFIN	NN	ART	NN	\$.

Linguistische Annotationen - Beispiele

- ▶ Text + Lemmatisierung +
- ▶ Part-of-speech (POS) (Wortarten-Tagging) +
- ▶ morphologische Information

Er	tritt	in	die	GM-Verwaltung	ein	und	wird	Großaktionär	des	Autokonzerns	.
er	treten	in	der	GM-Verwaltung	ein	und	werden	Großaktionär	der	Autokonzern	\$.
PPER	VVFIN	APPR	ART	NN	PTKVZ	KON	VAFIN	NN	ART	NN	
3.Nom.Sg.Masc	3.Sg.Pres.Ind		Acc.Sg.Fem	Acc.Sg.Fem			3.Sg.Pres.Ind	Nom.Sg.Masc	Gen.Sg.Masc	Gen.Sg.Masc	

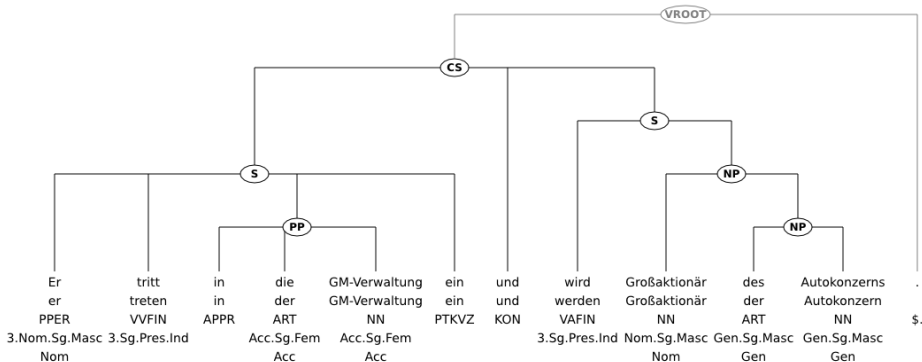
Linguistische Annotationen - Beispiele

- ▶ Text + Lemmatisierung +
- ▶ Part-of-speech (POS) (Wortarten-Tagging) +
- ▶ morphologische Information + Kasus

Er	tritt	in	die	GM-Verwaltung	ein	und	wird	Großaktionär	des	Autokonzerns	.
er	treten	in	der	GM-Verwaltung	ein	und	werden	Großaktionär	der	Autokonzern	\$.
PPER	VVFIN	APPR	ART	NN	PTKVZ	KON	VAFIN	NN	ART	NN	
3.Nom.Sg.Masc Nom	3.Sg.Pres.Ind		Acc.Sg.Fem Acc	Acc.Sg.Fem Acc			3.Sg.Pres.Ind	Nom.Sg.Masc Nom	Gen.Sg.Masc Gen	Gen.Sg.Masc Gen	

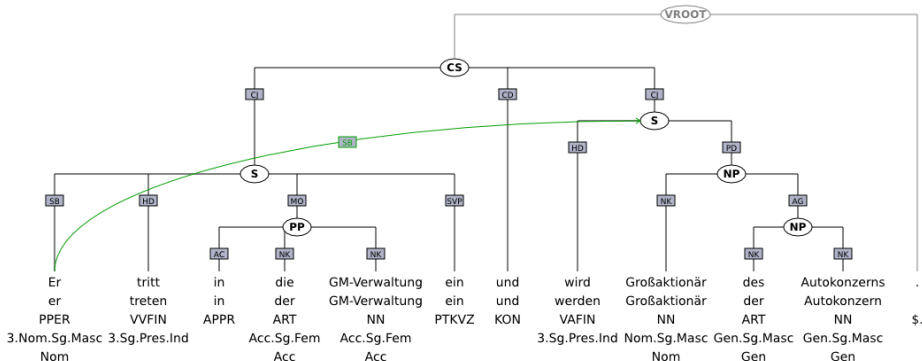
Linguistische Annotationen - Beispiele

- Text + Lemmatisierung +
- Part-of-speech (POS) (Wortarten-Tagging) +
- morphologische Information + Kasus + Syntax



Linguistische Annotationen - Beispiele

- ▶ Text + Lemmatisierung +
- ▶ Part-of-speech (POS) (Wortarten-Tagging) +
- ▶ morphologische Information + Kasus + Syntax +
- ▶ Grammatikalische Funktionen (GF) + sekundäre Kanten



STTS: Hintergrund

- ▶ Frühe 1990's: parallele Entwicklung von PoS-Tagsets für deutsch
 - ▶ in Tübingen (SfS)
 - ▶ und Stuttgart (IMS)
- ▶ Beide Versionen vereint
→ STTS (Stuttgart-Tübingen Tag Set), 1995/99

STTS: Hintergrund

- ▶ Frühe 1990's: parallele Entwicklung von PoS-Tagsets für deutsch
 - ▶ in Tübingen (SfS)
 - ▶ und Stuttgart (IMS)
- ▶ Beide Versionen vereint
→ STTS (Stuttgart-Tübingen Tag Set), 1995/99
- ▶ “(kleines) STTS”: POS (= default STTS)
- ▶ “großes STTS”: POS + morphology (nur Stuttgart)
- ▶ verwendet in NEGRA, TIGER, VERBMOBIL, DEREKO, DWDS,
...

STTS: Aufbau der Labels

- Tags spezifizieren Wortarten und tw. Flexion und Morphologie

@lacht VVFIN @ = Verb + voll + finit

STTS: Aufbau der Labels

- Tags spezifizieren Wortarten und tw. Flexion und Morphologie

@lacht VVFIN @ = Verb + voll + finit

- Ähnliches Prinzip in TIGER (treebank)

@OA @ = object + accusative

@OC @ = object + clausal

@OP @ = object + PP (vs. adjunct PP)

STTS: Aufbau der Labels

- ▶ Tags spezifizieren Wortarten und tw. Flexion und Morphologie
@lacht VVFIN @ = Verb + voll + finit
- ▶ Ähnliches Prinzip in TIGER (treebank)
@OA @ = object + accusative
@OC @ = object + clausal
@OP @ = object + PP (vs. adjunct PP)
- ▶ Systematische Namen sind leicht zu erfassen
- ▶ ermöglichen inkrementelle Annotation
- ▶ ermöglichen Unterspezifikation: VV (statt VVFIN, VVINFIN, ...)
- ▶ ermöglichen Anfragen als reguläre Ausdrücke: V.FIN

Aufbau der Labels

Wortart (N, V, ADJ, AP, P ...) plus:

- ▶ Lexikalische Information
 - ▶ Verben: V – M – A (full – modal – auxiliary)
 - ▶ Nomen: N – E (common – proper noun)
 - ▶ Pronomen: D – I – W – ... (demon. – indef. – interr.)

Aufbau der Labels

Wortart (N, V, ADJ, AP, P ...) plus:

- ▶ Lexikalische Information
 - ▶ Verben: V – M – A (full – modal – auxiliary)
 - ▶ Nomen: N – E (common – proper noun)
 - ▶ Pronomen: D – I – W – ... (demon. – indef. – interr.)
- ▶ Morpho-syntaktische Information
 - ▶ Verben: FIN – INF – IMP – PP
 - ▶ Adjektive: A – D (attributive vs. non-attributive)
 - ▶ Pronomen: S – AT (substantive/substituting vs. attributive)

Warum diese Information?

- ▶ Information leicht zu bestimmen
 - ▶ automatisch
e.g. ADJA vs. ADJD → flektiert
e.g. VA... für *jedes* Vorkommen *sein* und *haben*
 - ▶ manuell
e.g. ADJA vs. ADJD
- ▶ “Relevante” Information
 - ▶ e.g. VVFIN vs. VVIN

STTS Dokumentation

- ▶ Tagset:
<http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html>
- ▶ Guidelines und Wortlisten:
<http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html>

Andere Tagsets

- ▶ Auch andere Wortartenaufteilungen möglich
- ▶ POS Tagsets unterscheiden sich in der Anzahl der angenommenen Wortarten (feinkörnig vs. grobkörnig), z.B.:
 - ▶ Penn Treebank POS Tagset: 45 tags
 - ▶ kleines STTS-Tagset: 54 tags
 - ▶ Brown Corpus: 87 tags
 - ▶ Lancaster-Oslo/Bergen (LOB) Corpus: 135 tags
 - ▶ BNC Enriched Tagset (C7): 140 tags
 - ▶ London-Lund Corpus of Spoken English 197 tags

BNC Enriched Tagset (Auszug (1))

APPG	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
AT1	singular article (e.g. a, an, every)
CS	Subordinating conjunction, general (e.g. if, when, while, because)
DD	Central determiner, neutral for number (e.g. some, any, enough)
EX	existential there
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker - (' or's)
IF	for (as preposition)
II	general preposition
IO	of (as preposition)
IW	with, without (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. older, better, stronger)
JJT	general superlative adjective (e.g. oldest, best, strongest)
JK	catenative adjective (able in be able to, willing in be willing to)

BNC Enriched Tagset (Auszug (2))

ND1	singular noun of direction (e.g. north, southeast)
NN	common noun, neutral for number (e.g. sheep, cod)
NN1	singular common noun (e.g. book, girl)
NN2	plural common noun (e.g. books, girls)
NNA	following noun of title (e.g. M.A.)
NNB	preceding noun of title (e.g. Mr., Prof.)
NNL1	singular locative noun (e.g. Island, Street)
NNL2	plural locative noun (e.g. Islands, Streets)
NNO	numeral noun, neutral for number (e.g. dozen, hundred)
NNO2	numeral noun, plural (e.g. hundreds, thousands)
NNT1	temporal noun, singular (e.g. day, week, year)
NNT2	temporal noun, plural (e.g. days, weeks, years)
NNU	unit of measurement, neutral for number (e.g. in, cc)
NNU1	singular unit of measurement (e.g. inch, centimetre)
NNU2	plural unit of measurement (e.g. ins., feet)
RR	general adverb

BNC: Mehrwortlexeme

- ▶ Markierung von Mehrwortlexemen mit übergeordnetem Tag
- ▶ Jedes Token
 - < *Tag* > < *Anzahl-Token* > < *Nr_aktuelles-Token* >
- ▶ Beispiele:
 - ▶ all/RR41 of/RR42 a/RR43 sudden/RR44
 - ▶ in/II31 terms/II32 of/II33
 - ▶ at/RR21 length/RR22
 - ▶ a/DD21/RR21 lot/DD22/RR22
 - ▶ in/CS21/II that/CS22/DD1

Cosmas II

Über Cosmas II

Abmeldung

Recherche

Optionen

Archiv

Korpus

Suchanfrage

Wortformliste

Aktuelles Archiv

W – Archiv der geschriebenen Sprache

Aktuelles Korpus

W-öffentlich – alle öffentlichen Korpora des Archivs W

Suchanfrage

downloadet

Quellenansicht

Release: Deutsches Referenzkorpus (DeReKo-2012-II)

Ergel

Gesamt-KWIC

Gesamt-Volltext

Export

Kookkurrenzanalyse

Ergebnisse

Quellenansicht

Korpusansicht

Dokumentansicht

Ansicht vor/seit

Zeitpunkt

Jahrzehntansicht

Jahresansicht

Monatsansicht

Tagesansicht

Länderansicht

Textsortenansicht

Themenansicht

Stat. KWIC-Ausw.

Kookkurrenzanalyse


Gesamt-KWIC

Gesamt-Volltext

Export

Treffer	Texte	von	bis	Quelle
+	2	2	2007	2007 Braunschweiger Zeitung
+	1	1	2012	2012 Burgenländische Volkszeitung
+	2	2	1995	1996 COMPUTER ZEITUNG
+	1	1	2011	2011 Die Südschweiz
+	1	1	1998	1998 Frankfurter Rundschau
+	1	1	2011	2011 Hamburger Morgenpost
+	3	3	2004	2007 Mannheimer Morgen
+	1	1	2000	2000 Neue Kronen-Zeitung
+	11	11	2007	2012 Niederösterreichische Nachrichten
+	1	1	2010	2010 Rhein-Zeitung
+	1	1	2000	2000 Salzburger Nachrichten
+	1	1	1999	1999 Tiroler Tageszeitung
+	1	1	2008	2008 VDI nachrichten
+	2	2	1997	2000 Vorarlberger Nachrichten
+	2	1	2011	2011 Wikipedia.de 2011 Artikel
+	23	16	2006	2011 Wikipedia.de 2011 Diskussionen
+	1	1	2000	2000 Zürcher Tagesanzeiger
	55	47	1995	2012 17 Quellen

© 2003 - 2012 IDS Mannheim, COSMAS II



Tatjana Scheffler (Uni Potsdam)

Einführung in die Korpuslinguistik

29. April 2013

23 / 37

Über Cosmas II

Abmeldung

Recherche

Optionen

Archiv

Korpus

Suchanfrage

Wortformliste

Ergebnisse

→ Quellenansicht

Korpusansicht
 Dokumentansicht
 Ansicht vor/seit
 Zeitpunkt
 Jahrzehntansicht
 Jahresansicht
 Monatsansicht
 Tagesansicht
 Länderansicht
 Textsortenansicht
 Themenansicht
 Stat. KWIC-Ausw.

Kookkurrenzanalyse

Gesamt-KWIC

Gesamt-Volltext

Export

Aktuelles Archiv

W - Archiv der geschriebenen Sprache

Aktuelles Korpus

W-offentlich - alle öffentlichen Korpora des Archivs W

Suchanfrage










gedownloadet

Quellenansicht

Release: Deutsches Referenzkorpus (DeReKo-2012-II)

Ergebnisse

 Gesamt-KWIC
  Gesamt-Volltext
  Export
  Kookkurrenzanalyse

Treffer	Texte	von	bis	Quelle
	2	2	2005 2012	Die Südschweiz
	2	2	2006 2006	Hamburger Morgenpost
	1	1	2008 2008	Hannoversche Allgemeine
	1	1	2011 2011	Mannheimer Morgen
	1	1	2008 2008	Nürnberger Nachrichten
	1	1	2000 2000	Salzburger Nachrichten
	1	1	2010 2010	St. Galler Tagblatt
	2	2	2011 2011	Wikipedia.de 2011 Artikel
	21	12	2007 2011	Wikipedia.de 2011 Diskussionen
32	23	2000	2012	9 Quellen

Cosmas II: Organisation der Korpora

Archive

- ▶ W – geschriebene Korpora
- ▶ TAGGED-T – mit TreeTagger getaggte Korpora
- ▶ etc.

Cosmas II: Organisation der Korpora

Archive

- Korpora

- Dokumente (z.B. ein Monat einer Zeitung)

- Texte

Cosmas II: Suchanfragen

Registrierung

- ▶ kostenfrei über eine Emailadresse

Suche

- ▶ Suche nach Wörtern, Wortteilen, Lemmata
- ▶ Kombinationen, Abstandsoperatoren

Ausgabe:

- ▶ Belege, sortiert nach Quelle
- ▶ Frequenzen
- ▶ Konkordanzen
- ▶ Kollokationen
- ▶ Export von Belegen möglich

Cosmas II: Anfragesyntax

- ▶ zeilenbasiert, ähnlich den regulären Ausdrücken

Willkommen	Wortform <i>Willkommen</i>
Mond*	<i>Mond, Mondes, Mondenschein, ...</i>
?in	genau ein Zeichen, <i>bin, ein</i>
+in	0 oder 1 Zeichen, <i>bin, ein, in</i>
&Mond	Lemma
&be-	Präfix <i>be-</i>
&-chen	Suffix <i>-chen</i>
runder /+w1 Tisch	Wortabstand max 1 nach rechts (+)
anscheinend /s0 scheinbar	im selben Satz

www.ids-mannheim.de/cosmas2/web-app/hilfe/suchanfrage/