EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

PHILOSOPHISCHE
FAKULTÄT

Seminar für Sprachwissenschaft

Wortartentagging der Tübinger Ressourcen nach STTS

Erfahrungen mit verschiedenen Textgenres

STTS-Workshop Stuttgart
24.09.2012, Kathrin Beck, Erhard Hinrichs, Heike Telljohann, Yannick Versley

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

STTS-getaggte Tübinger Ressourcen (1)

- **POS-Tags manuell bearbeitet**
  - TüBa-D/S (Tübinger Baumbank des Deutschen / Spontansprache)
    - ca. 38 000 Sätze, Quelle: Verbmobil-Dialoge
  - TüBa-D/Z (Tübinger Baumbank des Deutschen / Zeitungskorpus)
    - ca. 65 500 Sätze, Quelle: die tageszeitung (taz)
  - Sample aus TüBa-D/DC (Diachrones Korpus)
    - ca. 3 800 Sätze aus insg. 6 Texten unterschiedlicher Epochen
    - Quelle: Ausgewähltes Material aus dem Projekt Gutenberg-DE
    - Nur intern verwendet
      (zur Evaluation der automatischen Annotation in TüBa-D/DC)

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

STTS-getaggte Tübinger Ressourcen (2)

- **POS-Tags automatisch annotiert**
  - TüPP-D/Z (Tüb. Partiell Geparstes Korpus des Dt. / Zeitungskorpus)
    - > 200 Mio. Wörter, Quelle: Textsammlung aus die tageszeitung (taz)
    - Kombination aus tnt-Modellen für verschiedene Genres
  - web-news
    - 1,7 Mia. Wörter, Quelle: Nachrichten- und Blogsites im WWW
    - RFTagger + Konversion nach STTS+morph
  - TüBa-D/DC (Tübinger Baumbank des Deutschen/Diachrones Korpus)
    - > 250 Mio. Wörter, Quelle: Projekt Gutenberg-DE
    - TreeTagger

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

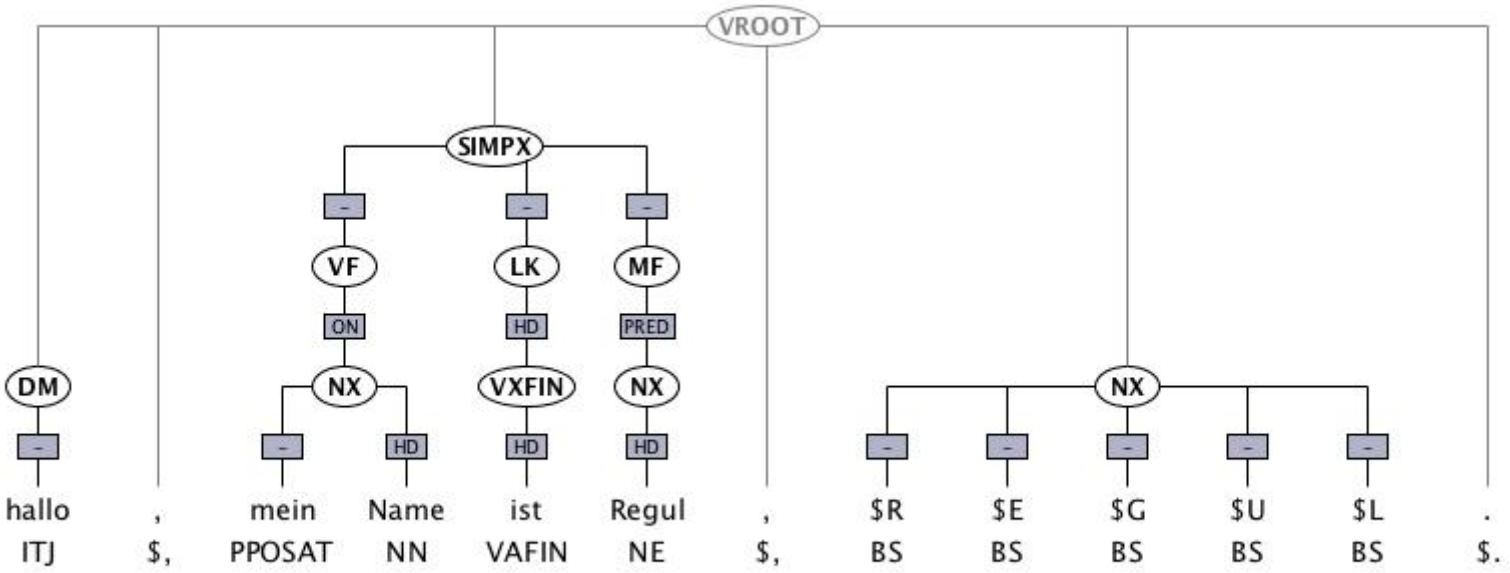
Abweichungen vom originalen STTS-Tagset in den manuell bearbeiteten Korpora

- **TüBa-D/S**
  - zusätzlich: BS (Buchstabe)
  - umbenannt: PAV (Pronominaladverb, z. B. *davon*, *deswegen*, *hierfür*)
    → PROP (pronominale Form einer PP)

- **TüBa-D/Z**
  - umbenannt: PAV → PROP (pronominale Form einer PP)

- **TIGER Corpus**
  - umbenannt: PAV → PROAV (Pronominaladverb)
  - nicht verwendet: PIDAT (attr. Indefinitpronomen m. Determiner)

Baumbeispiel aus der TüBa-D/S



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Anwendungsunterschiede von POS-Tags (Beispiele)

- **PIAT vs. PIDAT**
  - TIGER Corpus Bsp.: **viele (PIAT)** Menschen
    mehr (PIAT) Fragen
  - TüBa-D/Z Bsp: **viele (PIDAT)** Menschen
    mehr (PIAT) Fragen
- ***als* APPR vs. KOKOM**
  - TIGER Corpus Bsp.: *Powell galt nach Umfragen* ***als (APPR)*** *aussichtsreichster Gegner Clintons.*
  - TüBa-D/Z Bsp.: *Cuomo gilt* ***als (KOKOM)*** *exzellenter Redner.*

Anwendungsunterschiede von POS-Tags (Beispiele)

- **FM vs. NE am Beispiel von fremdsprachlichen Filmtiteln**

  - TIGER Corpus
    Bsp.: *Tous* (NE) *les* (NE) *matins* (NE) *du* (NE) *monde* (NE).

  - TüBa-D/Z
    Bsp.: *Knockin‘* (FM) *on* (FM) *Heaven‘s* (FM) *Door* (FM).

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Beispiel 1 für POS-Tag-Korrektur in TüBa-D/DC:

- Johann Wolfgang von Goethe, Die Leiden des jungen Werther (1774)

TEXT	TOKENS	POSTAGS
t_504	Das	PDS
t_505	bewog	VVFIN
t_506	den	ART
t_507	verstorbenen	ADJA
t_508	Grafen	NN
t_509	von	APPR
t_510	M.	NE
t_511	,	$,
t_512	einen	ART
t_513	Garten	NN
t_514	auf	APPR
t_515	einem	PIS
t_516	der	ART
t_517	Hügel	NN
t_518	anzulegen	VVIZU
t_519	,	$,
t_520	die	PRELS
t_521	mit	APPR
t_522	der	ART
t_523	schönsten	ADJA
t_524	Mannigfaltigkeit	NN
t_525	sich	PRF
t_526	kreuzen	VVINF
t_527	und	KON
t_528	die	ART
t_529	lieblichsten	ADJA
t_530	Täler	NN
t_531	bilden	VVINF
t_532	.	$.

POSTAGS
PDS
VVFIN
ART
ADJA
NN
APPR
NE
$,
ART
NN
APPR
PIS
ART
NN
VVIZU
$,
PRELS
APPR
ART
ADJA
NN
PRF
YYY-VVFIN
KON
ART
ADJA
NN
YYY-VVFIN
$.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Beispiel 2 für POS-Tag-Korrektur in TüBa-D/DC:

- Philipp Melanchthon, Die Augsburgische Konfession (1530)

TEXT	TOKENS	POSTAGS
t_1321	Hie	NE
t_1322	werden	VAFIN
t_1323	verworfen	VVPP
t_1324	die	PRELS
t_1325	,	$,
t_1326	so	ADV
t_1327	lehren	VVINF
t_1328	,	$,
t_1329	daß	KOUS
t_1330	diejenigen	PDS
t_1331	,	$,
t_1332	so	ADV
t_1333	einst	ADV
t_1334	seind	ADV
t_1335	fromm	ADJD
t_1336	worden	VAPP
t_1337	,	$,
t_1338	nicht	PTKNEG
t_1339	wieder	ADV
t_1340	fallen	VVINF
t_1341	mugen	VVINF
t_1342	.	$.

POSTAGS
YYY-ADV
VAFIN
VVPP
YYY-PDS
$,
YYY-PRELS
YYY-VVFIN
$,
KOUS
PDS
$,
YYY-PRELS
ADV
YYY-VAFIN
ADJD
VAPP
$,
PTKNEG
ADV
VVINF
YYY-VMFIN
$.

Beispiel 3 für POS-Tag-Korrektur in TüBa-D/DC:

- Gottfried von Straßburg, Tristan (1210)

TEXT	TOKENS	POSTAGS	POSTAGS
t_280	Tiure	NE	YYY-ADJD
t_281	unde	VVFIN	YYY-KON
t_282	wert	ADJD	ADJD
t_283	ist	VAFIN	VAFIN
t_284	mir	PPER	PPER
t_285	der	ART	ART
t_286	man	PIS	YYY-NN
t_287	,	$,	$,
t_288	der	PRELS	PRELS
t_289	guot	ADJD	YYY-NN
t_290	und	KON	KON
t_291	übel	ADJD	YYY-NN
t_292	betrahten	ADJA	YYY-VVINF
t_293	kan	NE	YYY-VMFIN
t_294	,	$,	$,
t_295	der	PRELS	PRELS
t_296	mich	PPER	PPER
t_297	und	KON	KON
t_298	iegelîchen	VVFIN	YYY-PIDAT
t_299	man	PIS	YYY-NN
t_300	nâch	NE	YYY-APPR
t_301	sînem	ADJA	YYY-PPOSAT
t_302	werde	VAFIN	YYY-NN
t_303	erkennen	VVINF	VVINF
t_304	kan	XY	YYY-VMFIN
t_305	.	$.	$.

Erweiterte STTS-Tagsets im PCFG-Parsing

- PCFG-Parsing braucht feinere Unterscheidung von Wortarten und Knoten (Schiehlen, 2004; Dubey, 2005; Versley 2005; Versley&Rehbein 2009)
- Unterschiedliche Verfeinerungen des STTS-Tagsets

<table>
<tr><th>STTS</th><th>Sch04</th><th>Dub05</th><th>Ver05</th><th>VR09</th></tr>
<tr><td>DET</td><td>DET</td><td>DET_SB</td><td>DET_nsm</td><td>DET_der</td></tr>
<tr><td>ADJA</td><td>ADJA</td><td>ADJA_SB</td><td colspan="2">ADJA_e</td></tr>
<tr><td>NN</td><td colspan="2">NN_SB</td><td colspan="2">NN_sm</td></tr>
<tr><td>VAFIN</td><td colspan="4">VAFIN_haben</td></tr>
<tr><td>VVFIN</td><td>VVFIN_OA_OC</td><td>VVFIN</td><td colspan="2">VVFIN_a</td></tr>
<tr><td>APPR</td><td>APPR</td><td>APPR_OA</td><td colspan="2">APPR_a</td></tr>
</table>

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Danke fürs Zuhören.