Swantje Westpfahl & Thomas Schmidt

POS für(s) FOLK

Problemanalyse des POS-Taggings für spontansprachliche Daten anhand des Forschungs- und Lehrkorpus Gesprochenes Deutsch

FOLK

- Forschungs- und Lehrkorpus für gesprochenes Deutsch
- Seit 2008 am Archiv für Gesprochenes Deutsch (AGD)
- Ziele:
  - Großes, computergestützt auswertbares, wissenschaftsöffentlich verfügbares Korpus authentischer Gespräche
  - Erarbeitung und Etablierung einer Best Practice für die Erstellung mündlicher Korpora

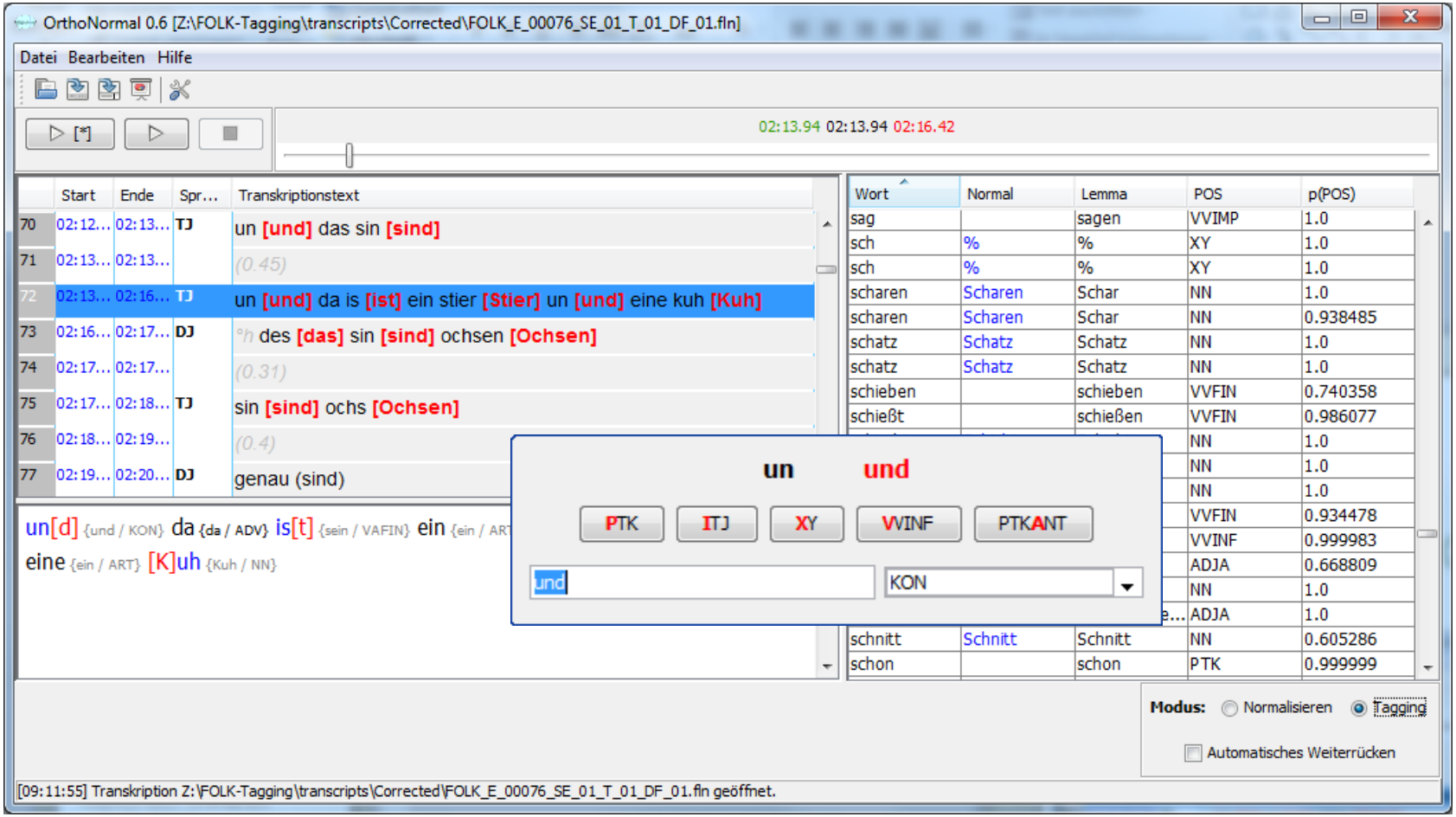
FOLK

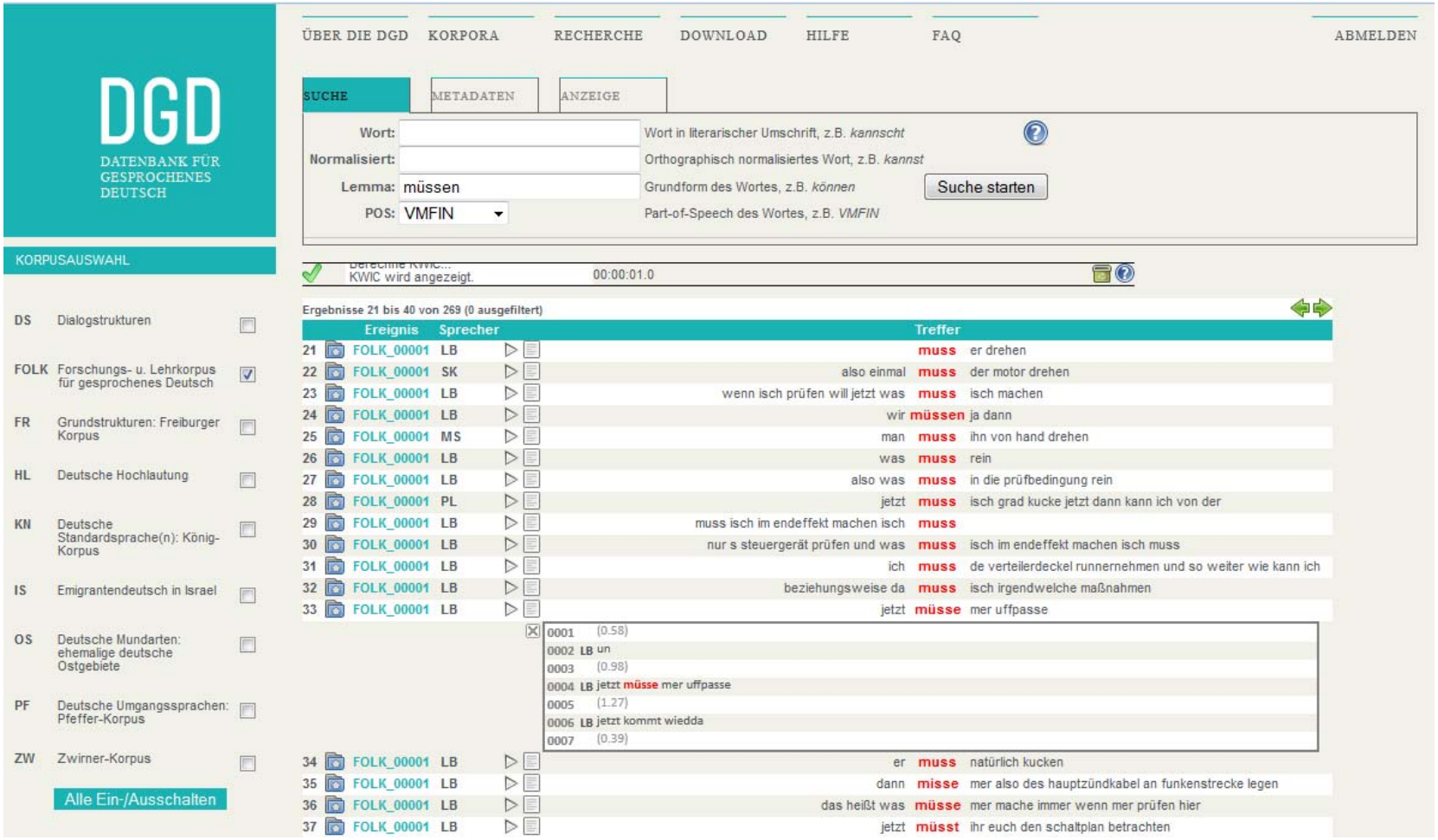
- Pilotveröffentlichung im Februar 2012 über die Datenbank Gesprochenes Deutsch (DGD2)
- Nächste Version: November 2012
  - ca. 75h Aufnahmen
  - ca. 800.000 Wort-Tokens
- Kontinuierlicher Ausbau in den nächsten Jahren

http://dgd.ids-mannheim.de
http://agd.ids-mannheim.de/folk.shtml

FOLK Workflow

- Audioaufnahmen → Maskierung
- Transkription nach GAT2 (Minimaltranskript) mit FOLKER
- Orthographische Normalisierung mit OrthoNormal (automatisch mit anschließender manueller Korrektur)
- Tagging und Lemmatisierung mit TreeTagger (Standard-Parameter für Deutsch, TT4J)
- **Manuelle Korrektur des Tagging**
- Veröffentlichung in der DGD2


OrthoNormal 0.6 [Z:\FOLK-Tagging\transcripts\Corrected\FOLK_E_00076_SE_01_T_01_DF_01.fln]
Datei Bearbeiten Hilfe
02:13.94 02:13.94 02:16.42
Start Ende Spr... Transkriptionstext
70 02:12... 02:13... TJ un [und] das sin [sind]
71 02:13... 02:13... (0.45)
72 02:13... 02:16... TJ un [und] da is [ist] ein stier [Stier] un [und] eine kuh [Kuh]
73 02:16... 02:17... DJ °h des [das] sin [sind] ochsen [Ochsen]
74 02:17... 02:17... (0.31)
75 02:17... 02:18... TJ sin [sind] ochs [Ochsen]
76 02:18... 02:19... (0.4)
77 02:19... 02:20... DJ genau (sind)
un[d] {und / KON} da {da / ADV} is[t] {sein / VAFIN} ein {ein / ART}
eine {ein / ART} [K]uh {Kuh / NN}
un und
PTK ITJ XY VVINF PTKANT
und
KON
Wort Normal Lemma POS p(POS)
sag sagen VVIMP 1.0
sch % % XY 1.0
sch % % XY 1.0
scharen Scharen Schar NN 1.0
scharen Scharen Schar NN 0.938485
schatz Schatz Schatz NN 1.0
schatz Schatz Schatz NN 1.0
schieben schieben VVFIN 0.740358
schießt schießen VVFIN 0.986077
NN 1.0
NN 1.0
NN 1.0
VVFIN 0.934478
VVINF 0.999983
ADJA 0.668809
NN 1.0
ADJA 1.0
schnitt Schnitt Schnitt NN 0.605286
schon schon PTK 0.999999
Modus: Normalisieren Tagging
Automatisches Weiterrücken
[09:11:55] Transkription Z:\FOLK-Tagging\transcripts\Corrected\FOLK_E_00076_SE_01_T_01_DF_01.fln geöffnet.


DGD
DATENBANK FÜR GESPROCHENES DEUTSCH
ÜBER DIE DGD
KORPORA
RECHERCHE
DOWNLOAD
HILFE
FAQ
ABMELDEN
SUCHE
METADATEN
ANZEIGE
Wort:
Wort in literarischer Umschrift, z.B. kannscht
Normalisiert:
Orthographisch normalisiertes Wort, z.B. kannst
Lemma: müssen
Grundform des Wortes, z.B. können
Suche starten
POS: VMFIN
Part-of-Speech des Wortes, z.B. VMFIN
KWIC wird angezeigt.
00:00:01.0
KORPUSAUSWAHL
DS Dialogstrukturen
FOLK Forschungs- u. Lehrkorpus für gesprochenes Deutsch
FR Grundstrukturen: Freiburger Korpus
HL Deutsche Hochlautung
KN Deutsche Standardsprache(n): König-Korpus
IS Emigrantendeutsch in Israel
OS Deutsche Mundarten: ehemalige deutsche Ostgebiete
PF Deutsche Umgangssprachen: Pfeffer-Korpus
ZW Zwirner-Korpus
Alle Ein-/Ausschalten
Ergebnisse 21 bis 40 von 269 (0 ausgefiltert)
Ereignis Sprecher Treffer
21 FOLK_00001 LB muss er drehen
22 FOLK_00001 SK also einmal muss der motor drehen
23 FOLK_00001 LB wenn isch prüfen will jetzt was muss isch machen
24 FOLK_00001 LB wir müssen ja dann
25 FOLK_00001 MS man muss ihn von hand drehen
26 FOLK_00001 LB was muss rein
27 FOLK_00001 LB also was muss in die prüfbedingung rein
28 FOLK_00001 PL jetzt muss isch grad kucke jetzt dann kann ich von der
29 FOLK_00001 LB muss isch im endeffekt machen isch muss
30 FOLK_00001 LB nur s steuergerät prüfen und was muss isch im endeffekt machen isch muss
31 FOLK_00001 LB ich muss de verteilerdeckel runnernehmen und so weiter wie kann ich
32 FOLK_00001 LB beziehungsweise da muss isch irgendwelche maßnahmen
33 FOLK_00001 LB jetzt müsse mer uffpasse
0001 (0.58)
0002 LB un
0003 (0.98)
0004 LB jetzt müsse mer uffpasse
0005 (1.27)
0006 LB jetzt kommt wiedda
0007 (0.39)
34 FOLK_00001 LB er muss natürlich kucken
35 FOLK_00001 LB dann misse mer also des hauptzündkabel an funkenstrecke legen
36 FOLK_00001 LB das heißt was müsse mer mache immer wenn mer prüfen hier
37 FOLK_00001 LB jetzt müsst ihr euch den schaltplan betrachten

Methodik

- Automatisiertes POS-Tagging mit dem Treetagger und dem STTS
- Manuelle Korrektur des Taggings
- Auswertung der Korrektur
- Analyse der Probleme des POS-Taggens von gesprochener Sprache
- Entwickeln von Ansätzen für Lösungen der Probleme

Manuelle Korrektur

- Grundsätzlich wird der Klassifizierung der Duden Grammatik gefolgt
- Problem: Zugehörigkeit zu Wortarten ist oft Interpretationssache z. B. ist „hoch“ in „Hand hoch“ Ellipse von „hochheben“, also PTKVZ oder Adverb?
- Fehler in der Normalisierung bzw. im Normalisierungsprozess z.B. Vergessen der Großschreibung etc.

Auswertung der Manuellen Korrektur

	Transkript 1	Transkript 2	Transkript 3
Wörter insgesamt	3976	5033	2020
Richtig	3229	4096	1626
Richtig in %	81,21	81,38	80,5
Korrigiert	747	937	394
Korrigiert in %	18,79	18,62	19,5
Richtig (Superkategorie)	3381	4295	1702
Richtig in % (Superkategorie)	85,04	85,34	84,26
Korrigiert (Superkategorie)	595	738	318
Korrigiert in % (Superkategorie)	14,96	14,66	15,74

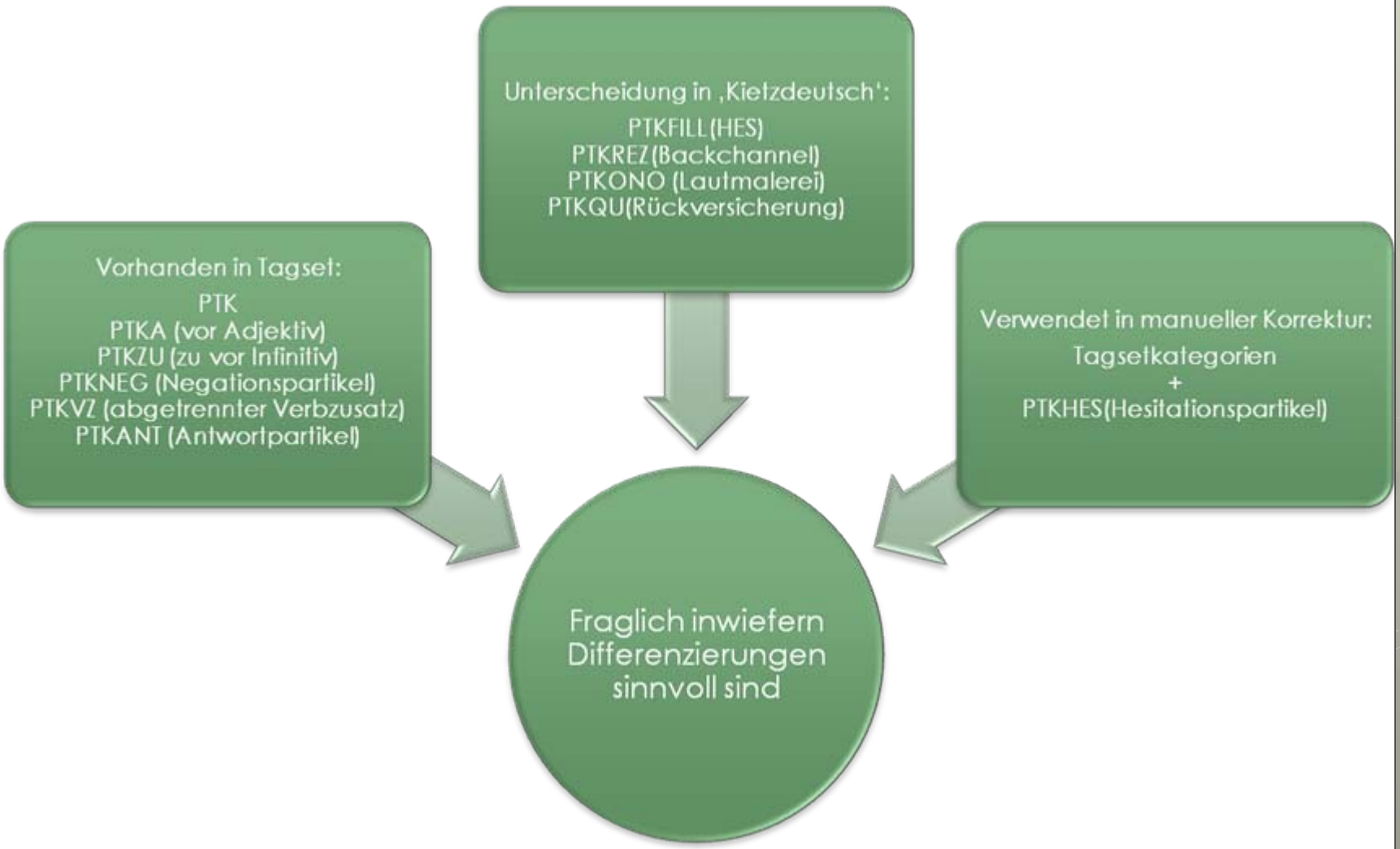
Auswertung der Korrektur

- Fehlerquote bei allen drei Gesprächstypen in etwa gleich
- Auffällig viele Partikeln in den „Top 10“ der Fehlerkorrekturen
- Relativ hohe Fehlerquote bei Verben, XY-Kategorien und Pronomen

Probleme des POS-Taggens von FOLK

- Probleme in Hinsicht auf systembedingte Grenzen
  - Probleme die durch FOLK-Workflow entstehen
  - Probleme die sich in Hinsicht auf Unterschiede zwischen geschriebener und gesprochener Sprache ergeben
- **Probleme in Hinsicht auf die Verwendung des STTS für gesprochener Sprache**

Partikeln


Unterscheidung in ‚Kietzdeutsch':
PTKFILL(HES)
PTKREZ(Backchannel)
PTKONO (Lautmalerei)
PTKQU(Rückversicherung)
Vorhanden in Tagset:
PTK
PTKA (vor Adjektiv)
PTKZU (zu vor Infinitiv)
PTKNEG (Negationspartikel)
PTKVZ (abgetrennter Verbzusatz)
PTKANT (Antwortpartikel)
Verwendet in manueller Korrektur:
Tagsetkategorien
+
PTKHES(Hesitationspartikel)
Fraglich inwiefern Differenzierungen sinnvoll sind

Kategorie Interjektion

- Laut Duden Unterkategorie der Partikeln
- Im STTS eigene Kategorie
- Kategorie nicht in Liste geschlossener Wortarten

Kategorie XY (Nichtwörter)

- Alles, was nicht erkannt wird, bekommt Tag XY
  - Buchstabiertes
  - Onomatopoetika
  - Unverständliches -> eigene Kategorie
  - Abbrüche -> eigene Kategorie

Kategorisierung der Verben

VVFIN	finites Verb, voll
VVIMP	Imperativ, voll
VVINF	Infinitiv, voll
VVIZU	Infinitiv mit *zu*, voll
VVPP	Partizip Perfekt, voll
VAFIN	finites Verb, aux
VAIMP	Imperativ, aux
VAINF	Infinitiv, aux
VAPP	Partizip Perfekt, aux
VMFIN	finites Verb, modal
VMINF	Infinitiv, modal
VMPP	Partizip Perfekt, modal

- Warum gibt es Kategorie für Imperativ, nicht aber für Indikativ bzw. Konjunktiv?
- Warum gibt es keine Infinitiv mit zu – Kategorie bei Modalverben oder Auxiliaren?

Liste der Wortformen geschlossener Wortarten

- Veraltet 1996 → alte Rechtschreibung
- Lückenhaft (bzw. unvollständig)
- Fehlerhaft (falsch Kategorisiertes)

Pronominaladverbien

- Pronominaladverbien in Liste der Wortformen geschlossener Wortarten des STTS unvollständig (s. Duden 660)
- „deshalb“ und „trotzdem“ sind in Liste als Pronominaladverbien PAV eingeordnet
- Fraglich inwiefern Unterscheidung zu regulären Adverbien notwendig ist

Konjunktionen

- „sondern“, „trotzdem“, „wo“ und „außer“ fehlen in der Liste bei den Konjunktionen
- Wenn es Kategorie für Pronominaladverbien gibt, müsste es dann nicht auch eine für Konjunktionaladverbien geben?
- „d.h.“, „bzw.“ und „z.B.“ sind nur als Abkürzungen bei den Konjunktionen aufgeführt -> Ausgeschriebene werden nicht als solche erkannt. Fraglich, ob sie wirklich zu den Konjunktionen gehören

Sonstige Unvollständigkeiten

- „selber“ fehlt bei Demonstrativpronomen PDS
- „wo“ fehlt bei dialektalem Gebrauch als substituierende Relativpronomen PRELS „die wo…“
- „irgendetwas“ und „irgendwas“ fehlt in Kategorie der attribuierenden Indefinitpronomen PIAT

Ansätze für Lösungen der Probleme

- Innerhalb des FOLK Projekts -> Workflow
- Post-Processing: Auffälligste Tokens z.B. „ja“ automatisiert als PTK taggen
- Weitere Transkripte manuell korrigieren und Tagger neu trainieren

Vielen Dank für Ihre Aufmerksamkeit

Häufigste Tagging-Fehler(>2%)

Transkript 1
Berufsschule

Korrektur-Tags	Prozentualer Anteil in der Fehlerquote
PTK	29.72%
PTKHES	9.5%
PTKANT	7.23%
ITJ	6.69%
XY	6.56%
VVINF	5.62%
NN	4.15%
ADJD	3.88%
VVFIN	2.54%
PWAV	2.28%
KON	2.14%
PDS	2.14%
PTKVZ	2.01%

Transkript 2
Studentengespräch

Korrektur-Tags	Prozentualer Anteil in der Fehlerquote
PTK	46.74%
PTKHES	7.04%
PPER	4.59%
VVINF	4.16%
PDS	4.16%
KON	3.52%
ADV	3.09%
XY	2.88%
ADJD	2.24%
NN	2.13%
PIS	2.13%
ITJ	2.13%

Transkript 3
Kind-Kind Vorlesen

Korrektur-Tags	Prozentualer Anteil in der Fehlerquote
XY	23.86%
PTK	18.78%
ADV	7.36%
PDS	4.57%
ITJ	4.57%
PPER	3.3%
NE	3.3%
PRELS	3.05%
KON	3.05%
VVFIN	3.05%
PTKANT	2.79%
VVINF	2.28%
PTKHES	2.28%
KOUS	2.03%
ADJD	2.03%

Auswertung Tokens Transkript 1

Form	# Korrekturen	Korrigiert in...
ja	132	PTK 95
		PTKANT 35
		ITJ 2
äh	69	PTKHES 69
mal	41	PTK 41
so	40	ITJ 40
also	27	PTK 27
gut	22	PTKANT 14
		PTK 8
wie	17	PWAV 17
ganz	14	ADJD 14
T	12	XY 12
als	12	KON 12
einfach	10	PTK 10
das	10	PDS 8
		KOUS 1
		PRELS 1
minus	10	NN 10
was	9	PWS 7
		PRELS 2

Auswertung Tokens Transkript 2

Form	# Korrekturen	Korrigiert in...	
ja	157	PTK	150
		PTKANT	7
äh	65	PTKHES	65
halt	49	PTK	49
hm	39	PTK	38
		PTKHES	1
aber	36	KON	24
		PTK	12
mal	31	PTK	31
doch	27	PTK	27
mich	22	PPER	22
einfach	21	PTK	21
auch	21	PTK	21
nein	19	PTK	19
schon	19	PTK	19
die	18	PDS	13
		ART	2
		PRELS	2
		PPER	1
das	17	PDS	15
		PRELS	2
+++	13	XY	13
mir	13	PPER	9
		PRF	4

Auswertung Tokens Transkript 3

Form	# Korrekturen	Korrigiert in...
ja	29	PTK 18 PTKANT 11
die	19	PDS 12 PRELS 5 ART 2
hm	17	PTK 9 XY 8
%	11	XY 11
einmal	10	PTK 10
sich	9	PPER 9
mh	9	PTK 7 XY 2
der	9	PRELS 4 PDS 4 ART 1
äh	9	PTKHES 9
hi	9	XY 9
+++	8	XY 8
als	8	KON 5 KOUS 3
Vesuv	7	NE 7
da	6	ADV 4 XY 2

ADV statt PTK

2.10.5 PTKANT: Antwortpartikel

Als Antwortpartikel werden die Wortformen *ja, nein, danke, bitte* bezeichnet, die im allgemeinen nur in direker Rede vorkommen und dann alleine einen Satz bilden oder in einem Antwortsatz als Bejahung, Verneinung oder Verstärkung verwendet werden.

Klassifikation von PTKANT

POS =	Beschreibung	Beispiele
PTVANT	Antwortpartikel	{*ja, nein, danke, bitte, doch* }
Aber:		
ADV	Abtönungspartikel	*er ist* **ja**/ADV *schon da*

Ergebnisse Pankow & Petterssons

4.2 Analyse der Fehler bei TreeTagger

In Tabelle 6 wird eine Übersicht der Fehler pro Wortart gegeben.

Tab. 6: Fehler pro Wortart bei TreeTagger

	Fehler	Tokens	Fehler-quote	Fehler Subklasse	Fehler Wortart	Fehler-quote Wortart
ADJEKTIVE	4	141	2,84%	1	3	2,13%
ADVERBIEN	105	543	19,34%	23	82	15,10%
ARTIKEL	11	303	3,63%	0	11	3,63%
INTERJEKTIONEN	3	4	75,00%	0	3	75,00%
KONJUNKTIONEN	10	215	4,65%	4	6	2,79%
PRONOMEN	133	535	24,86%	23	110	20,56%
PRÄPOSITIONEN	9	272	3,31%	1	8	2,94%
EIGENNAMEN	22	93	23,66%	0	22	23,66%
SUBSTANTIVE	13	474	2,74%	0	13	2,74%
VERBEN	129	571	22,59%	69	60	10,51%
VERBZUSATZ	15	21	71,43%	0	15	71,43%
ZU-KLASSE	1	10	10,00%	0	1	10,00%
ZAHLWÖRTER	7	36	19,44%	0	5	13,89%
INSGESAMT:	**462**	**3218**	**14,36%**		**339**	**10,53%**
KORREKTHEITS-RATE:			**85,64%**			**89,47%**

Bibliographie

- STTS-Tagtable (1995/1999). Online verfügbar unter http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html, zuletzt geprüft am 10.09.2012.
- Deppermann, Arnulf; Hartung, Martin (2011): Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des "Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK) am Institut für Deutsche Sprache (Mannheim). In: Ekkehard Felder, Marcus Müller und Friedemann Vogel (Hg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. Berlin: De Gruyter, S. 414–450.
- Feldweg, Helmut (1996): Die Wortformen der geschlossenen Wortarten im Stuttgart-Tübingen Tagset (STTS). Online verfügbar unter http://www.sfs.uni-tuebingen.de/Elwis/stts/Wortlisten/WortFormen.html, zuletzt geprüft am 10.09.2012.
- Institut für Deutsche Sprache (Hg.) (2012): FOLK. Forschungs- und Lehrkorpus Gesprochenes Deutsch. Online verfügbar unter http://agd.ids-mannheim.de/folk.shtml, zuletzt aktualisiert am 02.05.2012, zuletzt geprüft am 17.08.2012.
- Pankow, Christiane; Pettersson, Helena: Auswertung der Leistung von zwei frei zugänglichen POS-Taggern für die Annotation von Korpora des Gesprochenen Deutsch. Online verfügbar unter https://gupea.ub.gu.se/bitstream/2077/19368/1/gupea_2077_19368_1.pdf, zuletzt geprüft am 10.09.2012.
- Schiller, Anne; Teufel, Simone; Thielen, Christine; Stöckert, Christine (1995): Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Institut für maschinelle Sprachverarbeitung (Stuttgart); Universität Tübingen Seminar für Sprachwissenschaft (Tübingen). Online verfügbar unter ftp://ftp.ims.uni-stuttgart.de/pub/corpora/stts_guide.pdf, zuletzt aktualisiert am 14.10.1995, zuletzt geprüft am 17.08.2012.
- Schmid, Helmut: Improvements In Part-of-Speech Tagging With An Application To German. Institut für maschinelle Sprachverarbeitung (Stuttgart). Online verfügbar unter ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf, zuletzt geprüft am 17.08.2012.