Erweiterung des STTS für gesprochene Sprache

Ines Rehbein, Sören Schalowski und Heike Wiese

Institut für Deutsche Sprache
SFB 632 Informationsstruktur
Universität Potsdam

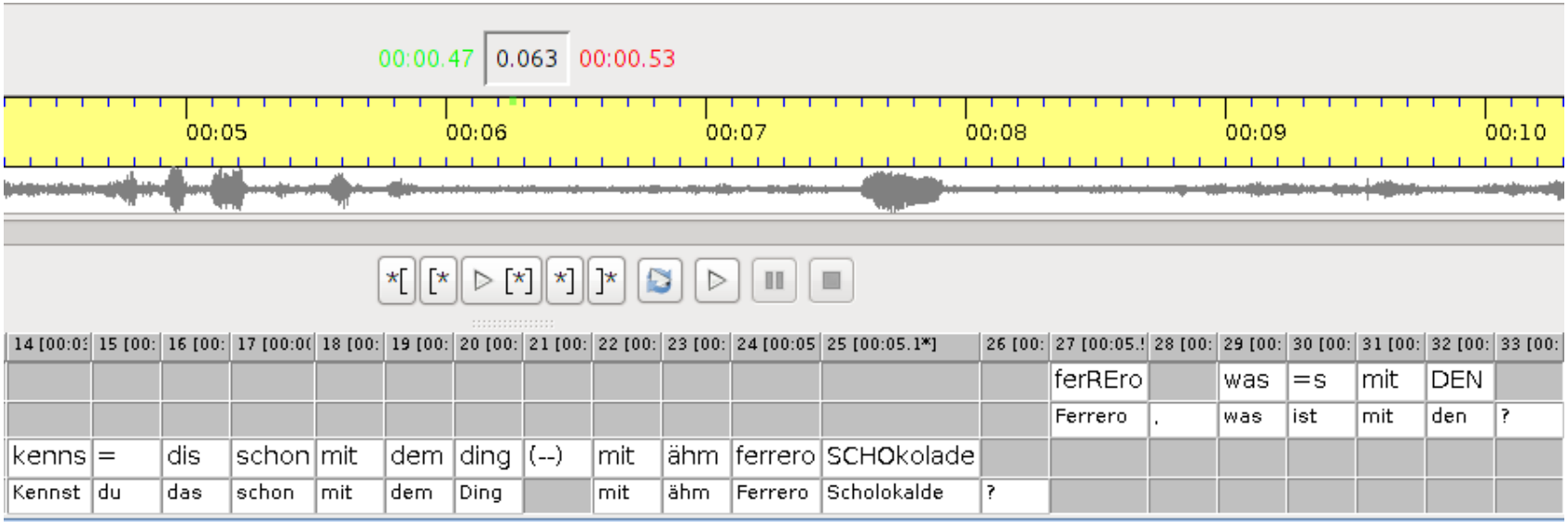
STTS Workshop am IMS Stuttgart
September 2012

Outline

Motivation

- Was?
  - Erweiterung des Stuttgart-Tübingen Tagsets (STTS) [1] für die Annotation von gesprochener Sprache
- Warum?
  - Besonderheiten mündlicher Kommunikation (z.B. gefüllte und nicht-gefüllte Pausen, Abbrüche, Rezeptionspartikeln, Fragepartikeln, ...) im STTS nicht angemessen berücksichtigt
- Wie?
  - STTS wird unverändert übernommen, um Interoperabilität mit existierenden Ressourcen zu gewährleisten
  - Einführung neuer Tags für Besonderheiten gesprochener Kommunikation

KiDKo – das KiezDeutsch-Korpus



Outline

Erweitertes Tagset (Übersicht)

	POS	**Beschreibung**
1	PTK	*unspezifische Partikeln*
2	PTKREZ	*Rezeptionspartikeln*
3	PTKONO	*Onomatopoeia*
4	PTKQU	*Fragepartikeln*
5	PTKPH	*Platzhalter*
6	PTKFILL	*gefüllte Pausen*
7	PAUSE	*stille Pausen*
8	NINFL	*Inflektive*
9	XYB	*Wortabbrüche*
10	XYU	*unverständliches Material*
11	$#	*abgebrochene Äußerungen*

PTK

- für unspezifische Partikeln wie *ja*, *na* in äußerungsinitialer (oder -finaler) Position (3, 4)

(1) A: Kommst Du auch ? B: **Ja**$_{PTKANT}$.

(2) Die hat **ja**$_{ADV}$ auch nicht funktioniert .

(3) **Ja**$_{PTK}$ wer bist du denn ?

(4) **Na**$_{PTK}$ den Gelben hab ich zu Hause .

PTKREZ

- für Rezeptionssignale (einfache, nicht-emotionale Reaktion der Rezipientin, die anzeigt, dass die Äußerung einer Sprecherin gehört und verstanden wurde)

(5) A: Stell dir das mal vor !
B: **M-hm**$_{PTKREZ}$.

PTKONO

- für Onomatopoeia und Formen von Lautmalerei

(6) Das Lied ging so **lalalala**$_{PTKONO}$.

(7) Interessant , **bla**$_{PTKONO}$ **bla**$_{PTKONO}$.

(8) **Eieieieia**$_{PTKONO}$!

(9) **Lalabillebillebille**$_{PTKONO}$!

(10) **Bam**$_{PTKONO}$, **bam**$_{PTKONO}$, **bam**$_{PTKONO}$!

PTKQU

- für Fragepartikeln wie *ne, gell, wa*, die am Ende einer positiven oder negativen deklarativen Äußerung stehen (11, 12, 13),

(11) Wir treffen uns am Kino , **ne**$_{PTKQU}$?

(12) Du kommst auch , **gell**$_{PTKQU}$?

(13) Italien hat gestern prima gespielt , **wa**$_{PTKQU}$?

- für Fragepartikeln nach Sprecherwechsel (14)

(14) A: Schwuchtel !
B: **Hä**$_{PTKQU}$?

PTKPH

- als Platzhalter, wenn die korrekte Wortklasse nicht aus dem Kontext inferiert werden kann

(15) Sie hat ein **Dings**$_{NN}$ gekauft .

(16) Er hat **dings**$_{PTKPH}$ hier .

a. Er hat MP3-Player$_{NN}$ hier .

b. Er hat gewonnen$_{VVPP}$ hier .

c. Er hat (Schuhe gekauft)$_{VP}$ hier .

PTKFILL

- für gefüllte Pausen (Filler)

(17) Das ist irgend so ein **äh**$_{PTKFILL}$ Rapper .

(18) Und **ähm**$_{PTKFILL}$ Servet ist **ähm**$_{PTKFILL}$ verletzt .

PAUSE

- für ungefüllte Pausen

(19) Das ist irgend so ein **(-)**$_{PAUSE}$ Rapper

(20) Wie , wie **(-)**$_{PAUSE}$ Arda .

(21) Ich werde das mit , **(-)**$_{PAUSE}$ mit meinem **(- -)**$_{PAUSE}$ Freund machen .

NINFL

- für Inflektive [2] (gebräuchliches Stilmittel in Comics und CMC, aber auch in gesprochener Sprache)

(22) Ich muss noch putzen . **Seufz** !

(23) Gleich haben wir Mathe . **Gähn** !

XYB und XYU

- **XYB**
  - für Wortabbruch

(24) Ich **ha**$_{XYB}$ # Sie kommt **Sams**$_{XYB}$ äh Sonntag .

- **XYU**
  - für unverständliches Material (schlechte Audioqualität, Codeswitching)

(25) So schnell mal **(unverständlich)**$_{XYU}$.

(26) Wir waren gestern bei **(fremdsprachlich)**$_{XYU}$.

  - für Nichtworte

(27) Oh ! **Viesch**$_{XYU}$!

(28) **Ndrrisch**$_{XYU}$!

Interpunktion – $#

- Interpunktion zur Markierung von abgebrochenen Äußerungen

(29) Sie war ge #$#

(30) Sie war #$#

Outline

Interannotator-Agreement

- 3 Annotator/innen
- erweitertes Tagset (65 Tags)
- Testset: spontane Dialoge ($\geq$ 2 Sprecher/innen) aus dem KiDKo Korpus [3] (3415 Token)

  $\Rightarrow$ Fleiss' κ: **0.975** (% Agr. 96.5)

- Häufige Fehlerquelle:
  - Annotation von *ja* (Verwechslung von Antwortpartikeln und Diskurspartikeln)

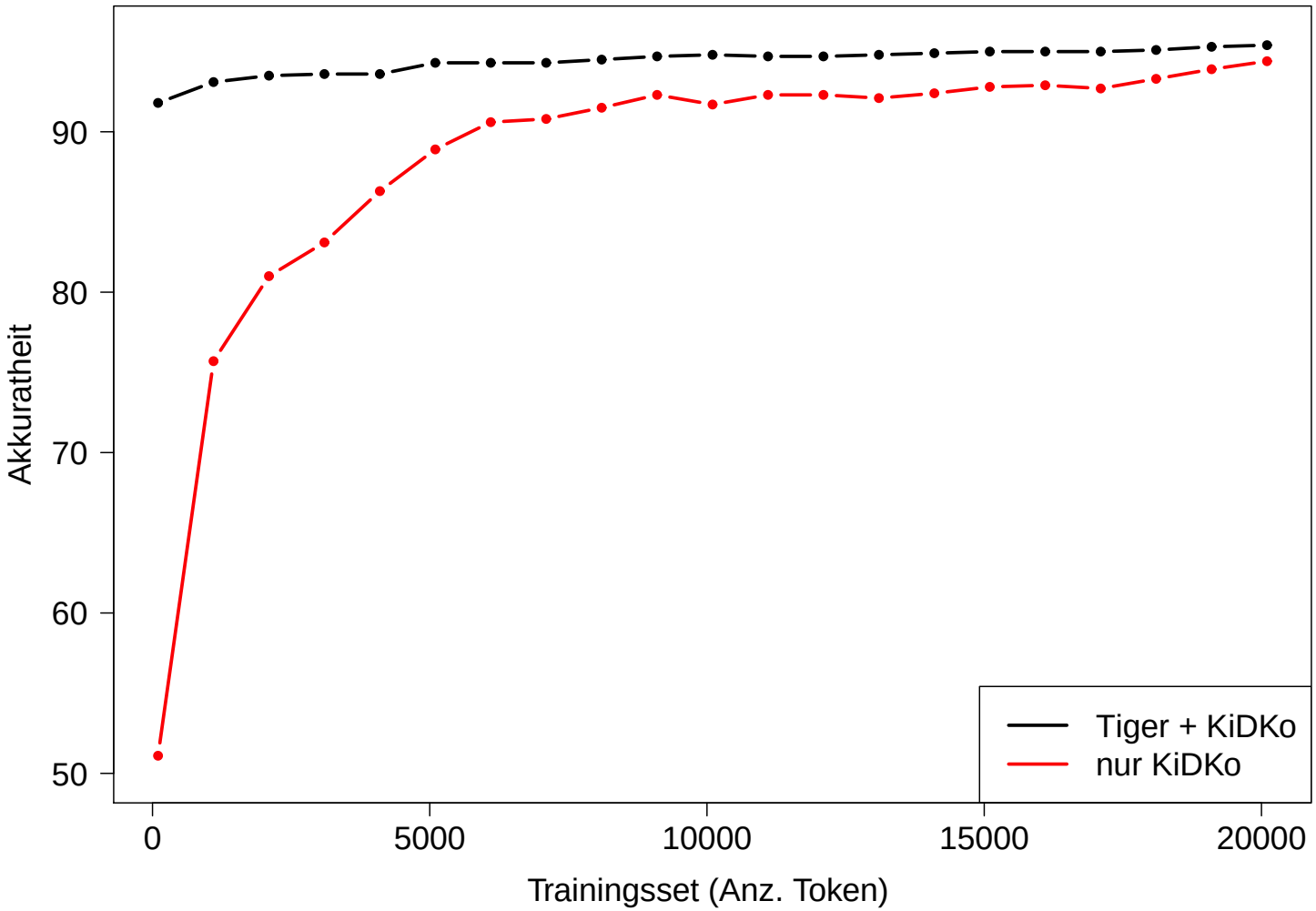
Evaluation POS-Tagger

Tagger:	Treetagger [4]
Trainingsdaten:	KiDKo (20.000 Token)
Testdaten:	KiDKo (8.800 Token)

- Tagger-Anpassung (Skript lookup.perl)
  - reguläre Ausdrücke für anonymisierte Namen (Personen, Straßen, ...)
  - Pausen, fremdsprachliches/unverständliches Material

Evaluation POS-Tagger

TIGER + KiDKo vs. nur KiDKo



Outline

Fazit

- Erweitertes Tagset ermöglicht eine adäquatere Beschreibung von Phänomenen gesprochener Sprache
- Interoperabilität mit existierenden linguistischen Ressourcen geschriebener Sprache ist gewährleistet
- Neu annotierte Sprachdaten können mit vorhandenen Trainingsdaten geschriebener Sprache kombiniert werden, um Systeme zur automatischen Verarbeitung natürlicher Sprache an die neue Domänen anzupassen

- Anne Schiller and Simone Teufel and Christine Thielen. *Guidelines für das Tagging deutscher Textkorpora mit STTS*, Universität Stuttgart, Universität Tübingen, 1995.
- Oliver Teuber. fasel beschreib erwähn – Der Inflektiv als Wortform des Deutschen. Germanistische Linguistik 6(26), S.141–142. 1998.
- Heike Wiese and Ulrike Freywald and Sören Schalowski and Katharina Mayr. Das Kiezdeutsch-Korpus. Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. In: Deutsche Sprache 40(2), S. 97–123. 2012.
- Helmut Schmid. Improvements In Part-of-Speech Tagging With an Application To German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland. 1995.
- Sabine Brants and Stefanie Dipper and Silvia Hansen and Wolfgang Lezius and George Smith. The TIGER Treebank. Proceedings of the First Workshop on Treebanks and Linguistic Theories. Sozopol, Bulgaria. 2002.

Danke fürs Zuhören!

Ambiguität in gesprochener Sprache

(31) Was **machs**$_{VVFIN}$ äh was mache ich falsch ?

(32) Ich habe ja nicht verstanden , wieso **euch**$_{PRF}$ äh wieso ihr euch gestritten habt .

(33) Wir **scha**$_{XYB}$ werden es nicht schaffen .

(34) Schni Schna Schnappi

PTK

- für unspezifische Partikeln wie *ja*, *na* in äußerungsinitialer (oder -finaler) Position

POS	**TIGER**	**TüBa-D/Z**	**TüBa-D/S**
PTKANT	43	147	27 986
ADV	154	372	4 679
ITJ	2	0	0
NN	0	16	0
total	199	536	32 664

Table: Distribution von *ja* in verschiedenen Korpora, normalisiert durch Korpusgröße

(35) A: Kommst Du auch ? B: **Ja**$_{PTKANT}$.

(36) Die hat **ja**$_{ADV}$ auch nicht funktioniert .

(37) **Ja**$_{PTK}$ wer bist du denn ?