Wortartenannotation im Kiezdeutschkorpus (KiDKo 1.0) – Draft

Ines Rehbein

October 28, 2014

Contents

1 Einleitung

Dieses Dokument beschreibt die Erweiterungen des Stuttgart-Tübingen Tagsets (STTS) [19] für die Annotation von gesprochener Sprache. Diese Erweiterung ist notwendig, da das Tagset ursprünglich für die Annotation von Registern geschriebener Sprache entwickelt wurde und folglich keine Wortartenkategorien für die Annotation von gesprochen-sprachlichen Phänomenen wie z.B. Hesitationen, Fragepartikeln oder von Rezeptionssignalen bereitstellt. Unsere Erweiterung ergänzt das STTS um 10 neue Tags und schlägt eine Neuordnung des hierarchisch aufgebauten Tagsets vor.

Bei der Erweiterung des Tagsets haben wir, soweit möglich, die Kategorien und Annotationsrichtlinien des Stuttgart-Tübingen Tagsets übernommen. Nur für lexikalischen Einheiten, die sich nicht mit den vorhandenen STTS-Kategorien beschreiben ließen, haben wir neue Tags zum Tagset hinzugefügt, um den Besonderheiten gesprochener Sprache gerecht zu werden. Dabei verfolgen wir einen Ansatz, den wir als *datengetrieben* und *bottom-up* bezeichnen. In einem ersten Schritt haben wir versucht, mit dem vorhandenen Taginventar des STTS Daten gesprochener Sprache zu annotieren. Dabei haben wir Problemfälle identifiziert und, wenn nötig, für diese neue Klassen eingeführt, die wir dann im weiteren Annotationsprozess getestet haben. Nur wenn die neuen Klassen von unseren Annotator_innen konsistent angewendet werden konnten, haben wir sie beibehalten. Schwierige Unterkategorien, die sich nicht verlässlich von ihren Nachbarklassen abgrenzen ließen, haben wir mit diesen in eine neue Klasse zusammengefasst, so z.B. die Rezeptionssignale und Antwortpartikeln in die neue Klasse der Responsive (SPRS).

Im Weiteren geben wir eine Übersicht über unsere Tagseterweiterung (Abschnitt 2) und diskutieren diese im Hinblick auf andere Vorschläge zur Annotation von konzeptionell mündlichen Daten (Abschnitt 4).

2 Übersicht über die Tagseterweiterung

Tabelle 2 zeigt die im STTS vorhandenen hierarchischen Oberklassen, die weiter unterteilt werden können in 54 feinkörnige Tags.

	STTS		**KiDKo**	
Beschreibung	Oberklasse	Unterkategorien	Oberklasse	Unterkategorien
Adjektive	ADJ	ADJA ADJD	ADJ	ADJA ADJD
Adverbien	ADV	ADV	ADV	ADV
Adpositionen	AP	APPR APPRART APPO APZR	AP	APPR APPRART APPO APZR
Artikel	ART	ART	ART	ART
Kardinalzahlen	CARD	CARD	CARD	CARD
Interjektionen	ITJ	ITJ	–	
Konjunktionen	KO	KOKOM KON KOUI KOUS	KO	KOKOM KON KOUI KOUS
Nomina	N	NE NN	N	NE NN
Pronomina	P	PDS PDAT PIS PIAT PIDAT PPER PPOSS PPOSAT PRELS PRELAT PRF PWS PWAT PWAV PAV	P	PDS PDAT PIS PIAT – PPER PPOSS PPOSAT PRELS PRELAT PRF PWS PWAT PWAV PROAV
Partikeln (Standarddeutsch)	PTK	PTKA PTKANT PTKNEG PTKVZ PTKZU	PTK	PTKA – PTKNEG PTKVZ PTKZU
Partikeln (gesprochener Sprache)	–		SP	SPFILL SPINI SPITJ SPONO SPQU SPRS
Verben	V	VAFIN VAIMP VAINF VAPP VMFIN VMINF VMPP VVFIN VVIMP VVINF VVPP – VVIZU	V	VAFIN VAIMP VAINF VAPP VMFIN VMINF VMPP VVFIN VVIMP VVINF VVPP VVINFL VVIZU

	STTS		**KiDKo**	
Beschreibung	Oberklasse	Unterkategorien	Oberklasse	Unterkategorien
Fremdsprachliches Material	FM	FM	FM	FM
Kompositionserstglied	TRUNC	TRUNC	TRUNC	TRUNC
Platzhalterpartikel (nicht eindeutig bestimmbar)	–		DINGS	DINGS
Nichtwörter	XY	XY	XY	XYB XYS XYU
Interpunktion	$	$. $, $(	$	$. $, $($#

2.1 Gesprächspartikeln – SPRS, SPFILL, SPINI, SPQU, SPITJ, SPONO

Die meisten Änderungen wurden im Bereich der Partikeln vorgenommen. Die Partikel-Klasse (PTK) im STTS beinhaltet die Unterklasse der Antwortpartikeln *ja, nein* (PTKANT), Partikeln bei Adjektiv oder Adverb (*zu* schön, *am* schönsten; PTKA), die Negationspartikel *nicht* (PTKNEG), abtrennbare Verbzusätze (PTKVZ), und *zu* vor Infinitiv (PTKZU). Für KiDKo haben wir zusätzlich eine neue Klasse eingeführt, die Partikeln in gesprochener Sprache beschreibt (SP).

Gemeinsames Merkmal aller Unterklassen von SP ist, dass ihre Elemente nicht flektierbar sind und vorwiegend in konzeptionell mündlicher Sprache auftreten. Unterschiede gibt es jedoch in Bezug auf ihre syntaktische Distribution. Während es sich bei den Responsiven und Interjektionen um satzfähige Einheiten handelt (Beispiele 1, 2), können Filler sowohl isoliert als auch innerhalb einer Äußerung auftreten und dort an jeder beliebigen Position stehen (Beispiel 3). Gliederungspartikeln leiten eine Äußerung ein, sind jedoch nicht syntaktisch in diese integriert (Beispiel 4). Fragepartikeln stehen üblicherweise am Ende einer deklarativen Äußerung (Beispiel7), stehen also weder isoliert, da sie sich auf eine vorhergehende Äußerung beziehen, noch sind sie syntaktisch in die Äußerung integriert. Es gibt also Unterschiede dahingehend, ob eine Form isoliert stehen kann oder nicht, jedoch sind alle unter SP zusammengefassten Formen nicht syntaktisch in die Äußerung integriert (d.h. sie füllen weder eine Argumentstelle, noch modifizieren sie Elemente der Äußerung), auch wenn sie innerhalb der Äußerung positioniert sind.

2.1.1 Responsive – SPRS

Aus den obigen Überlegungen heraus haben wir die Antwortpartikeln aus der Klasse der Partikeln herausgenommen, da diese selbständige Äußerungen bilden können, während die anderen vier Partikelklassen (PTKA, PTKNEG, PTKVZ, PTKZU) nur syntaktisch integriert auftreten. Die Antwortpartikeln haben wir der neuen SP-Klasse zugeordnet, die Phänomene dialogischer Kommunikation abdeckt, und zusammen mit den Rezeptionssignalen (z.B. *m-hm, mhh*) zu einer neuen Klasse SPRS (Responsiv) erweitert. Diese Verschmelzung der beiden Klassen haben wir vorgenommen, da in vielen Fällen eine eindeutige Unterscheidung zwischen der Funktion von Antwortpartikeln und Rezeptionssignalen nicht möglich ist. Beispiel 1 illustriert dies.

(1) Responsive (SPRS)

a. A: Kommst du heute auch ? B: **M-hm/Ja** . *(Antwortpartikel)*

b. A: Der Film gestern war super . B: **M-hm/Ja** . *(Backchannel)*

Die SP-Klasse beinhaltet neben der Klasse der Responsive noch Interjektionen, Filler (Hesitationen), Fragepartikeln, Gliederungspartikeln in äußerungsinitialer Position, und Onomatopoetika und andere Formen der Lautmalerei.

2.1.2 Interjektionen – SPITJ

Unsere Definitionen folgt der im STTS, die wiederum auf Bußmann [5] basiert.

¡

Interjektionen sind Wörter,

> "die zum Ausdruck von Empfindungen, Flüchen und Verwünschungen sowie zur Kontaktaufnahme dienen. [...] sie sind formal unveränderlich, stehen syntaktisch außerhalb des Satzzusammenhanges und haben im strengen Sinn keine lexikalische Bedeutung." *Bußmann (1990)*

(2) Interjektionen (SPITJ)

a. **Boah** !

b. **Ey** !

c. **Hallo** !

Die Einwortäußerung *Mann!*, die sowohl Anrede als auch Interjektion sein kann, wurde im Korpus konsistent als SPITJ annotiert, da sie in KiDKo überwiegend nicht zur Anrede verwendet, sondern als Exklamaktion eingesetzt wird. Andere häufig vorkommende Anredeformen wie z.B. *Süße, Dicker/Digga, Alter, Oğlum* haben wir als Nomen (NN) getaggt.

2.1.3 Filler/Hesitationen – SPFILL

Gefüllte Pausen (oder Hesitationen) können an jeder beliebigen Position innerhalb einer Äußerung auftreten. Sie können unterschiedliche Funktionen erfüllen, so z.B. als *floorholder* oder *turn-taking-Signale* eingesetzt werden [15, 9, 14, 6, 3, 7], oder um das Rederecht zu behalten oder zu bekommen [15, 18, 4]. Oft werden sie in der Planungsphase oder bei Wortfindungsstörungen produziert. Filler sind aber keine reinen Performanzphänomene. Sie können auch eine pragmatische Funktion in der Kommunikation haben, z.B. zur Markierung von Unsicherheit, als Höflichkeitsmarker oder auch zur Strukturierung des Diskurses eingesetzt werden [8, 3, 1].

Bei der Annotation unterscheiden wir nicht zwischen diesen verschiedenen Funktionen, sondern annotieren alle Fillerpartikeln mit dem Tag SPFILL. Wir beschränken uns auf Elemente ohne eigene lexikalische Bedeutung (*äh, ähm, öh, mh, ehm, ja*). Elemente mit lexikalischer Bedeutung können zwar im Diskurs eine ähnliche Funktionen erfüllen, werden aber in KiDKo nicht als Filler annotiert.

(3) Filler (SPFILL)

a. **Äh** .

b. **Äh** ich geh dann mal .

c. Ich **äh** geh dann mal .

d. Ich geh **äh** dann mal .

2.1.4 Äußerungsinitiale Gliederungspartikeln– SPINI

Das Tag SPINI wird zur Annotation von Gliederungspartikeln benutzt, die meist in äußerungsinitialer Position im Vor-Vorfeld stehen und (im Gegensatz zu den Interjektionen) selten Akzent tragen. In der Literatur werden sie als *Eröffnungssignale* [22] oder *Gliederungssignale in äußerungsinitialer Position* [23] beschrieben, oder als *Diskursmarker im Vor-Vorfeld* [2].

(4) Gliederungspartikeln (SPINI)

a. **Na** du bist ja gut drauf!

b. **Ja** wer bist du denn?

Wir benutzen diese Klasse vorwiegend für die Wortformen *ja/na*. Bei *ja* unterscheiden wir zwischen Modalpartikeln, die im Mittelfeld auftreten und nach den Richtlinien des STTS als ADV annotiert werden, und *ja* in äußerungsinitialer Position (4b), das wir als Diskursmarker oder Gliederungspartikel betrachten und mit dem Tag SPINI versehen.[1]

(5) Andere Wortklassen von *ja*

a. Wird **ja/ADV** mal Zeit .

b. Yippieh , yippieh yah , yah ! **Ja/SPITJ** !

c. A: Hast die Kohle ? B: **Ja/SPRS** .

d. Als ich gegangen bin , ruft meine **ja/SPFILL** äh eine Freundin von meiner Tante an

e. Also Montag ist Party . **Ja/SPQU** ?

2.1.5 Fragepartikeln – SPQU

Fragepartikeln (question tags) sind Partikeln wie z.B. *ne, wa, ja, gell* und Konjunktionen (*und, oder*, 7b), die üblicherweise am Ende einer deklarativen Äußerung stehen (6a-7b). Wir benutzten das SPQU-Tag aber nicht nur für Tagquestions im engeren Sinne, sondern auch für Fragepartikeln nach einem Sprecherwechsel (6d).

[1] Siehe auch Meer (2007) [16]) für eine Diskussion der verschiedenen Wortklassen von *ja*.

(6) Fragepartikeln (SPQU)
   a. Du kommst doch auch . **Ne** ?
   b. Der Film war voll geil . **Wa** ?
   c. Total geil . **Oder** ?
   d. SPK1: Du bist doof . SPK2: **Hä** ?

Adjektive wie *richtig, okay* können ebenfalls die Funktion einer Fragepartikel einnehmen. In diesem Fall annotieren wir jedoch ADJD, da die Abgrenzung zu elliptischen Äußerungen schwer zu treffen ist (siehe 7a-7b).

(7) Fragepartikeln (SPQU)
   a. Wir gehn jetzt los . **Okay/ADJD** ?
   b. Wir gehn jetzt los . Ist das **okay/ADJD** ?

2.1.6 Onomatopoetika und andere Formen der Lautmalerei – SPONO

Onomatopoetika verhalten sich ähnlich wie Interjektionen. Sie können sowohl isoliert stehen und so eine eigene, selbständige Äußerung bilden (8a) als auch in eine Äußerung integriert werden (8b). Es ist deshalb fraglich, ob die Einführung einer neuen Klasse SPONO gerechtfertigt ist, deren trennscharfe Abgrenzung von den Interjektionen schwierig ist, da sie viele Eigenschaften der Interjektionen teilt. Wir haben uns trotz dieses Problems dafür entschieden, Formen von Lautmalerei separat zu annotieren, da es sich um eine produktive Klasse handelt, die nur schwer durch Suchmuster zu erfassen und deshalb im Korpus schwer auffindbar ist. Wir behaupten also nicht, dass es sich hierbei um eine eigene, grammatisch definierte Wortklasse handelt, sondern betrachten die separate Annotatation als ein Hilfsmittel, das die Auffindbarkeit solcher Konstrukte im Korpus erleichtern soll.

Die Oberkategorie SP zeigt an, dass beide Klassen (SPITJ und SPONO) verwandte Phänomene beschreiben. Bei der Arbeit mit dem Korpus sollte sollten jedoch im Hinterkopf behalten werden, dass diese Auszeichnungen keinen Anspruch auf Vollständigkeit erheben und auch die Übereinstimmung zwischen den Annotator_innen bei der Annotation dieser Klassen aufgrund der problematischen Abgrenzung niedriger ist als zum Beispiel für die Unterscheidung von Präpositionen und Postpositionen.

(8) Onomatopoetika und Formen von Lautmalerei (SPONO)
   a. **Batsch** !
   b. Mach **wau , wau , wau** !

2.2 Nicht flektierte Verbstämme – VVINFL

Eine weitere neue Unterkategorie ist die der unflektierten Verbformen (VVINFL). Bei diesen so genannten *Inflektiven* handelt es sich um nicht flektierte, prädikativ gebrauchte Verbstämme [25]. Sie sind ein häufig eingesetztes Stilmittel in Comics und in computer-vermittelter Kommunikation, treten jedoch auch in gesprochener Sprache auf. Während sowohl einfache Inflektive als auch komplexe Inflektivkonstruktionen in computervermittelter Kommunikation hoch frequent sind, kommen sie in gesprochener Sprache eher selten vor. Beispiel 9 zeigt zwei Belege aus KiDKo.

Inflektive werden durch die Tilgung der Flektionsendung gebildet und lassen sich leicht ins Verbparadigma einfügen. Die Verben behalten ihre lexikalische Semantik und ihre Argumentstellen (wobei jedoch die des Subjekts typischerweise ungefüllt bleibt und per default mit der Sprecherin identifiziert wird (9b) (siehe hierzu auch [20]:208)).

(9) a. **Knutsch** !
   b. Daumen **drück** .

2.3 Platzhalterpartikel – DINGS

Ein weiteres neues Tag, DINGS, beschreibt keine grammatische Wortklasse, sondern kann (genau wie das TRUNC-Tag im STTS) auf Elemente verschiedener Wortklassen angewendet werden. Wir annotieren DINGS, wenn es aufgrund von fehlendem Kontext nicht möglich ist, die Wortklasse eines lexikalischen Elements (vorwiegend mit der Wortform *dings*) zu bestimmen.

Während in Beispiel (10a) aufgrund des Artikels eindeutig bestimmt werden kann, dass es sich um ein Nomen handeln muss, gibt es für (10b) mehrere verschiedene Möglichkeiten (10b, i-iii). In diesen Fällen annotieren wir nicht die wahrscheinlichste Lesart, sondern weisen das Tag DINGS zu.

(10) Platzhalterpartikel (DINGS)

a. Er hat ein **Dings/NN** hier .
b. Er hat **dings/DINGS** hier .
   i. Er hat MP3-Player/NN hier .
   ii. Er hat gewonnen/VVPP hier .
   iii. Er hat [Schuhe gekauft]/VP hier .

2.4 Nichtworte – XY

Die Oberklasse XY der Nichtworte im STTS beinhaltet genau ein Tag, nämlich XY. Wir haben diese Oberklasse weiter aufgeteilt in die Tags XYS, XYB und XYU.

XYS entspricht dem XY-Tag im STTS, das allgemein zur Auszeichnung von Nichtworten/Sonderzeichen verwendet wird (11a), XYB dient zur Auszeichnung von Wortabbrüchen (11c), und XYU wird zur Auszeichnung von unverständlichem Material verwendet, meist aufgrund von schlechter Audioqualität oder auch für nicht verständliche fremdsprachliche Äußerungen (11c).

(11) Nichtwort (XY)
   a. Motorola/NN V8/XYS
   b. N/XYB äh nächste Woche
   c. Ich habe kein UNINTERPRETABLE/XYU

2.5 Interpunktion – $#

Unsere letzte Erweiterung betrifft den Bereich der Interpunktion. Dort haben wir ein neues Tag für die Markierung von abgebrochenen Äußerungen hinzugefügt. Diesem Abbruch kann ein Wortabbruch vorausgehen (12b), muss aber nicht (12b).

(12) Ende einer abgebrochenen Äußerung ($#)
   a. Ich bin fer **#/$#**
   b. Ich will auch ein **#/$#**

Tabelle 2.5 zeigt die Klassen des STTS, die in unserer Tagseterweiterung modifiziert wurden.

STTS	ITJ	–	PTK	V	XY	–	Interpunktion
	ITJ		PTKA PTKANT PTKNEG PTKVZ PTKZU	VVFIN VVINF VVIMP VAFIN ...	XY		$. $, $(

KiDKo	ITJ	**SP**	PTK	V	XY	DINGS	Interpunktion
		SPRS SPITJ **SPFILL** **SPQU** **SPINI** **SPONO**	PTKA PTKNEG PTKVZ PTKZU	VVFIN VVINF VVIMP **VVINFL** VAFIN ...	XYS **XYB** **XYU**	**DINGS**	$. $, $(**$#**

3 Anmerkungen – Unterschiede zum STTS

3.1 PAV/PROAV/PROP

Während die beiden großen, deutschen Baumbanken das STTS für die Annotation von Wortarten benutzen, gibt es kleine Unterschiede in der Benennung und Interpretation der Tags. Zum Beispiel heißen Pronominaladverbien im STTS PAV, in TiGer PROAV, und in der TüBa-D/Z PROP. In KiDKo sind wir der Benennung in TiGer gefolgt und haben Pronominaladverbien mit dem Tag PROAV ausgezeichnet.

3.2 PIAT/PIDAT

Weitere Unterschiede zwischen den Baumbanken betreffen die Indefinitpronomen. TüBa-D/Z benutzt hier beide im STTS vorhandenen Tags, PIAT (Indefinitpronomen ohne Determinierer) und PIDAT (Indefinitpronomen mit Determinierer), während in TiGer bei attribuierenden Indefinitpronomen nicht weiter unterschieden wird, ob diese mit einem Determinierer vorkommen können oder nicht, und alle attribuierenden Indefinitpronomen mit dem Label PIAT ausgezeichnet werden. Auch hier sind wir TiGer gefolgt und haben alle attribuierenden Indefinitpronomen mit dem PIAT-Tag versehen.

In gesprochener Sprache (und auch in anderen konzeptionell mündlichen Texten) finden wir häufig die Verwendung von *so* und, verschmolzen mit einem indefiniten Artikel, *sone/n/r* in Determinierer-Position. Hole und Klumpp [13] bezeichnen das adnominale *so* auch als dritten Standardartikel im gesprochenen Deutsch. In dieser Position kann *so* allerdings viele verschiedene Funktionen übernehmen, die nicht leicht voneinander abgrenzbar sind und sich zum Teil auch überlappen. In Beispiel (13b) dient *so* als Intensivierunspartikel, Beispiel (13d) kann als Vagheitsmarker oder Heckenausdruck analysiert werden, und in Beispiel (13e) dient *so* als deiktischer Vergleichsausdruck (siehe hierzu auch Umbach und Ebert [26] und Hirschmann [10]). Da eine Unterscheidung in der Praxis oft nicht eindeutig getroffen werden kann (und die Klassen auch theoretisch noch keineswegs klar voneinander abgrenzbar sind), sind wir der Annotationspraxis in der TüBa-D/Z gefolgt (Beispiel 13f) und haben auch adnominale Instanzen von *so* mit dem Label ADV versehen.

Bei der Verschmelzung von *so* mit indefinitem Artikel (*sone*) vor Pluralnomen hingegen haben wir, wenn die Auflösung dieser Verschmelzung aufgrund von fehlender Kongruenz zwischen *sone* und Nomen in einer ungrammatikalischen Äußerung resultieren würde (Beispiel 13g), diese Instanzen analog zu *solche* analysiert und mit dem Tag PIAT ausgezeichnet.[2]

(13) a. um GELD reingeben in so KORB
Um Geld reingeben , in so Korb .
⇒ *so* als indefiniter Artikel

b. ich hab so HUNger
Ich habe so Hunger .
⇒ *so* als Intensivierunspartikel

c. so um ZWEI is do okay
So um Zwei ist doch okay .
⇒ *so* als Vagheitsmarker

d. so TECHno oder so
So Techno oder so .
⇒ *so* als Vagheitsmarker?

e. Ich hab so TEXT gemacht
Ich habe so Text gemacht .
⇒ Vagheitsmarker, Indefinitpronomen oder deiktische Referenz (so, auf diese Art und Weise)?

f. Mein Körper hat schon so Risse überall .
⇒ so/ADV (TüBa-D/Z, s10254)

g. sone sone men in BLACK
Sone , sone Men in Black .
⇒ solche/PIAT Men in Black

4 Verwandte Arbeiten – POS-Annotation im FOLK-Korpus

Die eingeschränkte Nutzbarkeit des STTS für nicht-kanonische und insbesondere konzeptionell mündliche Sprache ist hinlänglich bekannt. Es gibt Bemühungen, dieses Problem anzugehen und Lösungen zu diskutieren (siehe hierzu das Themenheft zu Stand und Perspektiven des STTS [29] und der Beitrag zur Anpassung des STTS für die Annotation von nicht-kanonischer Sprache [28]). Unsere Arbeit versteht sich als Teil dieser Bemühungen auf dem Weg zu einem neuen, erweiterten Wortartentagset für das Deutsche.

Unsere Erweiterungsvorschläge weisen Überschneidungen mit denen von Westpfahl und Schmidt [24] auf, die an der Wortartenannotation des FOLK-Korpus [21] arbeiten.[3] Westpfahl schlägt drei neue Oberkategorien für

[2] Zwar haben wir in unserem Korpus viele Beispiele für nicht-grammatikalische (nach den Regeln der Standardgrammatik der geschriebenen Sprache) Äußerungen, jedoch konnten wir im Korpus keine Belege für derartige Kongruenzverletzungen finden, was uns darin bestärkt hat, *sone* nicht als Kombination aus *so + indefinitem Artikel* zu analysieren, sondern als eine neue Form, die nicht tokenisiert wurde.

[3] Dies ist vielen vorausgegangenen Diskussionen mit Swantje Westpfahl und Hagen Hirschmann geschuldet, die wir als sehr hilfreich und fruchtbar empfunden haben.

die Annotation von Partikeln gesprochener Sprache vor, nämlich 1) Partikeln (PTK), die sich syntaktisch und distributionell beschreiben lassen, 2) nicht-grammatischen Elementen (NG), die sowohl isoliert stehen als auch an jeder beliebigen Position in einer Äußerung vorkommen können, und 3) zwischen satzklammer-externen Elementen (SE, Nachfeld- und Vor-Vorfeld-Elemente), die zwar nicht syntaktisch integriert, jedoch auf Diskursebene pragmatisch gebunden sind.

Unser Annotationsschema ist zum Teil weniger feinkörnig als das FOLK-Tagset.[4] Unsere Klasse SP subsumiert die NG-Tags und die unter SE eingeordneten Rückversicherungssignale/Questiontags, während die PTK-Klasse in KiDKo nach den Regeln des herkömmlichen STTS als ADV annotiert wurden. Jedoch bietet die Annotation in KiDKo eine Unterscheidung zwischen Interjektionen (SPITJ) und Responsiven (SPRS), die in Westpfahls Annotationsschema unter NGIRR zusammengefasst werden. Beide Tagsets bieten Tags zur Annotation von Fillern/Hesitationen und von Onomatopoetika, das FOLK-Tagset stellt ein weiteres Tag für Aktionswörter bereit, das sowohl für nicht flektierte Verbformen als auch für Akronyme verwendet wird. Erstere werden in KiDKo mit einem eigenen Tag (VVINFL) annotiert, während Akronyme wie *lol* (laughing out loud) oder *OMG* (oh my god) als Interjektionen getaggt werden.

5 Ausblick

Während mit den hier vorgestellten Erweiterungen des STTS für die Annotation von gesprochener Sprache neue, zusätzliche Kategorien für Gesprächspartikeln definiert wurden, die vorwiegend auf Diskursebene fungieren und pragmatische Funktion haben, wurde die im STTS vorhendene Kategorie ADV zur Auszeichnung von Adverbien übernommen. Die ADV-Kategorie des STTS umfasst Adverbien und Partikeln mit unterschiedlicher Funktion und syntaktischer Distribution wie z.B. Satzadverbien, Intensivierunspartikel, fokusassoziierte Partikeln und Modalpartikeln. In diesem Bereich ist eine weitere Aufteilung und Unterscheidung auf jeden Fall wünschenswert. Sowohl das im FOLK-Korpus verwendete Tagset [24, 27] als auch Arbeiten von Hirschmann [11] und Rehbein et al. [17] schlagen hier Verbesserungen vor, die zur Zeit in Annotationsexperimenten und automatischen Tagging- und Parsingexperimenten erprobt werden. Langfristiges Ziel ist eine einheitliche Erweiterung des STTS, die dann als neuer Quasi-Standard für die Annotation von Wortarten im gesprochenen Deutsch dienen kann.

References

[1] J.E. Arnold, Maria Fagnano, and Michael K. Tanenhaus. Disfluencies signal thee, um, new information. *Journal of Psycholinguistic Research*, 32(1):25–36, 2003.

[2] Peter Auer and Susanne Günthner. Die Entstehung von Diskursmarkern im Deutschen – ein Fall von Grammatikalisierung? In Torsten Leuschner, Tanja Mortelmans, and Sarah de Groodt, editors, *Grammatikalisierung im Deutschen (Linguistik - Impulse und Tendenzen; 9.)*, pages 335–362. Berlin: de Gruyter, 2005.

[3] D.J. Barr. Trouble in mind: paralinguistic indices of effort and uncertainty in communication. In *Oralité et gestualité, communication multimodal, interaction*, pages 597–600. Paris: L'Harmattan, 2001.

[4] G.W. Beattie. *Talk: an analysis of speech and non-verbal behaviour in conversation.* Milton Keynes: Open University Press, 1983.

[5] Hadumod Bussmann. *Lexikon der Sprachwissenschaft.* Alfred Kroner Verlag, Stuttgart, 1990.

[6] H.H. Clark. *Using language.* Cambridge: Cambridge University Press, 1996.

[7] H.H. Clark and J.E. Fox Tree. Using uh and um in spontaneous speech. *Cognition*, 84:73–111, 2002.

[8] Kerstin Fischer. *From cognitive semantics to lexical pragmatics: the functional polysemy of discourse particles.* Mouton de Gruyter: Berlin, New York, 2000.

[9] Erving Goffman. Radio talk. In *Forms of talk*, pages 197–327. Philadelphia, PA: University of Pennsylvania Press, 1981.

[10] Hagen Hirschmann. Außenseiter oder alleskönner? die problematik der klassifikation pränominaler wortarten im deutschen am beispiel des adnominalen so. talk presented at the dgfs-jahrestagung 2014, ag 10, 2014.

[4] Dies ist weniger theoretisch motiviert als der begrenzten Laufzeit des Projekts sowie der begrenzten Anzahl an AnnotatorInnen geschuldet, so dass auf eine wünschenswerte, jedoch zeitintensive feinkörnige Annotation der ADV-Klasse des STTS verzichtet wurde. Siehe hierzu jedoch fortführende Arbeiten wie z.B. [12, 17].

[11] Hagen Hirschmann. *Modifikatoren im Deutschen. Studien zur deutschen Grammatik.* Tübingen, Stauffenburg, to appear.

[12] Hagen Hirschmann, Nadine Lestmann, Ines Rehbein, and Swantje Westpfahl. Erweiterung der Wortartenkategorien des STTS im Bereich ADV und PTK. Präsentation auf dem STTS-Workshop, Hildesheim, Germany, 2013.

[13] Daniel Hole and Gerson Klumpp. Definite type and indefinite token: the article son in colloquial german. *Linguistische Berichte*, 182:231–244, 2000.

[14] W.J.M. Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41–104, 1983.

[15] H. Maclay and C. Osgood. Hesitation phenomena in spontaneous english speech. *Word*, 15:19–44, 1959.

[16] Dorothee Meer. "ja er redet nur MÜLL hier." – Funktionen von 'ja' als Diskursmarker in Täglichen Talkshows. *gidi Arbeitspapierreihe*, 11, 2007.

[17] Ines Rehbein and Hagen Hirschmann. Towards a syntactically motivated analysis of modifiers in german. In *Conference on Natural Language Processing (KONVENS)*, 2014.

[18] S.R. Rochester. The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, 2(1):51–81, 1973.

[19] Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart, Universität Tübingen, 1999.

[20] Peter Schlobinski. *knuddel zurueckknuddel dich ganzdollknuddel*. Inflektive und Inflektivkonstruktionen im Deutschen. *Zeitschrift für Germanistische Linguistik*, 29(2):192–218, 2001.

[21] Thomas Schmidt. The research and teaching corpus of spoken German – FOLK. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).

[22] Johannes Schwitalla. Dialogsteuerung. Vorschläge zur Untersuchung. In Franz Josef Berens, Karl-Heinz Jäger, Gerd Schank, and Johannes Schwitalla, editors, *Projekt Dialogstrukturen. Ein Arbeitsbericht*, pages 73–104. Hueber, München, 1976.

[23] Johannes Schwitalla. *Gesprochenes Deutsch. Eine Einführung.* Erich Schmidt Verlag, Berlin, 2006.

[24] Thomas Schmidt Swantje Westpfahl. POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. *JLCL*, 28(1):139–153, 2013.

[25] Oliver Teuber. fasel beschreib erwähn – Der Inflektiv als Wortform des Deutschen. *Germanistische Linguistik*, 26(6):141–142, 1998.

[26] C. Umbach and C. Ebert. German demonstrative 'so' – intensifying and hedging effects. *Sprache und Datenverabeitung (International Journal for Language Data Processing)*, 1(2), 2009.

[27] Swantje Westpfahl. STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 1–10, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[28] Heike Zinsmeister, Ulrich Heid, and Kathrin Beck. Adapting a part-of-speech tagset to non-standard text: The case of STTS. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4097–4104, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA).

[29] Heike Zinsmeister, Ulrich Heid, and Kathrin Beck. Das Stuttgart-Tübingen Tagset – Stand und Perspektiven. *Journal for Language Technology and Computational Linguistics*, 28(1), 2014.