

Tagset und Richtlinie für das PoS- Tagging von Sprachdaten aus Genres internetbasierter Kommunikation

Michael Beißwenger ▪ Thomas Bartz ▪ Angelika Storrer ▪ Swantje Westpfahl
(Stand: 13.09.2015)

1. Status dieses Dokuments
 2. Erweiterungen und Modifikationen gegenüber STTS (1999) im Überblick
 3. Tags und PoS-Kategorien für IBK-spezifische Phänomene
 - 3.1 Emoticons (EMOASC, EMOIMG)
 - 3.2 Aktionswort (AKW)
 - 3.3 Hashtags und Adressierungen (HST, ADR)
 - 3.4 URLs und E-Mail-Adressen (URL, EML)
 4. Tags und PoS-Kategorien für Phänomene der konzeptionellen Mündlichkeit
 - 4.1 Kontraktierte Formen: Tags für die häufigsten Bildungsmuster (APPRART, VVPPER, VMPPER, VAPPER, KOUSPPER, PPERPPER, ADVART)
 - 4.2 Partikeln
 - 4.2.1 Intensitäts-, Fokus- und Gradpartikeln (PTKIFG)
 - 4.2.2 Modal- und Abtönungspartikeln (PTKMA)
 - 4.2.3 Partikeln als Teile von Mehrwort-Lexemen (PTKMWL)
 - 4.3 Diskursmarker (DM)
 - 4.4 Onomatopoetikon (ONO)
 5. Erwähnte Literatur
- ANHANG: Vollständige Übersicht über das Tagset *STTS_IBK*

1. Status dieses Dokuments

Dieses Dokument beschreibt das Part-of-speech-Tagset, das die Grundlage für die PoS-Annotation im Rahmen der *Shared Task zur automatischen linguistischen Annotation internet-basierter Kommunikation (EmpiriST2015)* bildet. Die dargestellten Tags und Kategorien liegen der manuellen Annotation der Trainings- und Evaluationsdatensets zugrunde. Ihre korrekte Zuordnung zu Tokens in den ausgegebenen Datensets bildet den Zielhorizont für Tagging-Verfahren, die im Rahmen der Shared Task entwickelt werden.

Das Tagset basiert auf dem *Stuttgart-Tübingen Tagset (STTS)* (Schiller et al. 1999). Gegenüber der kanonischen Version des STTS umfasst es Erweiterungen für typische Elemente internet-basierter Kommunikation (IBK) sowie – abgestimmt auf STTS-basierte Tagsets, die für die Annotation von Korpora gesprochener Sprache entwickelt wurden – Erweiterungen und Modifikationen für eine Erfassung von Phänomenen „konzeptioneller Mündlichkeit“, die in gesprochener Sprache und in der schriftlichen internetbasierten Kommunikation gleichermaßen auftreten.¹ Das Tagset wurde 2012–2014 im Kontext des DFG-Netzwerks „Empirische Erforschung internetbasierter Kommunikation“ (*Empirikom*)² sowie in Zusammenhang mit drei CLARIN-D-Workshops zur Erweiterung des STTS erarbeitet.³

Explizit dargestellt werden im vorliegenden Dokument nur diejenigen PoS-Kategorien, die gegenüber der kanonischen Version von STTS (Schiller et al. 1999) als IBK-spezifische Erweiterungen hinzutreten oder die Modifikationen existierender Kategorien darstellen. Für diejenigen Bereiche des STTS, die von den beschriebenen Erweiterungen und Modifikationen nicht betroffen sind, gelten die in Schiller et al. (1999) formulierten Richtlinien.

2. Erweiterungen und Modifikationen gegenüber STTS (1999) im Überblick

Die im Folgenden beschriebene STTS-Version für IBK (fortan: **STTS_IBK**) erweitert STTS (1999) um eine Reihe von IBK-spezifischen Tags, die Phänomene beschreiben, die mit den Kategorien aus STTS (1999) nicht erfasst werden können (EMO, AKW, HST, ADR, URL, EML). Daneben gibt es Tags für Phänomene der konzeptionellen Mündlichkeit, die in redigierten Tex-

1 Vorversionen der Tagset-Erweiterungen für IBK-spezifische Phänomene und für Phänomene gesprochener Sprache sind in Bartz et al. (2013), Westpfahl/Schmidt (2013) und Westpfahl (2014) beschrieben.

2 <http://www.empirikom.net>

3 Für Anregungen und Diskussionen zu Vorversionen des hier beschriebenen Tagsets danken wir den Mitgliedern und Gästen des *Empirikom*-Netzwerks sowie den TeilnehmerInnen der CLARIN-D-Workshops in Stuttgart (2012), Tübingen (2013) und Hildesheim (2013).

ten entweder keine Rolle spielen oder Randerscheinungen darstellen, die aber in Korpora internetbasierter Kommunikation und in Korpora gesprochener Sprache häufig vorkommen und als typische Merkmale gesprochener bzw. konzeptionell mündlicher Sprache Gegenstand der Forschung sind. Die Erweiterungen für diese Phänomene differenzieren bestimmte, in STTS (1999) vorhandene Kategorien für die Zwecke der Annotation von IBK- und Gesprächskorpora weiter aus. Betroffen von diesen Modifikationen sind der Bereich der sog. „kontraktierten Formen“ (der eine Erweiterung um Tags für konzeptionell mündliche Formen erfährt), der Bereich der Partikeln (dem Kategorien für die Darstellung von Abtönungs-/Modalpartikeln, Intensitäts-/Fokus-/Gradpartikeln und Partikeln als Teilen von Mehrwort-Lexemen hinzugefügt werden)⁴ und der Bereich der Diskursmarker. Neu hinzugefügt wird außerdem eine Kategorie ONO für die Annotation von Onomatopoetika.

Die Abbildung auf S. 4 gibt eine Kurzübersicht über die Erweiterungen und Modifikationen in STTS_IBK gegenüber STTS (1999). Eine Übersicht über das komplette Tagset findet sich im Anhang dieses Dokuments.

4 Die Restrukturierung im Bereich der Partikeln orientiert sich an einem Vorschlag, der im Rahmen der Diskussion zur Erweiterung von STTS von Hagen Hirschmann, Nadine Lestmann, Ines Rehbein und Swantje Westpfahl für die gesprochene Sprache formuliert wurde und der u.a. am Institut für deutsche Sprache (Mannheim) bei der Annotation von Korpora gesprochener Sprache umgesetzt wird.

Tag	Kategorie	Beispiele
-----	-----------	-----------

I. Tags für IBK-spezifische Phänomene:

EMO ASC	Emoticon, als Zeichenfolge dargestellt (Typ „ASCII“)	:-) :-(^^ O.O
EMO IMG	Emoticon, als Grafik-Ikon dargestellt (Typ „Image“)	😊 😘, <i>kodiert als:</i> emojiQsmilingFaceWithSmilingEyes emojiQkissingCatFaceWithClosedEyes
AKW	Aktionswort	*lach*, freu, grübel, *lol*
HST	Hashtag	Kreta war super! <u>#urlaub</u>
ADR	Adressierung	<u>@lothar</u> : Wie isset so?
URL	Uniform Resource Locator	http://www.tu-dortmund.de
EML	E-Mail-Adresse	peterklein@web.de

II. Tags für Phänomene der konzeptionellen Mündlichkeit:

VV PPER	Tags für die häufigsten Bildungsmuster kontraktierter Formen (APPRART ist in STTS bereits vorhanden)	schreibste, machste
APPR ART		vorm, überm, fürn
VM PPER		willste, darfste, musste
VA PPER		haste, biste, isses
KOUS PPER		wenns, weils, obse
PPER PPER		ichs, dus, ers
ADV ART		son, sone
PTK IFG	Intensitäts-, Fokus- oder Gradpartikel	<u>sehr</u> schön, <u>höchst</u> eigenartig, <u>nur</u> sie, <u>voll</u> geil
PTK MA	Modal- oder Abtönungspartikel	Das ist <u>ja</u> / <u>vielleicht</u> doof. Ist das <u>denn</u> richtig so? Das war <u>halt</u> echt nicht einfach.
PTK MWL	Partikel als Teil eines Mehrwort-Lexems	keine <u>mehr</u> , <u>noch</u> mal, <u>schon</u> wieder
DM	Diskursmarker	<i>prototypisch: weil, obwohl, nur, also als Einheiten mit projektivem Potenzial im Vorvorfeld von V2-Sätzen</i>
ONO	Onomatopoetikon	boing, miau, zisch

3. Tags und PoS-Kategorien für IBK-spezifische Phänomene

3.1 Emoticons (EMOASC und EMOIMG)

„Klassische“, tastaturschriftlich erzeugte Emoticons werden typischerweise durch die Kombination von Interpunktions-, Buchstaben- und Sonderzeichen gebildet. In unterschiedlichen Kulturkreisen haben sich unterschiedliche Stile herausgebildet (z.B. westlicher, japanischer, koreanischer Stil), deren Verwendung aber nicht auf die jeweiligen Ursprungskulturen beschränkt geblieben ist. So sind in vielen deutschsprachigen Online-Communities neben den „klassischen“ Emoticons westlichen Stils inzwischen u.a. auch japanische Emoticons gebräuchlich.

Emoticons können am Ende eines Satzes bzw. einer satzwertigen kommunikativen Einheit oder in Form von Parenthesen auftreten sowie alleine eine kommunikative Äußerung realisieren. Sie werden u.a. zur emotionalen Kommentierung, zur Respondierung von Vorgängeräußerungen oder als Illokutions- und Ironiemarker verwendet.

Emoticons treten in verschiedenen Realisierungsformen auf:

- als tastaturschriftlich erzeugte und am Bildschirm als Schriftzeichenfolgen dargestellte Einheiten;
- als tastaturschriftlich erzeugte, vom verwendeten Kommunikationswerkzeug in Grafik-Icons umgewandelte und am Bildschirm als Grafik-Icons angezeigte Einheiten;
- als durch Auswahl aus einem Menü mit Grafik-Icons erzeugte und am Bildschirm als Grafik-Icons angezeigte Einheiten.

STTS_IBK unterscheidet Emoticons danach, wie sie am Bildschirm angezeigt werden, in

- Emoticons, die als Zeichenfolge dargestellt werden (EMOASC mit dem Tag-Bestandteil 'ASC' für 'ASCII');
- Emoticons, die als Grafik-Icon dargestellt werden (EMOIMG mit dem Tag-Bestandteil 'IMG' für 'Image'). In den Ausgangsdaten ist dieser Typ von Emoticons in Form von Zeichenfolgen kodiert, die auf <emojiQ...> beginnen, keinen Whitespace enthalten und die eine standardisierte, eindeutige Beschreibung des Grafik-Icons beinhalten, das an der betreffenden Stelle am Bildschirm bzw. Smartphone-Display angezeigt wurde. Der komplette auf <emojiQ...> beginnende Ausdruck ist als EMOIMG zu taggen.

Beispiele für Exemplare des Typs *EMOASC*:

- (1) *och, die fischbude am heumarkt is ok:-)*
- (2) *Mit mir will einfach keiner chatten!:((((*
- (3) *Ach nee, jetze isses plötzlich wieder eine Stadt? :-P*
- (4) *:-/ Nein, nicht wirklich. Na ja, aber was ist den der Sinn des ganzen?*
- (5) *Find ich echt super! \O/*
- (6) *Klar mein ich das ernst. ^^*

Beispiele für Exemplare des Typs *EMOIMG* (WhatsApp-Nachrichten):

- (7) Huhu! :) soll ich nachher noch irgendwas mitbringen?
emojiQsmilingFaceWithSmilingEyes
 ⇒ Darstellung im Display:
 Huhu! :) soll ich nachher noch irgendwas mitbringen? 😊
- (8) Ja, natürlich. Muss nur schauen wegen Uni. **emojiQkissingCatFaceWithClosedEyes**
 ⇒ Darstellung im Display:
 Ja, natürlich. Muss nur schauen wegen Uni. 😘

Bisweilen werden einzelne Bestandteile von Emoticons des Typs *EMOASC* von den Schreibern iteriert:

:-) ⇒ :-)) , :-)))))) ...
 :-(⇒ :-(, :-(...

3.2 Aktionswort (AKW)

Die Kategorie ‚Aktionswort‘ (AKW) umfasst Einheiten wie *grins*, *freu*, *lach*, *grübel*, *lol*, *rofl*, *stirnrunzel*, *malaufschreib*, die als selbstständige Einheiten der Interaktion fungieren. Prototypischerweise haben Aktionswörter die Form von einfachen Inflektiven (*grins*, *freu*, *lach*, *grübel*). Sie treten aber auch in der Form erweiterter Inflektive (*stirnrunzel*, *malaufschreib*) oder von Akronymen auf (*lol*, *rofl*). Bisweilen fungieren anstelle von Inflektiven auch Vertreter anderer Wortarten (**schock**) oder Verbformen in der 1. Person Singular als Basis (*beidirseinwill*). Häufig, aber nicht immer, sind Aktionswörter durch Asteriske markiert (**freu**, **grübel**, **lol**).

Bei komplexen Aktionswörtern tritt bisweilen Getrennschreibung auf (**vor mich hindämmer**, **gewissensbisse krieg**). In solchen Fällen wird lediglich der Inflektiv als AKW getaggt und werden die übrigen Teile entsprechend ihrer Zugehörigkeit zu anderen PoS-Kategorien behandelt.

Zusammengeschriebene komplexe Aktionswörter (*stirnrunzel*, *malaufschreib*) werden hingegen beim PoS-Tagging nicht künstlich in Tokens zerlegt, sondern als Ganze als Einheiten des Typs AW annotiert.

Asteriske oder sonstige Zeichen (z.B. Spitzklammern), die der Ein- und Ausleitung von Aktionswörtern dienen, werden bereits bei der Tokenisierung vom Wort abgetrennt. Das PoS-Tag AKW wird entsprechend nur für den sprachlichen Ausdruck vergeben.

3.3 Hashtags und Adressierungen (HST, ADR)

Hashtags und Adressierungen werden beim PoS-Tagging je nach Distribution unterschiedlich behandelt:

- a) Syntaktisch integrierte Verwendungen werden nach der PoS-Zugehörigkeit des sprachlichen Ausdrucks getaggt, mit dem das Thema (bei Hashtags) bzw. die Adressaten (bei Adressierungen) bezeichnet werden:

Ich war neulich im #urlaub \Rightarrow $\langle \#urlaub \rangle$ = NN

Ich habe @lothar getroffen \Rightarrow $\langle @lothar \rangle$ = NE

- b) Syntaktisch nicht integrierte Verwendungen werden mit spezifischen Tags (HST, ADR) ausgezeichnet:

Kreta war super! #urlaub \Rightarrow $\langle \#urlaub \rangle$ = HST

@lothar: Wie isset so? \Rightarrow $\langle @lothar \rangle$ = ADR

3.4 URLs und E-Mail-Adressen (URL, EML)

Für Tokens, mit denen eine URL angegeben wird, steht das Tag *URL* zur Verfügung. Tokens, mit denen eine E-Mail-Adresse angegeben wird, können mit dem Tag *EML* ausgezeichnet werden.

Volle URLs haben die folgende Struktur, wobei entweder Teil 1) oder Teil 3) wegfallen kann (aber nicht beide, andernfalls handelt es sich um bloße Domainnamen):

- 1) $\langle http:// \rangle$ oder $\langle https:// \rangle$
- 2) Angabe eines Domainnamens, der aus einer optionalen Subdomain (z.B. $\langle www. \rangle$), einem zentralen Namensbestandteil (z.B. $\langle spiegel-online \rangle$) und einer Top-Level-Domain-Endung (z.B. $\langle .de \rangle$) besteht

- 3) Angabe von Unterverzeichnis(sen) und Dateiname, z.B.

<http://www.spiegel.de/politik/deutschland/frank-walter-steinmeier-bereit-fuer-gauck-nachfolge-a-1051431.html>

Volle URLs werden grundsätzlich als URL klassifiziert, unabhängig davon, ob sie syntaktisch integriert sind oder nicht:

Schau mal hier: <http://www.spiegel.de/politik/deutschland/frank-walter-steinmeier-bereit-fuer-gauck-nachfolge-a-1051431.html> ⇒ URL

Lies dir mal <http://www.spiegel.de/politik/deutschland/frank-walter-steinmeier-bereit-fuer-gauck-nachfolge-a-1051431.html> *durch.* ⇒ URL

Begründung: URLs haben, wenn sie syntaktisch integriert sind, immer eine Doppelfunktion: Einmal wird mit ihnen der Adressat der Äußerung auf eine URL verwiesen, andererseits repräsentieren sie ein Element der syntaktischen Struktur. Im Kontext der Annotation von Daten aus Genres internetbasierter Kommunikation werden wir erstere Funktion höher als zweite.

Bloße Domainnamen wie z.B. <www.spiegel-online.de> oder <[spiegel-online.de](http://www.spiegel-online.de)> werden hingegen nur dann als URL getaggt, wenn sie syntaktisch nicht integriert sind. Im Falle syntaktisch integrierter Verwendungen werden sie entsprechend ihrer syntaktischen Funktion klassifiziert:

Schau mal hier: [spiegel.de](http://www.spiegel.de) ⇒ <[spiegel.de](http://www.spiegel.de)> = URL

Schau mal auf [spiegel.de](http://www.spiegel.de) ⇒ <[spiegel.de](http://www.spiegel.de)> = NE

Schau mal auf www.spiegel.de ⇒ <[spiegel.de](http://www.spiegel.de)> = NE

E-Mail-Adressen werden grundsätzlich als URL klassifiziert, unabhängig davon, ob sie syntaktisch integriert sind oder nicht; die Begründung ist dieselbe wie für volle URLs (s.o.)

Meine E-Mail: peter@schmitz.de ⇒ EML

Schreib mir bitte an die peter@schmitz.de, *nicht an die alte Adresse.* ⇒ EML

4. Tags und PoS-Kategorien für Phänomene der konzeptionellen Mündlichkeit

4.1 Kontraktierte Formen: Tags für die häufigsten Bildungsmuster

(APPRART, VPPER, VMPPER, VAPPER, KOUSPER, PPERPPER, ADVART)

STTS (1999) kennt keine Tags für *umgangssprachliche kontraktierte Formen*. Darunter verstehen wir Reduktionsformen wie *haste, biste, kannste, fürn, auf'm, wenns, weil's, obse, son, sone*, die beim Allegrosprechen in gesprochener Sprache rein koartikulationsbedingt gebildet werden und die in aller Regel frei für ihre Ausgangsformen – *hast du, bist du, kannst du* etc. – austauschbar sind. Umgangssprachliche kontraktierte Formen kommen auch in der schriftlichen internetbasierten Kommunikation vor und sind daher beim PoS-Tagging zu behandeln.

Für obligatorische kontraktierte Formen im Bereich der Präposition-Artikel-Verschmelzungen (*am, ans, im, zur, zum*) kennt STTS (1999) bereits die Kategorie APPRART. *STTS_IBK* erweitert STTS (1999) um Tags für sechs weitere Typen von kontraktierten Formen, von denen angenommen wird, dass sich mit ihnen die überwiegende Mehrheit der Vorkommen von kontraktierten Formen in der internetbasierten Kommunikation erfassen lässt. Die Benennung der Tags erfolgt analog zur Benennung des schon vorhandenen Tags APPRART und setzt sich aus den Kürzeln derjenigen PoS-Kategorien zusammen, aus denen die jeweilige kontraktierte Form gebildet ist.

Die Auswahl der Formtypen, für die *STTS_IBK* eigene Tags vorsieht, erfolgte auf Basis einer Auswertung der umgangssprachlichen kontraktierten Formen im Plauderchat-Teilkorpus des Dortmunder Chat-Korpus⁵. 92% aller Vorkommen im untersuchten Korpus entfallen auf sieben Bildungsmuster. Umgangssprachliche Fälle von Präposition-Artikel-Verschmelzungen werden in *STTS_IBK* mit der schon in STTS (1999) vorhandenen Kategorie APPRART erfasst. Für Formen, die nach den übrigen sechs der sieben häufigsten Bildungsmuster gebildet sind, stehen eigene Tags zur Verfügung:

Tag:	Kategorie (Bildungsmuster):	Beispiele:
APPRART	Präposition + Artikel	<i>vorm, überm, fürn, auf'm, mit'm</i>
VPPER	Vollverb + Personalpronomen	<i>schreibste, machste, kommste</i>
VMPPER	Modalverb + Personalpronomen	<i>willste, darfst, musste</i>

5 <http://www.chatkorpus.tu-dortmund.de>

VAPPER	Auxiliarverb + Personalpronomen	<i>haste, biste, isses</i>
KOUSPPER	unterordnende Konjunktion mit Satz + Personalpronomen	<i>wenns, weils, obse, dasste</i>
PPERPPER	Personalpronomen + Personalpronomen	<i>ichs, dus, ers</i>
ADVART	Adverb + Artikel	<i>son, sone</i>

Umgangssprachliche kontraktierte Formen, die sich *nicht* mit einem dieser Tags beschreiben lassen, sollen nach der PoS-Zugehörigkeit desjenigen Wortes getaggt werden das innerhalb der kontraktierten Form den Host stellt (z.B.: Negationspartikel+Adverb *nimmer* „nicht mehr“ ⇒ PTKNEG, Vollverb+Artikel *isn* „ist ein“ ⇒ VV).

4.2 Partikeln

In konzeptionell mündlichen Äußerungen sind verschiedene Typen von Partikeln hochfrequent, die in STTS (1999) nicht berücksichtigt sind, deren Darstellung für die Analyse (gesprochener wie geschriebener) dialogischer Interaktion aber von großem Interesse ist.

STTS_IBK erweitert die Subklassen von Partikeln aus STTS (1999) um Kategorien für die Annotation von Intensitäts-, Fokus- und Gradpartikeln (PTKIFG) und für die Annotation von Modal- und Abtönungspartikeln (PTKMA). Das bereits vorhandene Inventar an Partikelkategorien (PTKZU, PTKNEG, PTKVZ, PTKANT, PTKA) bleibt intensional wie extensional unangetastet. Durch die Einführung der beiden neuen Klassen ergibt sich allerdings ein veränderter Zuschnitt für die Klasse der Adverbien (ADV), die in STTS (1999) die hier unter PTKIFG und PTKMA gefassten Ausdrucksklassen mitumfasst.

PTKIFG und PTKMA sind als Sammelklassen zu verstehen, die nicht eine, sondern mehrere in der grammatischen Literatur (in z.T. unterschiedlichem Zuschnitt) unterschiedene Partikelklassen nach distributionellen Kriterien zusammenfassen: Intensitäts-, Fokus- und Gradpartikeln (PTKIFG) ist gemeinsam, dass sie Phrasen modifizieren und daher nicht alleine, wohl aber als Teil der Phrase, die sie modifizieren, ins Vorfeld verschoben werden können. Modal- und Abtönungspartikeln (PTKMA) lassen sich hingegen überhaupt nicht umstellen.

4.2.1 Intensitäts-, Fokus- und Gradpartikeln (PTKIFG)

Intensitäts-, Fokus- und Gradpartikeln weisen distributionell sehr ähnliche Eigenschaften auf: Sie bilden Teile von Phrasen und können typischerweise nicht alleine, sondern nur zusammen mit der gesamten Phrase ins Vorfeld verschoben werden. Intensitätspartikeln stehen *immer*, Fokus- bzw. Gradpartikeln *meist* vor ihrem Bezugsausdruck. Beim PoS-Tagging werden die beiden Klassen unter der Kategorie PTKIFG zusammengefasst.

Intensitätspartikeln:

Funktion, morphologische und syntaktische Charakteristik nach GRAMMIS⁶:

- Als Intensitätspartikeln bezeichnen wir eine Klasse von Partikeln wie *sehr* und *überaus*, die die von einem Adjektiv oder Adverb ausgedrückte Charakterisierung intensivierend-steigernd oder abschwächend-abstufend modifizieren: *überaus schön*, *kaum gefährlich*, *einigermaßen gern*.
- Intensitätspartikeln sind unflektiert, können keine Phrasen bilden und nicht im Vorfeld stehen.
- Der Bezugsausdruck der Intensitätspartikeln im engeren Sinne ist ein Adjektiv oder Adverb: *sehr glücklich*, *überaus gern*, *zu oft*. In wenigen Fällen ist der Bezugsausdruck auch ein Verb: *das schmerzt sehr*, *er leidet ziemlich*. Im Unterschied zu den Fokuspartikeln (*sogar die Katze*) ist der Bezugsausdruck aber nie ein Nomen.
- Im Gegensatz zu den Fokuspartikeln stehen Intensitätspartikeln unmittelbar vor dem modifizierten Ausdruck.

Typische Vertreter sind:

sehr, ausgesprochen, beileibe, einigermaßen, etwas, fast, kaum, nahezu, recht, überaus, ungemein, vollauf, weitaus.

Daneben gehören der Kategorie auch Ausdrücke adjektivischen Ursprungs an, die zur Intensivierung verwendet werden:

absolut, außerordentlich, außergewöhnlich, enorm, extrem, ganz, höchst, komplett, total, ungewöhnlich, völlig, weit, ziemlich, ...

6 Referenz hier und im Folgenden: Komponente „Systematische Grammatik“ in GRAMMIS 2.0 – Das grammatische Informationssystem des Instituts für deutsche Sprache (IDS). <http://hypermedia.ids-mannheim.de/call/public/sysgram.ansicht>

Fokus- bzw. Gradpartikeln:

Funktion, morphologische und syntaktische Charakteristik nach GRAMMIS:

- Mit Fokuspartikeln wie *sogar, bereits, nur, selbst* wird eine Einstufung des Gesagten bzw. bestimmter hervorgehobener Aspekte des Gesagten auf Skalen vorgenommen.
- Fokuspartikeln sind unflektiert, können keine Phrasen bilden, sind nicht selbständig verwendbar und können nicht allein im Vorfeld stehen.
- Fokuspartikeln stehen in der Regel vor (a), einige auch unmittelbar nach dem Bezugsausdruck (b). Auch Distanzstellung vom Bezugsausdruck kommt vor (c):

(a) Nur zwei Jahre muss er sitzen.

(b) Zwei Jahre nur muss er sitzen.

(c) Zwei Jahre muss er nur sitzen.

Beispiele:

allein, allenfalls, annähernd, auch, ausgerechnet, bereits, besonders, bestenfalls, bloß, einzig, erst, etwa, frühestens, gar, gerade, lediglich, mindestens, noch, nur, schon, selbst, sogar, spätestens, vor allem, wenigstens, zumindest.

4.2.2 Modal- und Abtönungspartikeln (PTKMA)

Unter der Kategorie der Modal- und Abtönungspartikeln (PTKMA) werden beim PoS-Tagging solche Ausdrücke zusammengefasst,

- die die Geltung einer Proposition einschränken und explizit wertend gebraucht werden können („Das kann man *ja/doch/fei/halt/eigentlich* nicht machen“, „Du bist *vielleicht* gerissen!“) oder
- die auf Erwartungen und Einstellungen der Adressaten zielen und dazu beitragen, Äußerungen in den jeweiligen Handlungszusammenhang zu integrieren (*halt, doch, nur, eben, denn*).

Morphologische und syntaktische Charakteristik nach GRAMMIS:

Abtönungs-/Modalpartikeln ...

- sind distributionell an das Mittelfeld im Satz gebunden. Dort werden sie je nach Fokus verschieden positioniert;

- können keine Phrasen bilden;
- können nicht durch w-Fragen erfragt werden;
- können untereinander kombiniert werden: *Du hast doch wohl nicht etwa Angst?*

Verwendungsbeispiele aus GRAMMIS:

- *Das war vielleicht eine Schweinerei!*
- *Möchtest du etwa in meiner Haut stecken?*
- *Wie heißt eigentlich dein Hund?*
- *Und man muss sich nur vor einem hüten, dass man eben dann wirklich sagt, alle Leut' sind blöd, die etwas über einen schreiben, denn es gibt halt auch die wahnsinnig Guten.*
(Jürgen von der Lippe 1995 in SDR 3 Leute)
- *Nun sei doch froh, dass wir hier in Ruhe frühstücken können.* (Marietta Meguid 1997 in SDR 3 Die Schwabensaga, 2. Staffel)
- *Wenn ich doch nur die Kraft hätte, Peter!* (Domenica 1994 in SDR 3 Leute)
- *Das kann man immer wieder beobachten eben, dass RTL eben dann eher mit der Katastrophe aufmacht und die politische Nachricht erst an zweiter Stelle bringt, und beim ARD und ZDF wäre es dann doch umgekehrt gewichtet.* (Petra Gerster 1998 in SWR1 Leute)
- *Am 21. Juni saß er in unserem Berliner Studio und hörte sich die Frage an, ob er damals, als Stasi-General, denn auch die Bundesrepublik besucht habe.* (Wolfgang Heim 1995 in SDR 3 Leute)

4.2.3 Partikeln als Teile von Mehrwort-Lexemen (PTKMWL)

Die Kategorie PTKMWL umfasst eine kleine Gruppe von Partikeln, für die charakteristisch ist, dass sie gemeinsam mit anderen Einheiten Mehrwort-Lexeme bilden, mit denen typischerweise Aspekt ausgedrückt wird. Den Kopf des Mehrwort-Lexems bildet ein Wort anderer Wortart (z.B. ein Adverb). Die Partikel modifiziert diesen Kopf nicht wie eine Intensitäts-, Fokus- oder Gradpartikel, sondern konstituiert in Einheit mit dem Kopf die Bedeutung des Mehrwort-Lexems. Einzelne Teile der Mehrwort-Konstruktion können unter Beibehaltung der Position des jeweils ande-

ren Teils nicht ohne Bedeutungsveränderung alleine ins Vorfeld verschoben werden. Zu PTKMWL gibt es homonyme Wörter in anderen Wortartenklassen.

Beispiel (Mehrwort-Lexem *immer noch*):

Baba ist immer noch brummelig.

→ * *Noch ist Baba immer brummelig.*

→ * *Immer ist Baba noch brummelig.*

Immer ist im Beispiel weder ein Adverb (= kann nicht alleine ins Vorfeld verschoben werden), noch hat es die Funktion eines Intensivierers. Vielmehr markiert es Aspekt zum Adverb *noch* (= dass der Zustand des Brummelig-Seins andauert).

Schwierige Fälle sind *schon* und *noch*, die homonym auch als Gradpartikeln, Adverbien und Abtönungspartikeln vorkommen:

Noch der dümmste Kopf kriegt es hin. ein Buch zu kaufen. (PTKIFG)

Noch haben wir Ferien. (Adverb)

Ich habe mir noch nie ein Buch gekauft.

* *Noch habe ich mir nie ein Buch gekauft. (PTKMWL)*

Ich fuhr nur 5 km/h zu schnell. Schon bei der Ampel haben sie mich rausgewunken. (PTKIFG)

Wir haben schon Ferien. (Adverb)

Dein Verhalten gestern war schon doof. (Abtönungspartikel, weil (a) keine temporale Lesart und (b) nicht alleine und ohne Bedeutungsveränderung ins Vorfeld verschiebbar)

Ich habe Brahms schon immer geliebt.

* *Schon habe ich immer Brahms geliebt. (PTKMWL)*

PTKMWL: Bestand und Beispiele:

auch noch, dazu noch, dann noch, doch noch, Zeitangabe + noch (z.B. in: *im Juli noch, nächstes Jahr noch, zuerst noch*), gerade noch, immer noch, immer mehr (PTKMWL + PIS), immer wieder, noch immer, keine mehr, nachher noch, nicht mehr, nichts mehr, x noch (im Sinne von „dazu“, z.B. in *den Pfeffer noch*), noch x (im Sinne von „dazu“, z.B. in *noch den Pfeffer*), noch ein/e/r, noch so, noch jemand, noch ein/mal, noch etwas, noch welche, noch zwei/drei/etc., noch dazu, noch mal, noch mehr, noch nie, noch + gesteigertes Adjektiv (z.B. in: *noch schlim-*

mer), nur mehr, nur noch, schon + gesteigertes Adjektiv (z.B. in: schon länger), schon mal, schon öfter/oft, heute schon, schon wieder, schon immer, immer schon, vorhin schon, erst mal, gerade erst, kaum erst, gar nicht erst, was/wohin/woher/wer/wie/wo (auch) immer (PWAV/PWS (ADV) + PTKMWL), Adjektiv + genug (z.B. in: *früh genug*, *alt genug*, *schnell genug*).

ABER: *nicht gerade* (PTKNEG PTKIFG), *viel mehr* X (PIS PIAT), *noch nicht* (zeitl.) (ADV PTKNEG), *auch mal* (ADV ADV).

4.3 Diskursmarker (DM)

Diskursmarker sind Einheiten, die im Vorvorfeld (oder: linken Außenfeld) von Sätzen stehen und die projizierende Funktion haben. Sie leiten keinen Nebensatz an, sondern schließen eine eigenständige Äußerung (prototypischerweise einen Verbzweitsatz) an das zuvor Gesagte an. Ihre Funktion ist die Verknüpfung von Diskurseinheiten. Zu den Diskursmarkern rechnen wir auch Fälle des „epistemischen *weil*“.

Diskursmarker können einfach und komplex sein. Im einfachen Fall liegt eine Einworteinheit vor, im komplexen Fall eine Mehrworteinheit. Prototypische Vertreter für einfache Diskursmarker sind *weil*, *obwohl*, *nur*, *also*. Beispiele für komplexe Diskursmarker sind *Ich mein* und *ehrlich gesagt*.

Bei der PoS-Annotation werden nur einfache Diskursmarker als solche ausgezeichnet. Im Falle von Mehrworteinheiten in der Funktion von Diskursmarkern werden die einzelnen Wort-Tokens entsprechend ihrer Zugehörigkeit zu anderen PoS-Kategorien behandelt.

Die einfachen Diskursmarker haben Homonyme in anderen Wortartenklassen:

- ***weil***: subordinierende Konjunktion (Einleitung von Kausalsätzen)
- ***obwohl***: subordinierende Konjunktion (Einleitung von Konzessivsätzen)
- ***nur***: Fokuspartikel, Abtönungspartikel
- ***also***: Adverb

***weil* und *obwohl* als Diskursmarker:**

Das Kriterium für das Vorliegen einer Verwendung der Ausdrücke ***weil*** und ***obwohl*** als Diskursmarker ist, dass der darauf folgende Satz keine Verbletzststellung aufweist (sondern typischerweise Verbzweitstellung):

- (a.) *Ich war gestern nicht in der Vorlesung, weil ich krank war.* (Subjunktor)
- (b.) *Ich war gestern nicht in der Vorlesung, weil ich war krank.* (DM)
- (c.) *Ich komme heute zur Vorlesung, obwohl ich krank bin.* (Subjunktor)
- (d.) *Ich komme heute zur Vorlesung. Obwohl – ich bin krank... Dann wohl eher doch nicht.* (DM)

nur als Diskursmarker:

Das Kriterium für das Vorliegen einer Verwendung des Ausdrucks **nur** als Diskursmarker ist, dass der Ausdruck (a) in Initialposition steht (Vorvorfeld bzw. linkes Außenfeld) und (b) dass der darauf folgende Satz nur dann Verberststellung aufweist, wenn es sich dabei um einen Frage-satz handelt:

- (e.) *Ich komme mit ins Kino. Nur diesmal suche ich den Film aus.* (DM)
- (f.) *Ich komme mit ins Kino. Nur: Diesmal suche ich den Film aus.* (DM)
- (g.) *Ich find das schon OK. Nur: Habt ihr euch schon mal überlegt, was das kostet?* (DM)
- (h.) *Ich find das schon OK, nur frage ich mich, was das Ganze soll.* (ADV)
- (i.) *Ich find das schon OK, ich frage mich nur, was das Ganze soll.* (ADV)
- (j.) *Nur Blonde kamen an diesem Abend in die Disco rein.* (Fokuspartikel am Satzanfang: „nur“ dient in diesem Fall nicht der Verknüpfung von Diskurseinheiten, sondern hat als Skopus das nachfolgende Nomen. Entsprechend steht „nur“ hier nicht im Vorvorfeld, sondern ist Teil einer Nominalphrase, die insgesamt das Vorfeld des Satzes besetzt.)

also als Diskursmarker:

Das Kriterium für das Vorliegen einer Verwendung des Ausdrucks **also** als Diskursmarker ist, dass der Ausdruck (a) in Initialposition steht (Vorvorfeld bzw. linkes Außenfeld) und (b) dass der darauf folgende Satz typischerweise Verbzweitstellung aufweist:

- (k.) *Ich fang dann mal an. Also (,) als ich neulich in die Klasse kam, da herrschte vielleicht ein Chaos!* (DM, Vorvorfeld)
- (l.) *Also ich sag mal so: Petra und Thomas mögen sich nicht besonders.* (DM, Vorvorfeld)
- (m.) *Wir können den Wagen heute Nachmittag drannehmen, Sie können ihn also gegen Abend abholen.* (Adverb im Mittelfeld)
- (n.) *Radio gab es damals noch nicht. Also mußten die Heilbronner warten, bis sie am nächsten Montag von dem Signal erfuhren.* (Adverb in Vorfeldposition)

- (o.) Also *willst du jetzt mit mir ins Kino oder nicht?* (Grenzfall; kann in der gesprochenen Sprache je nach Intonation entweder als Adverb oder als DM gewertet werden.)

Fälle wie (k.) und (l.), in denen also einen V2-Satz einleitet, sollen grundsätzlich als Diskursmarkerverwendungen getaggt werden (Achtung: Im Falle von (k.) ist der V2-Satz „*da herrschte vielleicht ein Chaos*“; „*als ich neulich in die Klasse kam*“ ist ein dem V2-Satz vorangestellter Adverbialsatz mit Verbletzstellung!).

Umstellprobe: Wenn sich ein initial stehendes also ins Mittelfeld verschieben lässt, liegt *kein* Diskursmarker vor. Ist Umstellung (ohne Bedeutungsveränderung) möglich, so handelt es sich um ein Adverb.

In Fällen wie (o.), in denen Verberststellung vorliegt, sollen in den IBK-Daten nur diejenigen Vorkommen von also als Diskursmarkerverwendungen gewertet werden, bei denen die Platzierung des also im Vorvorfeld durch typographische Markierung explizit angezeigt ist (Komma, Doppelpunkt oder Gedankenstrich):

- (o.1) Also, *willst du jetzt mit mir ins Kino oder nicht?*
(o.2) Also: *Willst du jetzt mit mir ins Kino oder nicht?*
(o.3) Also – *willst du jetzt mit mir ins Kino oder nicht?*

4.4 Onomatopoetikon (ONO)

Onomatopoetika – Nachahmungen von Schallereignissen mit lautlichen bzw. schriftlichen Mitteln – werden mit dem Tag ONO ausgezeichnet. Beispiele für Onomatopoetika sind *miau*, *kike-riki*, *platsch*, *plopp*, *boing*, *zisch* und *peng*.

5. Erwähnte Literatur



- Bartz, Thomas; Beißwenger, Michael; Storrer, Angelika (2013): Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. In: Journal for Language Technology and Computational Linguistics 28 (1) (Themenheft „Das STTS-Tagset für Wortartentagging – Stand und Perspektiven“, hrsg. v. Heike Zinsmeister, Ulrich Heid & Kathrin Beck), 157-198. http://www.jlcl.org/2013_Heft1/7Bartz.pdf
- Schiller, A./Teufel, S./Stöckert, Ch./Thielen, Ch. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technischer Bericht. Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. <http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf>
- Westpfahl, Swantje; Schmidt, Thomas (2013): POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. In: Journal for Language Technology and Computational Linguistics 28 (1) (Themenheft „Das STTS-Tagset für Wortartentagging – Stand und Perspektiven“, hrsg. v. Heike Zinsmeister, Ulrich Heid & Kathrin Beck), 193-153. http://www.jlcl.org/2013_Heft1/6Westpfahl.pdf
- Westpfahl, Swantje (2014): STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In: Lori Levin und Manfred Stede (eds.): Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, 1–10.

ANHANG: Vollständige Übersicht über das Tagset *STTS_IBK*

Blau hinterlegte Tabellenzeilen kennzeichnen die Erweiterungen gegenüber STTS (1999).

Tag	Beschreibung	Beispiele
ADJA	attributives Adjektiv	<i>[das] große [Haus]</i>
ADJD	adverbiales oder prädikatives Adjektiv	<i>[er fährt] schnell</i> <i>[er ist] schnell</i>
ADV	Adverb	<i>schon, bald, heute, jetzt</i>
APPR	Präposition, Zirkumposition links	<i>in [der Stadt], ohne [mich]</i>
APPRART	Präposition mit Artikel	<i>im [Haus], zur [Sache], vorm, überm, fürn</i>
APPO	Postposition	<i>[ihm] zufolge, [der Sache] wegen</i>
APZR	Zirkumposition rechts	<i>[von jetzt] an</i>
ART	bestimmter oder unbestimmter Artikel	<i>der, die, das, ein, eine</i>
CARD	Kardinalzahl	<i>zwei [Männer], [im Jahre] 1994</i>
FM	Fremdsprachliches Material	<i>[Er hat das mit“] A big fish [“übersetzt]</i>
ITJ	Interjektion	<i>mhm, ach, tja</i>
ONO	Onomatopoetikon	<i>boing, miau, zisch</i>
DM	Diskursmarker	prototypisch: <i>weil, obwohl, nur, also als Einheiten mit projektivem Potential im Vorfeld von V2-Sätzen</i>
KOUI	unterordnende Konjunktion mit „zu“ und Infinitiv	<i>um [zu leben]</i> <i>anstatt [zu fragen]</i>
KOUS	unterordnende Konjunktion mit Satz (VL-Stellung)	<i>weil, dass, damit wenn, ob</i>
KON	nebenordnende Konjunktion	<i>und, oder, aber</i>
KOKOM	Vergleichspartikel ohne Satz	<i>als, wie</i>
NN	Appellativa	<i>Tisch, Herr, [das] Reisen</i>
NE	Eigennamen	<i>Hans, Hamburg, HSV</i>
PDS	substituierendes Demonstrativpronomen	<i>dieser, jener</i>
PDAT	attributierendes Demonstrativpronomen	<i>jener [Mensch]</i>
PIS	substituierendes Indefinitpronomen	<i>keiner, viele, man, niemand</i>
PIAT	attributierendes Indefinitpronomen ohne Determiner	<i>kein [Mensch]</i> <i>irgendein [Glas]</i>

Tag	Beschreibung	Beispiele
PIDAT	attributierendes Indefinitpronomen mit Determiner	<i>[ein] wenig [Wasser] [die] beiden [Brüder]</i>
PPER	irreflexives Personalpronomen	<i>ich, er, ihm, mich, dir</i>
PPOSS	substituierendes Possesivpronomen	<i>meins, deiner</i>
PPOSAT	attributierendes Possesivpronomen	<i>mein [Buch], deine [Mutter]</i>
PRELS	substituierendes Relativpronomen	<i>[der Hund,] der</i>
PRELAT	attributierendes Relativpronomen	<i>[der Mann,] dessen [Hund]</i>
PRF	reflexives Personalpronomen	<i>sich, einander, dich, mir</i>
PWS	substituierendes Interrogativpronomen	<i>wer, was</i>
PWAT	attributierendes Interrogativpronomen	<i>welche [Farbe]</i>
PWAV	adverbiales Interrogativ- oder Relativpronomen	<i>warum, wo, wann worüber, wobei</i>
PAV	Pronominaladverb	<i>dafür, dabei, deswegen. trotzdem</i>
PTKZU	„zu“ vor Infinitiv	<i>zu [gehen]</i>
PTKNEG	Negationspartikel	<i>nicht</i>
PTKVZ	abgetrennter Verbzusatz	<i>[er kommt] an, [er fährt] Rad</i>
PTKANT	Antwortpartikel	<i>ja, nein, danke, bitte</i>
PTKA	Partikel bei Adjektiv oder Adverb	<i>am [schönsten], zu [schnell]</i>
PTKIFG	Intensitäts-, Fokus- oder Gradpartikel	<i>sehr [schön], höchst [eigenartig], nur [sie], voll [geil]</i>
PTKMA	Modal- oder Abtönungspartikel	<i>[Das ist] ja / vielleicht [doof] [Ist das] denn [richtig so?] [Das war] halt [echt nicht einfach]</i>
PTKMWL	Partikel als Teil eines Mehrwort-Lexems	<i>keine <u>mehr</u>, <u>noch</u> mal, <u>schon</u> wieder</i>
TRUNC	Kompositions-Erstglied	<i>An- [und Abreise]</i>
VVFIN	finites Verb, voll	<i>[du] gehst, [wir] kommen [an]</i>
VVIMP	Imperativ, voll	<i>komm [!]</i>
VVINF	Infinitiv, voll	<i>gehen, ankommen</i>
VVIZU	Infinitiv mit „zu“, voll	<i>anzukommen, loszulassen</i>
VVPP	Partizip Perfekt, voll	<i>gegangen, angekommen</i>
VAFIN	finites Verb, aux	<i>[du] bist, [wir] werden</i>
VAIMP	Imperativ, aux	<i>sei [ruhig!]</i>
VAINF	Infinitiv, aux	<i>werden, sein</i>

Tag	Beschreibung	Beispiele
VAPP	Partizip Perfekt, aux	<i>gewesen</i>
VMFIN	finites Verb, modal	<i>dürfen</i>
VMINF	Infinitiv, modal	<i>wollen</i>
VMPP	Partizip Perfekt, modal	<i>[er hat] gekonnt</i>
VVPPER	Kontraktion: Vollverb + irreflexives Personalpronomen	<i>schreibste, machste</i>
VMPPER	Kontraktion: Modalverb + irreflexives Personalpronomen	<i>willste, darfst, musste</i>
VAPPER	Kontraktion: Auxiliarverb + irreflexives Personalpronomen	<i>haste, biste, isses</i>
KOUSPPER	Kontraktion: unterordnende Konjunktion mit Satz (VL-Stellung) + irreflexives Personalpronomen	<i>wenns, weils, obse</i>
PPERPPER	Kontraktion: irreflexives Personalpronomen + irreflexives Personalpronomen	<i>ichs, dus, ers</i>
ADVART	Kontraktion: Adverb + Artikel	<i>son, sone</i>
EMOASC	Emoticon, als Zeichenfolge dargestellt (Typ „ASCII“)	<i>:-) :-(^^ O.O</i>
EMOIMG	Emoticon, als Grafik-Ikon dargestellt (Typ „Image“)	  , kodiert als: <i>emojiQsmilingFaceWithSmilingEyes</i> <i>emojiQkissingCatFaceWithClosedEyes</i>
AKW	Aktionswort	<i>*lach* freu, grübel *lol*</i>
HST	Hashtag	<i>[Kreta war super!] #urlaub</i>
ADR	Adressierung	<i>@lothar [: Wie isset so?]</i>
URL	Uniform Resource Locator	<i>http://www.tu-dortmund.de</i>
EML	E-Mail-Adresse	<i>peterklein@web.de</i>
XY	Nichtwort, Sonderzeichen enthaltend	<i>D2XW3</i>
\$,	Komma	<i>,</i>
\$.	Satzbeendende Interpunktion	<i>. ? ! ; :</i>
\$(sonstige Satzzeichen; satzintern	<i>- [] ()</i>