Empiri*ST* 2015

Shared Task of the *Empirikom*-Network for the automatic linguistic annotation of computer-mediated communication / social media

Tagset and guidelines for the PoS tagging of language data from genres of computer-mediated communication / social media

Michael Beißwenger ▪ Thomas Bartz ▪ Angelika Storrer ▪ Swantje Westpfahl

(Version: 13.09.2015 | Translation by Sabine Bartsch)

1. Status of this document

This document describes the part of speech tagset that forms the basis for PoS annotations in the shared task for automatic linguistic annotation of computer-mediated communication (CMC) (EmpiriST 2015). The tags and categories described here have been used for the manual annotation of the training and evaluation data released as part of the Shared Task.

The tagset is based on the *Stuttgart-Tübingen Tagset* (*STTS*, Schiller et al. 1999). In contrast to the canonical version of the STTS, it comprises extensions for elements typical of CMC as well as extensions and modifications for phenomena of „conceptually oral" language use that also occur in corpora of spoken language. For the latter, the tagset used in EmpiriST 2015 is compatible with STTS extensions introduced by Westpfahl/Schmidt (2013) and Westpfahl (2014) for the part of speech annotation of spoken language corpora.

The tagset was developed in 2012-13 in the context of the DFG scientific network „Empirische Erforschung internetbasierter Kommunikation" („Empirical research on internet-based communication", *Empirikom*)[1] and in the context of three CLARIN-D workshops for the extension of STTS.[2]

This document explicitly treats only those PoS categories that extend the canonical version of STTS (Schiller et al. 1999) or that represent modifications of existing categories. For those areas of STTS not affected by these extensions and modifications the guidelines set out in Schiller et al. (1999) are binding.

2. Overview of extensions and modifications relative to STTS (1999)

The STTS version for CMC or *internet-based communication* (abbreviated: STTS_IBK) described henceforth extends STTS (1999) by a set of specific tags for the representation of phenomena of CMC that are not covered by any of the categories in STTS (1999). Examples are: EMO, AKW, HST, ADR, URL, EML. There are, furthermore, tags for phenomena of conceptually oral language. Extensions treating these phenomena distinguish certain categories present in STTS (1999) for the purpose of annotating corpora of CMC and of spoken language. These modifications include the representation of contractions, of particles (by means of an extended set of categories for downtoners / modal particles, intensifiers / focus and gradation particles

1 http://www.empirikom.net

2 Thanks go to members and guests of the *Empirikom*-Netzwerk as well as the participants of the CLARIN-D-Workshops in Stuttgart (2012), Tübingen (2013) and Hildesheim (2013).

that are constituents of multi-word lexemes) [3] as well as the domain of discourse markers. In addition, a category ONO for the annotation of onomatopoetica has been added.

The following table gives an overview of these additional or extended categories over and above the STTS (1999) tag set. An overview of the complete tagset is given as an appendix to this document.

PoS tag	Category	Examples
I. Tags for phenomena which are specific for CMC / social media discourse:		
EMO ASC	ASCII emoticon	:-) :-(^^ O.O
EMO IMG	Graphic emoticon	
AKW	Interaction word	*lach*, freu, grübel, *lol*
HST	Hash tag	Kreta war super! #urlaub
ADR	Addressing term	@lothar: Wie isset so?
URL	Uniform resource locator	http://www.tu-dortmund.de
EML	E-mail address	peterklein@web.de
II. Tags for phenomena which are typical for spontaneous spoken language in colloquial registers:		
VV PPER	Tags for types of colloquial contractions which are frequent in CMC (APPRART is already existing in STTS 1999)	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	‚Intensitätspartikeln', Fokuspartikeln', ‚Gradpartikeln'	sehr schön, höchst eigenartig, nur sie, voll geil
PTK MA	Modal particles	Das ist ja / vielleicht doof. Ist das denn richtig so? Das war halt echt nicht einfach.
PTK MWL	Particle as part of a multi-word lexeme	keine mehr, noch mal, schon wieder
DM	Discourse markers	weil, obwohl, nur, also, ... *with V2 clauses*
ONO	Onomatopoeia	boing, miau, zisch

3 The restructuring in the domain of particles goes back to suggestions from Hagen Hirschmann, Nadine Lestmann, Ines Rehbein und Swantje Westpfahl for spoken language in the context of the aforementioned CLARIN-D workshops on the extension of STTS (1999).

3. Tags and PoS categories for phenomena specific to CMC

3.1 Emoticons (EMOASC and EMOIMG)

Keyboarded emoticons are typically combinations of punctuation marks, alphabetical characters and special characters. In different cultural spheres, different styles have evolved (e.g. the Western, Japanese or Korean style), whose usage is not confined to their original cultural domains. Thus, in many German online communities, classical emoticons are used alongside e.g. Japanese emoticons.

Emoticons can occur at the end of a sentence, as sentence-equivalent communicative units or as parentheses; they can also be used to represent a communicative act by themselves. They are used, among other things, for the purpose of emotional comment, in response to previous utterances or as illocution or irony markers.

Emoticons occur in different forms of realization:

- as keyboarded expressions displayed as characters on screen,
- as keyboarded expressions that are converted into graphic icons by the communication tool and displayed as such on screen,
- as a selection of graphical icons from a software menu displayed on screen as graphical icons.

STTS_IBK distinguishes emoticons according to the way they are displayed on screen into

- emoticons displayed as sequence of characters (EMOASC with the tag constituent ‚ASC' for ‚ASCII');
- emoticons displayed as graphic icons (EMOIMG with the tag constituent ‚IMG' for ‚Image'). In the source data, this type of icon is encoded as a character sequence starting with <emojiQ...>, contain no whitespace and comprise a standardized, unique description of the graphic icon that was displayed on the screen or smartphone display. The full expression beginning with <emojiQ…> must be tagged as EMOIMG.

Examples of the type E*MOASC* are:

(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.* ***^^***

Examples of the type *EMOIMG* (in WhatsApp messages) are:

(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. 😽

Occasionally, individual units of emoticons of the type EMOASC are iterated by the writers:

:-) ⇨ **:-))** , **:-))))))))** ...
:-(⇨ **:-((** , **:-(((((** ...

3.2 Action words (‚Aktionswort' ‚AKW)

The category ‚Aktionswort' (AKW) comrises units such as

grins, *freu*, *lach*, *grübel*, *lol*, *rofl*, *stirnrunzel*, *malaufschreib*.
(EN: grin, happy, laughing, lol, rofl, wondering, taking notes)

that function as independent units in the interaction. Prototypically, they occur in the form of simple inflectives (*grins*, *freu*, *lach*, *grübel*). They can also occur in the form of extended inflectives (*stirnrunzel*, *malaufschreib*) or as acronyms (*lol*, *rofl*). Occasionally, instead of inflectives other word classes (**schock**) or verb forms in the 1st person singular are used as a basis (*beidirseinwill*). Action wirds are frequently, but not always, marked by asterisks (**freu**, **grübel**, **lol**).

Some action words are written as complex words delimited by white space (**vor mich hindämmer**, **gewissensbisse krieg**). In these cases, only the inflective is tagged as AKW, the remainder of the expression is treated according to their regular PoS membership.

Action words spelled as consecutive sequences (*stirnrunzel*, *malaufschreib*) are not artificially divided into tokens for PoS tagging, but treated as units of type AKW.

Asterisks or similar characters (e.g. brackets <>), that are used to mark the beginning and end of action words, will already have been removed from the word in the tokenization process. The PoS tag AKW is only attached to the linguistic expression.

3.3 Hashtags and addressing terms (HST, ADR)

Hashtags and addressing terms are treated differently according to their distribution:

a) Syntactically integrated uses are tagged according to their PoS of the linguistic expression that refers to the topic (in the case of Hashtags) or the form of address (in the case of addressing terms:

Ich war neulich im #urlaub ⇨ <#urlaub> = NN
Ich habe @lothar getroffen ⇨ <@lothar> = NE

b) Syntactically non-integrated uses are tagged with specific tags (HST, ADR):

Kreta war super! #urlaub ⇨ <#urlaub> = HST
@lothar: Wie isset so? ⇨ <@lothar> = ADR

3.4 URLs and e-mail addresses (URL, EML)

For tokens giving an URL, the tag *URL* is used. Tokens giving an e-mail address are tagged as *EML*.

Full URLs have the following structure in which either part 1) or part 3) can be omitted (but never both; otherwise it is a domain name and not an URL):

1) <*http://*> or <*https://*>
2) Domain name consisting of an optional subdomain (e.g. <*www.*>), a central name element (e.g. <*spiegel-online*>) and a Top-Level-Domain extension (e.g. <*.de*>)
3) Subdirectories and file names, e.g. *http://www.spiegel.de/politik/deutschland/frank-walter-steinmeier-bereit-fuer-gauck-nachfolge-a-1051431.html*

Full URLs are classified as URL, regardless of whether they are syntactically integrated or not:

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

Rationale: When URLs are syntactically integrated, they invariably serve a double function: on the one hand, they are used to refer the addressee to the URL, on the other hand, they represent an element of the syntactic structure. In the context of the annotation of data from genres of computer-mediated communication the first function has priority over the second one.

Domain-names, such as <*www.spiegel-online.de*> oder <*spiegel-online.de*>, are only tagged as URL when they are not syntactically integrated. In cases of syntactic usage they are classified according to their syntactic function:

Schau mal hier: spiegel.de ⇨ <spiegel.de> = URL
Schau mal auf spiegel.de ⇨ <spiegel.de> = NE
Schau mal auf www.spiegel.de ⇨ <spiegel.de> = NE

E-mail addresses are always classified as EML, regardless of their syntactic function; the rationale is the same as for full URLs (see above):

Meine E-Mail: peter@schmitz.de ⇨ EML
Schreib mir bitte an die peter@schmitz.de, nicht an die alte Adresse. ⇨ EML

4. Tags and PoS categories for conceptually oral phenomena

4.1 Contracted forms: tags for the most frequent patterns (APPRART, VVPPER, VMPPER, VAPPER, KOUSPPER, PPERPPER, ADVART)

STTS (1999) does not have tags for colloquial contracted forms such as *haste, biste, kannste, fürn, auf'm, wenns, weil's, obse, son, sone*, which are realized in rapid spoken articulation due to co-articulation phenomena and are mostly freely substitutable for their original full forms – *hast du, bist du, kannst du* etc. Colloquial contracted forms occur also in written CMC and thus have to be treated in PoS tagging.

For obligatorily contracted forms in the domain of preposition-article-contractions (*am*, *ans*, *im*, *zur*, *zum*) STTS (1999) already has the category APPRART. *STTS_IBK* extends STTS (1999) by six types of tags for additional contracted forms. It is assumed that these cover the majority of occurrences of contracted forms in CMC. Tags are named in analogy to already existing tags of type APPRART. They are comprised of abbreviations of the PoS categories which form the basis for the respective contracted form.

Form types covered by STTS_IBK were selected on the basis of an analysis of colloquial contracted forms from a subcorpus of the Dortmunder Chat-Korpus[4]. 92% of all occurrences in the corpus under study are covered by the seven patterns given in the following table. Colloquial expressions of the pattern preposition-article contractions are already covered by the STTS category APPRART. For the remaining six forms the following tags have been defined:

Tag:	***category (pattern of formation):***	***Examples:***
APPRART	preposition + article	*vorm, überm, fürn, auf'm, mit'm*
VVPPER	full verb + personal pronoun	*schreibste*, *machste*, *kommste*
VMPPER	modal verb + personal pronoun	*willste*, *darfste*, *musste*
VAPPER	auxiliary verb + personal pronoun	*haste*, *biste*, *isses*
KOUSPPER	subordinating conjunction with sentence + personal pronoun	*wenns*, *weils*, *obse*, *dasste*
PPERPPER	personal pronoun + personal pronoun	*ichs*, *dus*, *ers*
ADVART	adverb + article	*son*, *sone*

Colloquial contracted forms that cannot be described by any of these tags, should be tagged according to the PoS membership of the word that serves as a host of the contracted form (e.g.: negation particle + adverb *nimmer* „nicht mehr" ⇨ PTKNEG, full verb + article *isn* „ist ein" ⇨ VV).

4.2 Particles

There are some high frequency particles in conceptually oral utterances that are not covered by STTS (1999). STTS_IBK introduces tags for the following particle classes:

- Categories for the annotation of intensifier, focus and gradation particles: PTKIFG
- Modal and downtoner particles: PTKMA
- Particles as parts of multi-word lexemes: PTKMWL

The class 'adverb' (ADV) has a changed coverage due to the new particle classes. The existing inventory of particle categories stays in place and can be looked up in the description of standard STTS (1999): PTKZU, PTKNEG, PTKVZ, PTKANT, PTKA.

4 http://www.chatkorpus.tu-dortmund.de

4.2.1 Intensifier, focus, gradation particles (PTKIFG)

These classes are distributionally very similar. They form parts of phrases and typically do not occur on their own and can only be moved to the pre-field (*Vorfeld)* with the entire phrase. Intensifier particles are always, focus and gradation particles mostly located in front of their reference expression. In PoS tagging, both classes are subsumed under the category PTKIFG.

Intensifier particles:

Functional, morphological and syntactic characteristics according to GRAMMIS[5]:

- intensifier particles such as s*ehr* and *überaus*, intensify or downtone a characterization made by an adjective or adverb: *überaus schön, kaum gefährlich, einigermaßen gern;*
- intensifier particles are not inflected, cannot form phrases and cannot occur in the pre-field;
- reference expression of these particles is an adjective or adverb: *sehr glücklich*, *überaus gern*, *zu oft*. In very few cases it is a verb: *das schmerzt sehr*, *er leidet ziemlich*. In contrast to focus particles (*sogar die Katze*) it is never a noun.
- In contrast to focus particles, intensifier particles occur immediately in front of the modified expression.

Typical examples are:

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

Also, expressions derived from adjectives, used as intensifier:

> *absolut, außerordentlich, außergewöhnlich, enorm, extrem, ganz, höchst, komplett, total, ungewöhnlich, völlig, weit, ziemlich, …*

Focus or gradation particles:

Functional, morphological and syntactic characteristics according to GRAMMIS:

- focus particles such as *sogar, bereits, nur, selbst* are used for scalar modification,

5 „Systematische Grammatik“ in GRAMMIS 2.0 – Das grammatische Informationssystem des Instituts für deutsche Sprache (IDS). http://hypermedia.ids-mannheim.de/call/public/sysgram.ansicht

- focus particles are uninflected, cannot form phrases, are not independently useable and cannot occur independently in the pre-field.
- focus particles typically occur (a) before, in some cases also (b) immediately after the reference expression. Distance from the reference expression also occurs (c):

  (a) *Nur zwei Jahre* muss er sitzen.
  (b) *Zwei Jahre nur* muss er sitzen.
  (c) *Zwei Jahre* muss er *nur* sitzen.

Examples:

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 and downtoner particles (PTKMA)

The category PTKMA subsumes expressions

- that limit the scope of the proposition („Das kann man *ja/doch/fei/halt/eigentlich* nicht machen“, „Du bist *vielleicht* gerissen!“), or
- that refer to (assumed) expectations and attitudes of the addressees and integrate the utterance into the context of the interaction (*halt, doch, nur, eben, denn*).

Morphological and syntactic characteristics according to GRAMMIS:

Downtoners / modal particles ...

- are distributionally bound to the middle field (*Mittelfeld*) of the sentence. There, they are positioned according to their focus;
- cannot build phrases;
- cannot be queried by means of w-questions;
- can be combined with each other: *Du hast doch wohl nicht etwa Angst?*

Examples from *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 Particles as parts of multi-word lexemes (PTKMWL)

The category PTKMWL covers a small group of particles that form multi-word lexemes with other lexical units which are typically used for the expression of aspect. The head of the multi-word lexeme is a word of another word class (e.g. an adverb). The particle modifies the head, but unlike an intensifier, focus or gradation particle, it constitutes the meaning of the multi-word lexeme in combination with the head. The individual constituents of the multi-word construction cannot be moved to the pre-field by themselves without altering the meaning of the expression. Homonyms of PTKMWL tokens exist in other word classes:

Example (multi-word lexeme *immer noch*):

Baba ist immer noch brummelig.
→ * *Noch ist Baba immer brummelig.*
→ * *Immer ist Baba noch brummelig*.

Immer is neither an adverb in this example (= cannot be moved to the pre-field) nor does it serve the function of an intensifier. It marks aspect of the adverb *noch* (= an ongoing state of being 'brummelig' ('*in a bad mood'*)).

Difficult cases are *schon* and *noch* that are also homonyms of gradation particles, adverbs and downtoners:

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. (Downtoner particle, because (a) no temporal reading and (b) cannot be moved to the pre-field on its own without a change of meaning)

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

PTKMWL:Examples:

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 schlimmer*), 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*).

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

4.3 Discourse markers (DM)

Discourse markers are units that occur in the pre-field (*Vorfeld*) (or the left periphery) of sentences and have projecting function. They do not introduce subordinate clauses, but connect an expression to the previous context (prototypically a sentence with verb second ordering). They serve as connectors of discourse units. We also count cases of epistemic "weil" among the discourse markers.

Discourse markers can be simple or complex:

- simple: *weil, obwohl, nur, also.*
- complex: *Ich mein* und *ehrlich gesagt*.

In the simple case, they are constituted by a single lexical item, in the complex case they are constituted by a multi-word lexical units. Prototypical cases of discourse markers are *weil*, *obwohl*, *nur*, *also*. Examples for complex discourse markers are *Ich mein* und *ehrlich gesagt*.

In the PoS annotation, only simple discourse markers are annotated. In the case of multi-word lexemes functioning as complex discourse markers the individual word tokens are tagged according to their PoS class.

Simple discourse markers have homonyms in other word classes:

- ***weil:*** subordinating conjunction (introducing causal sentences)
- ***obwohl:*** subordinating conjunction (introducing concessive sentences)
- ***nur:*** focus particle, downtoners
- ***also:*** adverb

***weil* and *obwohl* as discourse markers:**

The criterion for ***weil*** and ***obwohl*** in their function as discourse markers is that the following sentence does not display verb last order (but, instead, typically V2 (= verb second)):

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

***nur* as discourse marker:**

The criterion for ***nur*** as discourse marker is that the expression (a) occurs in initial position (pre-pre-field, or left periphery), and (b) that the following sentence can only display verb first order when it is a question:

(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.* (Sentence initial focus particle: the function of „nur“ in this case is not that of a connector of discourse units, but its scope extends over the following noun. Accordingly, “nur” is not in the pre-pre-field, but part of a noun phrase which constitutes the pre-field of the sentence.)

***also* as discourse marker:**

The criterion for ***also*** as discourse marker is (a) that it occurs in initial position (pre-pre-field or left periphery),and (b) that the following sentence typically displays verb second order:

(k.) *Ich fang dann mal an. Also (,) als ich neulich in die Klasse kam, da herrschte vielleicht ein Chaos!* (DM, pre-pre-field)

(l.) *Also ich sag mal so: Petra und Thomas mögen sich nicht besonders.* (DM, pre-pre-field)

(m.) *Wir können den Wagen heute Nachmittag drannehmen, Sie können ihn also gegen Abend abholen.* (advber in middle field)

(n.) *Radio gab es damals noch nicht. Also mußten die Heilbronner warten, bis sie am nächsten Montag von dem Signal erfuhren.* (adverb in pre-field position)

(o.) *Also willst du jetzt mit mir ins Kino oder nicht?* (borderline case, in spoken language can be either classified as an advber or a DM depensing on the intonation.)

Cases such as (k.) and (l.), where *also* opens a V2 sentence should be tagged as a discourse marker.

Permutation test: When an initial *also* can be moved to the mid field, it is NOT a discourse marker. When permutation is possible (without change of meaning), it is an adverb.

In cases such as (o.), where the verb is in initial position, only those occurrences should be tagged as discourse marker, where the pre-pre-field (*Vorvorfeld*) position is typographically indicated (comma, colon or hyphen):

(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 Onomatopoetica (ONO)

Forms of sound imitation by phonetic or graphemic means are tagged as ONO (onomatopoeicon). Examples: *miau, kikeriki, platsch, plopp, boing, zisch* und *peng*.

5. References

- Bartz, Thomas; Beißwenger, Michael; Storrer, Angelika (2013): Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvor-schläge. 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 For-schungs- 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.

APPENDIX: Complete Tagset *STTS_IBK*

Cells with blue background color represent extensions to STTS (1999).

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

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

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