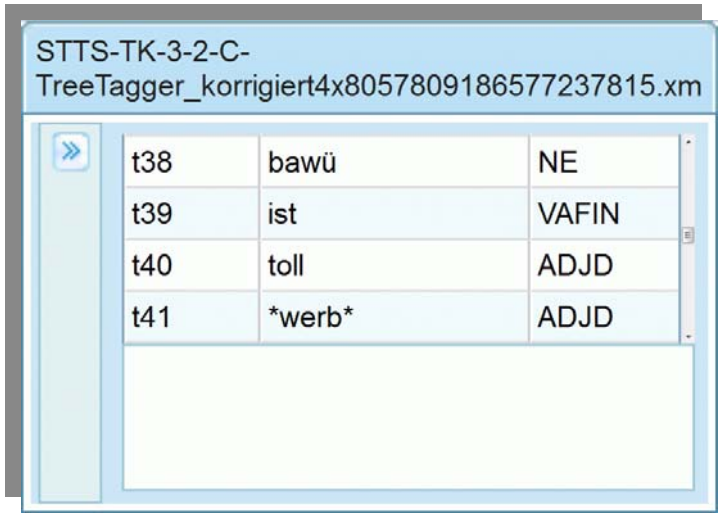
Überlegungen zur Modifikation und Erweiterung von STTS für das Tagging von Korpora zur internetbasierten Kommunikation


STTS-TK-3-2-C-TreeTagger_korrigiert4x8057809186577237815.xm

t38	bawü	NE
t39	ist	VAFIN
t40	toll	ADJD
t41	*werb*	ADJD

Thomas Bartz
Michael Beißwenger
Angelika Storrer

technische universität dortmund

CLARIN-D-Workshop:

Das STTS-Tagset für Wortartentagging: Stand und Perspektiven

Stuttgart, 24. September 2012

Interaktionsorientiertes Schreiben in der internetbasierten Kommunikation (IBK)

Leitlinie bei der **Gestaltung schriftlicher Kommunikationsbeiträge in dialogischer internetbasierter Kommunikation** ist weniger die situationsunabhängige Verständlichkeit des sprachlichen Produkts als vielmehr der kommunikative Erfolg der damit realisierten sprachlichen Handlung(en) im Kontext der laufenden Interaktion. Die sprachliche Form ist optimiert für Adressaten, die die im Kommunikationsgeschehen vorausgegangenen Schritte (Äußerungen) kennen und über den aktuellen Stand des Geschehens auf dem Laufenden sind. Die Planung und Versprachlichung geschieht häufig schnell und spontan.

⇨ **Typische Merkmale des *interaktionsorientierten Schreibens* im Internet:**

- Schnellschreib-Phänomene (Tippfehler)
- sprachliche Ökonomie: liberaler Umgang mit orthographischen Normen, die auf eine Verständnissicherung in der Distanzkommunikation hin optimiert sind (z.B. GKS, Interpunktion); Akronyme
- Orientierung am Duktus der gesprochenen Umgangssprache (Lexik, Syntax)
- „Verschriftete Umgangssprache“: Verschriftungen, die sich an der umgangssprachlichen Lautung anstatt am schriftlichen Standard orientieren
- Häufige Verwendung innovativer semiotischer und sprachlicher Formen, die sich in der IBK als Mittel zur emotionalen und evaluativen Kommentierung, zur Kohärenzsicherung in dialogischer Interaktion und zum spielerischen Rekurs auf Körperlichkeit herausgebildet haben (Emoticons, Aktionswörter, Adressierungsausdrücke)

Für welche Bereiche ist ein POS-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation von Interesse?

1. für die korpusgestützte Analyse sprachlicher Besonderheiten in der internetbasierten Kommunikation (IBK): Automatische linguistische Annotation verbessert die Möglichkeiten der qualitativen und quantitativen Analyse;
2. für die korpusgestützte Analyse aktueller Tendenzen in der deutschen Gegenwartssprache: Die Integration von IBK-Daten bzw. -Teilkorpora in annotierte Korpora zur deutschen Gegenwartssprache ermöglicht Untersuchungen zum Sprachwandel *durch* IBK;
3. für alle, die in Linguistik, Computerlinguistik und Informatik mit linguistisch aufbereiteten Webkorpora arbeiten (und dabei auch mit IBK-Phänomenen umgehen müssen).

DeRiK – Deutsches Referenzkorpus zur internetbasierten Kommunikation

KobRA – Korpusbasierte linguistische Recherche und Analyse mit Hilfe von Data-Mining (BMBF)

empirikom
DFG-Netzwerk „Empirische Erforschung internetbasierter Kommunikation“
http://www.empirikom.net

POS-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation: Tests mit gängigen Taggern

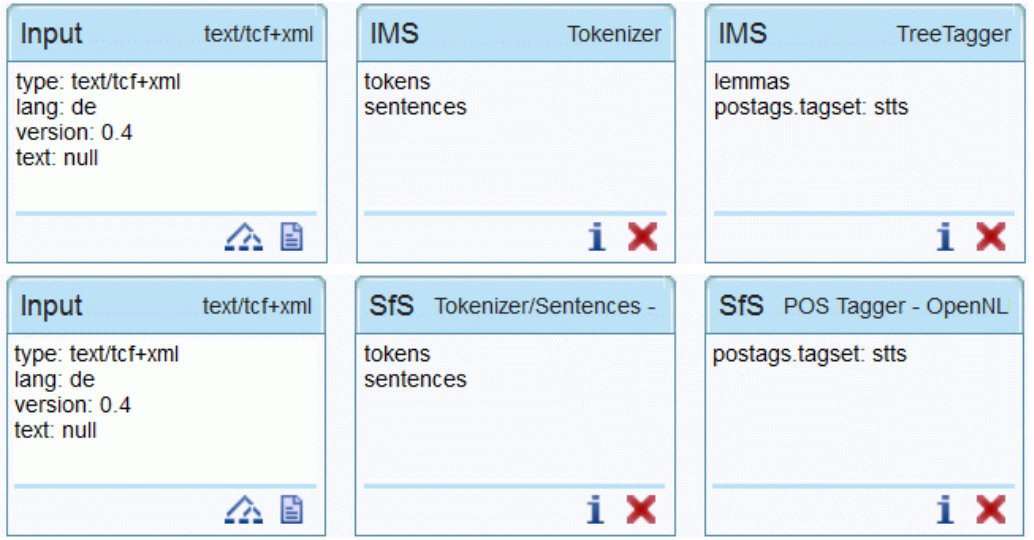
Testdatenset mit Belegen für ausgewählte Phänomene IBK-spezifischer Sprachverwendung

Phänomentyp	*Wikipedia-Diskussion*	*Chat*	*DWDS*
Verschriftete Umgangsprache I: Wortschreibung	20	20	(20)
IBK-typische oder nicht konventionalisierte Akronyme	20	20	
Verschriftete Umgangssprache II: Kontraktive Formen (VVFIN/VAFIN/VMFIN + PPER)	20	20	
IBK-spezifische Elemente I: Emoticons	20	20	
IBK-spezifische Elemente II: Aktionswörter	20	20	
Postings Gesamt:	**100**	**100**	
	200		

CLARIN-D | WebLicht – Web-Based Linguistic Chaining Tool

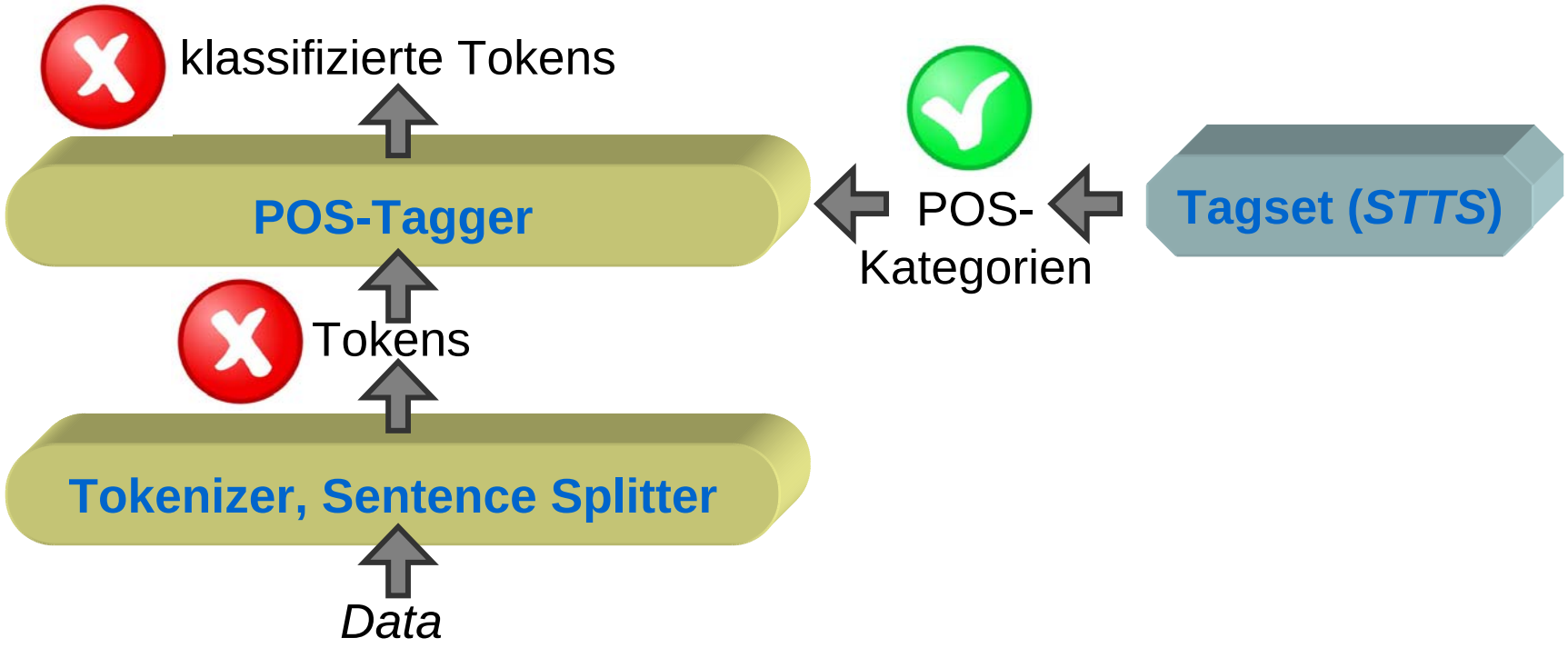
Toolchain 1: Kombinierter Tokenisierer und Satzgrenzenerkenner + TreeTagger des IMS

Toolchain 2: Kombinierter Tokenisierer und Satzgrenzenerkenner + Tagger aus dem OpenNLP-Projekt (SfS)



Problemtyp I: Tokenisierungs-Problem: Die Daten lassen sich in Tokens zerlegen, zu denen es Kategorien im Tagset (STTS) gibt. Der Tokenisierer liefert aber Tokens, die so segmentiert sind, dass sie sich nicht sinnvoll klassifizieren lassen.

⇨ Grund: Nicht-standardkonforme Verwendung von Spatien und Interpunktionszeichen (bei der Wort- und Satzschreibung).



Problemtyp I: Tokenisierungs-Problem: Die Daten lassen sich in Tokens zerlegen, zu denen es Kategorien im Tagset (STTS) gibt. Der Tokenisierer liefert aber Tokens, die so segmentiert sind, dass sie sich nicht sinnvoll klassifizieren lassen.

⇨ Grund: Nicht-standardkonforme Verwendung von Spatien und Interpunktionszeichen (bei der Wort- und Satzschreibung).

wieso stoeps?biste losgerannt einkaufen udn ahst vergessen dich anzuziehen vorher?*G*

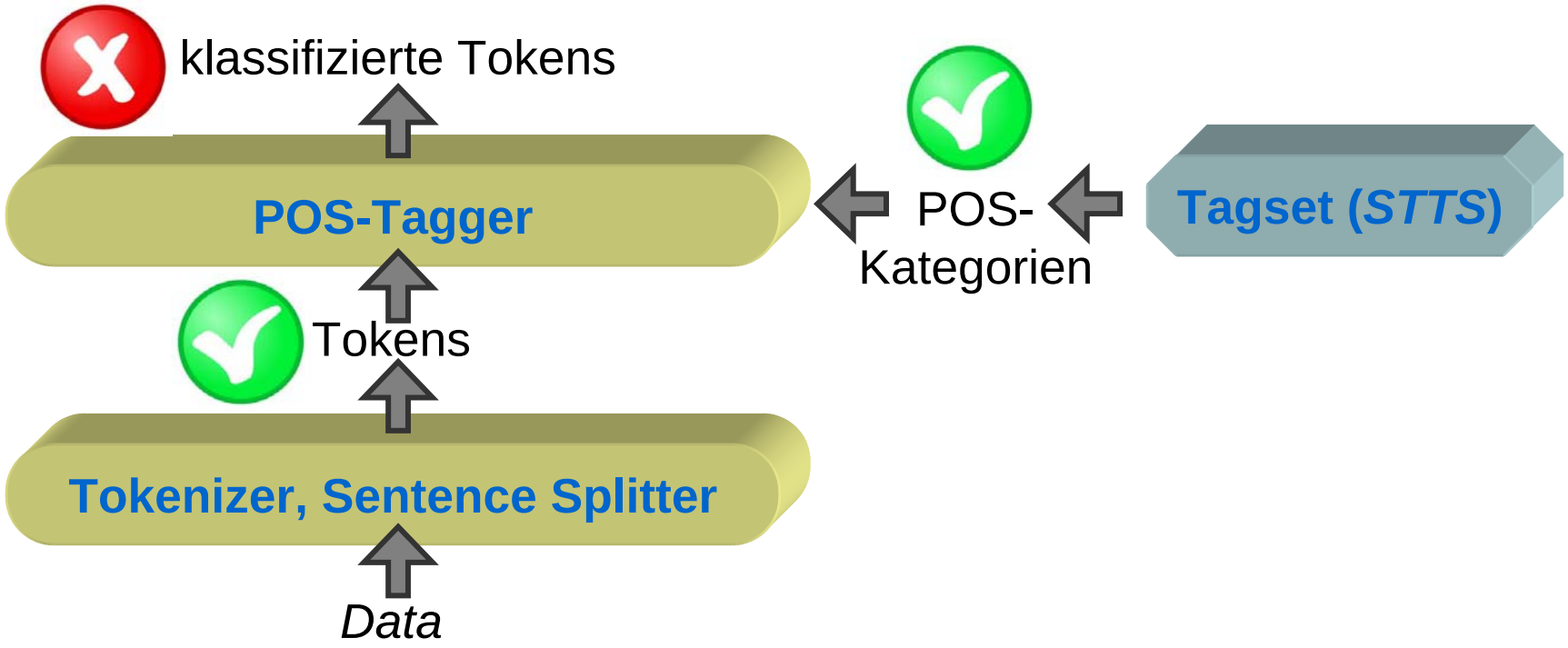
Dortmunder Chat-Korpus, Dok. Nr. 2221007

```
<token ID="t154">wieso</token>
<token ID="t155">stoeps?biste</token>
<token ID="t156">losgerannt</token>
<token ID="t157">einkaufen</token>
<token ID="t158">udn</token>
<token ID="t159">ahst</token>
<token ID="t160">vergessen</token>
<token ID="t161">dich</token>
<token ID="t162">anzuziehen</token>
<token ID="t163">vorher?*G*</token>
```

```
<token ID="t154">wieso</token>
<token ID="t155">stoeps</token>
<token ID="t156">?</token>
<token ID="t157">biste</token>
<token ID="t158">losgerannt</token>
<token ID="t159">einkaufen</token>
<token ID="t160">udn</token>
<token ID="t161">ahst</token>
<token ID="t162">vergessen</token>
<token ID="t163">dich</token>
<token ID="t164">anzuziehen</token>
<token ID="t165">vorher?*G*</token>
```

Problemtyp II: Klassifizierungs-Problem: Die Daten lassen sich in Tokens zerlegen, zu denen es Kategorien im Tagset (STTS) gibt. Der Tokenisierer segmentiert korrekt, der Tagger kann die Tokens aber nicht als Vertreter der vorhandenen Kategorien identifizieren.

⇨ tritt auf z.B. bei nicht-standardkonformen (an der umgangssprachlichen Lautung orientierten oder kreativen) Wortschreibungen sowie bei nicht konventionalisierten/okkasionellen Akronymen.



Verschriftete Umgangssprache: Wortschreibung

DWDS	Wikipedia-Diskussionen	Chat
Ja, Meg Ryan habe tatsächlich ...	**Jut**, ich find die Variante mit ...	**jo**, mach das mal...
... **ja**, der Brownie koste zwei Dollar.	**Jo**, gute Vorbereitung ist ...	**japp** tom, stimmt. ...
... Prinzipiell **ja**, auch wenn ...	**Joh**, da hast Du sicher nicht ...	**jepp** zora, das bin ich ;)
... **ja**, in ihm offenbare sich der ...	**Jap**, geht klar!	**jau** das auto fährt ...
Goethe-Jahr? Aber **nein**: eine...	**Jupp**, aber Hinweise zu ...	**nope**,die 10000 gesamt sind ...
... dann sage ich: **nein**.	**Nee** dann müsste ich ja ...	@quaki, **nee**,bin ...
... **Nein**, nein: Der normale ...	Ach **nee**, jetze isses ...	**nööö** is er nich
... **Okay**, okay, sie ist ein ...	**Nö**, hat er nicht mehr ;-)	**nö**,dat ebste findeste ...
..., droht **jetzt** der Bankrott. ...	**Nööö** (Zitat Benutzer:Orientalist)...	**oki**...mach`s gut
Gut machst du **das**!, ruft ...	**okidoki**, sag Bescheid, wenn du ...	nö,**dat** ebste findeste ...
... gefroren ist, **das** ist schon ...	Ach nee, **jetze** isses plötzlich ...	**dat** ist donald duck
Als meine Augen wieder **gucken**...	Um Gottes Willen, geh **fott** mit ...	... einfach **guckst** was da ist ...
... Die Rose verblüht ihm **nicht**.	**Weia**, Augenkrebs hoch drei. ...	**tach** tomcat
..., in der Künstlerkolonie Worpswede **nicht** genug ...	**Tach** Wurm, geh mich doch ...	**nööö** is er nich
... mit dieser Spende **nichts** zu tun...	... geh mich doch **fott** mit ...	ich mag **net** wissen wie ...
... Bergtouren **nichts** anderes als ...	**Guck** Dir genau den kompletten ...	und sagt **nix**, der sack
Darum kann **ich** es ...	Hehe, **sowatt** kütt vüür ...	...kann man hier **nischt** mehr ...
..., wenn **ich** täglich einige...	Hehe, sowatt **kütt vüür** ...	... **isch** hab bestanden
... Mein richtiger **Vater** war ...	**Isch** ja gut, es hier noch ...	mach **isch** glatt :)
..., **aber** auch mit großem Aufwand...	... mit **Vadder** is hier Kim Il Sung...	ich auch **aba** bei mir ...

Verschriftete Umgangssprache: Wortschreibung

Korrekt klassifizierte Beleg-Instanzen...	*TreeTagger*	*POS Tagger OpenNLP*
... aus dem **DWDS-Korpus:**	**18** (20)	**15** (20)
... aus **Wikipedia-Diskussionsseiten:**	**1** (20)	**1** (20)
... aus **Chats:**	**2** (20)	**3** (20)

VVIMP: Guck Dir genau den kompletten Vereinsnamen an.

VVFIN: ... wenn du gar nich suchst sondern einfach guckst was da ist...

PIS: und sagt nix, der sack

PTKNEG: nöö is er nich

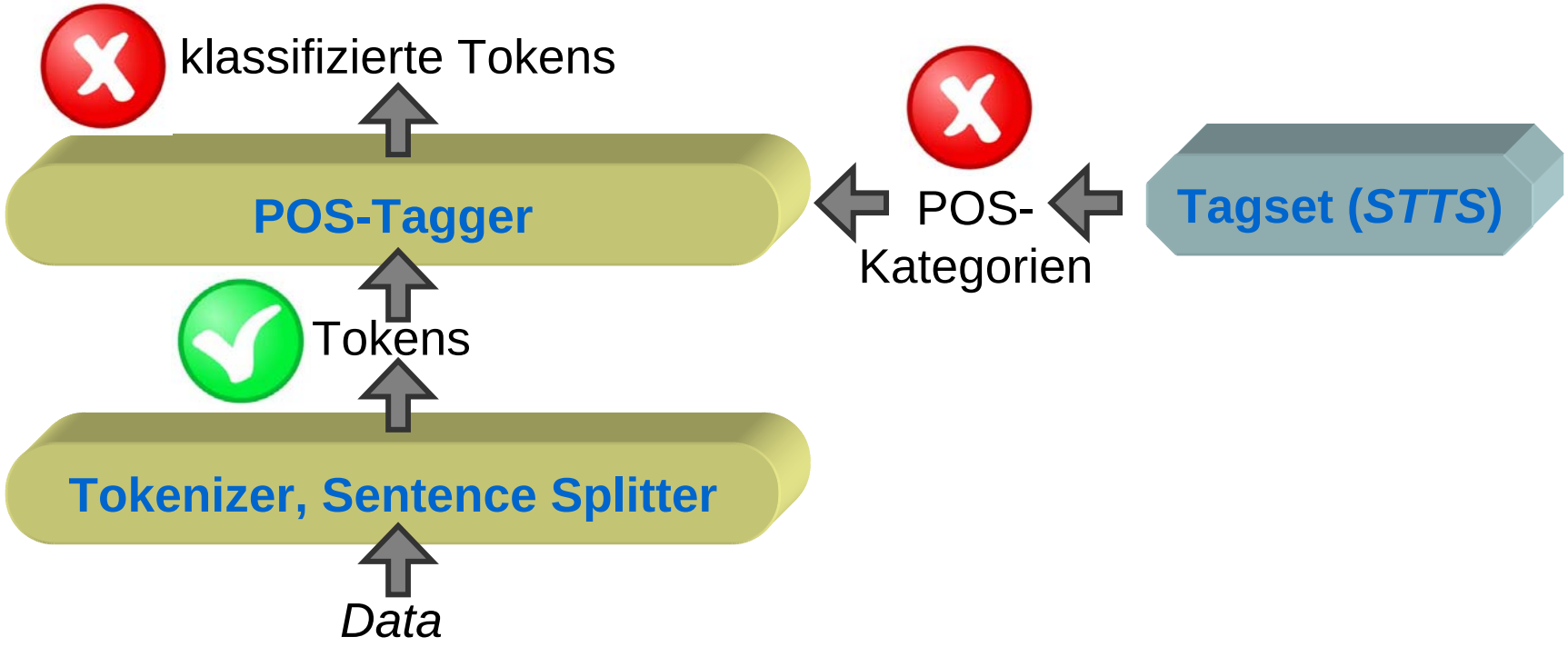
VVFIN: sowatt kütt vüür

Vergleichbare Ergebnisse liefern die Tests zum Tagging IBK-typischer bzw. nicht-knoventionalisierter Akronyme:
IMHO, bspw., b.t.w., Btw., vllt, evt., mE, zB, Thx, jmd, LG, POV, ...

Problemtyp III: Kategorien-Problem: Der Tokenisierer segmentiert korrekt, der Tagger kann die Tokens aber nicht sinnvoll klassifizieren, da es im Tagset (STTS) keine Kategorien dafür gibt.

⇨ tritt auf in folgenden Fällen:

(1) Tokens sind keine (oder keine prototypischen) *Wort*-Tokens;
(2) Tokens gehören zu Kategorien, die erst noch an existierende POS-Einteilungen anzubinden sind.



Verschriftete Umgangssprache: Kontraktive Formen

„Es ist bis jetzt (aus technischen Gründen) nicht möglich, Mehrwortlexeme als Ganzes zu taggen, oder kontraktive Formen mit einer Kombination aus mehreren Tags zu versehen.“ (STTS-Guidelines: 9) Gerade in der internetbasierten Kommunikation sind kontraktive Formen aber häufig anzutreffen.

Beispiel: **Kontraktive Formen des Typs VVFIN / VAFIN / VMFIN + PPER (+ PPER)**

haste, biste, findeste, könnteste, magste, meinste, denkste, machste, machstes, isses, hats, kanns, kenns, gehts, habs, sags, schreibs, machs, machts, wärs, wirds

	Wikipedia-Diskussionen:		Chat:	
Beleg-Instanzen klassifiziert als...	*TreeTagger*	*OpenNLP*	*TreeTagger*	*OpenNLP*
VVFIN / VAFIN:	**8** (20)	**7** (20)	**7** (20)	**10** (20)
NN:	**8** (20)	**1** (20)	**6** (20)	**0** (20)
ADJA / ADJD:	**4** (20)	**5** (20)	**7** (20)	**4** (20)
ADV:	**0** (20)	**0** (20)	**0** (20)	**3** (20)
andere:	**0** (20)	**7** (20)	**3** (20)	**3** (20)

Emotikons

:)
(:
:-)
:-))
:-)))
:o)
;-)
;-)))))
:(
:-(
:O)
:P
:-P
8)
=o)
^^
-.-
o_O
O-O

Tokenisierung/Tagging, 1. Durchgang (IMS-Tokenisierer + TreeTagger)**:**
27 von 40 Emoticon-Tokens inkorrekt segmentiert; Tokens entsprechend uneinheitlich klassifiziert.

```
<token ID="t99">:-</token>
<token ID="t100">)</token>
<token ID="t101">)</token>
```

Tagging, 2. Durchgang (TreeTagger) nach manueller Normalisierung der Tokenisierung:

POS-Tag	Anzahl
NN	20
ADJD	12
ADJA	6
NE	1
VVFIN	1
Gesamt	**40**

Auffällig: Nicht vergeben werden die Kategorien

- **XY** („Nichtwort“, auch für größere Symbolgruppen oder Kombinationen aus Ziffern und Zeichen, die sich nicht als CARD oder ADJA einordnen lassen)
- **ITJ** (Interjektion – obwohl den Emoticons positional und funktional ähnlich *... dazu gleich noch mehr*)

„Aktionswörter“ (Inflektive und Inflektivkonstruktionen)

freu
lach
lächel
grins
fiesgrins
wink
Gähn
Seufz
werb
wunder
stotter
rotwerd
einrück
lol
LOL
lol
rofl
Grummel
kopfschüttel
duck
g
ggg
lernenmuss
*feuerzeug an
reb weiterreich*

Tokenisierung/Tagging, 1. Durchgang:
automatische Tokenisierung; die meisten Aktionsausdrücke werden dabei zusammen mit den Asterisken als *ein* Token behandelt und getaggt.

Tokenisierung/Tagging, 2. Durchgang:
manuelle Normalisierung der Tokenisierung: Eliminierung aller Asterisken; auch Mehrwort-Ausdrücke mit Spatien werden als *ein* Token ausgewiesen. Automatisches Tagging der Postings mit den normalisierten Tokens.

Tagging-Ergebnis Durchgang 1:

POS-Tag	Anzahl
NN	20
ADJD	13
ADJA	4
NE	2
VVIMP	1
Gesamt	**40**

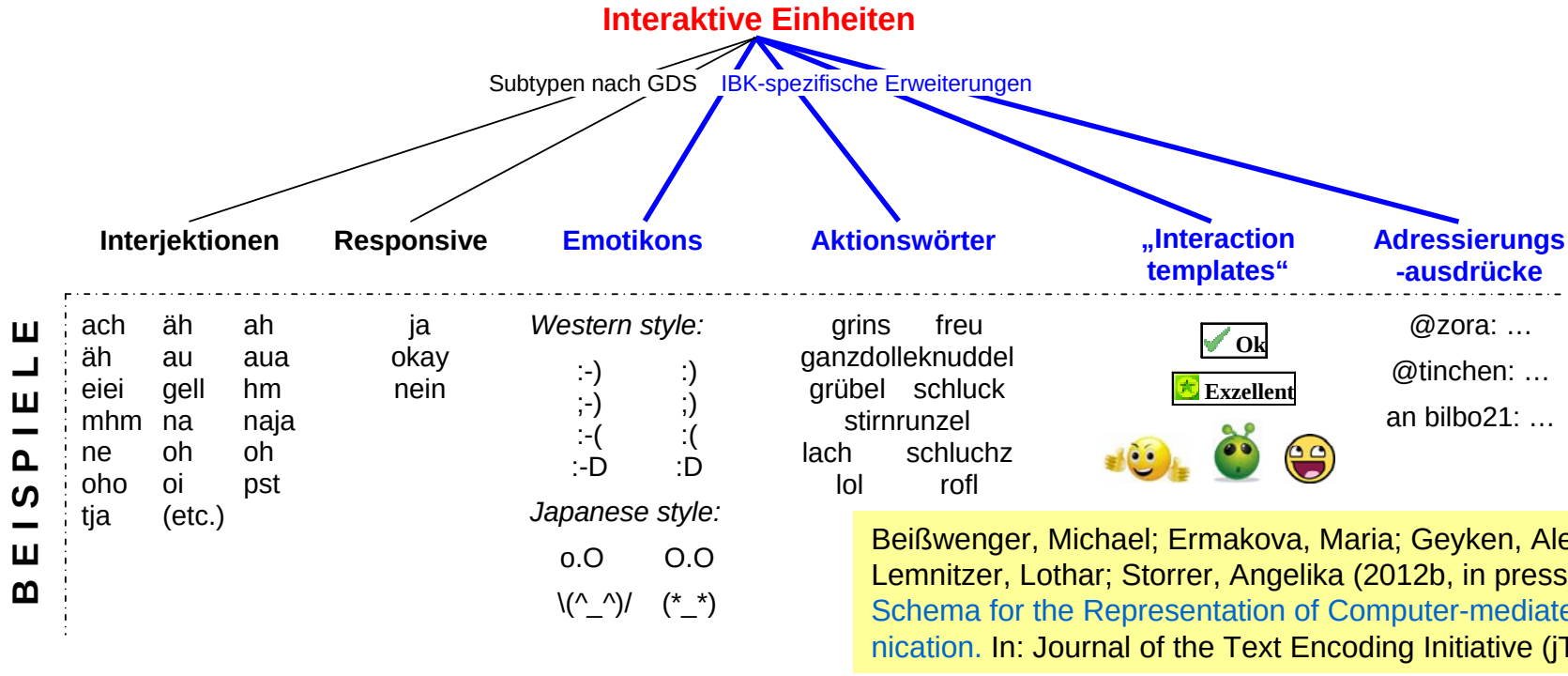
Tagging-Ergebnis Durchgang 2:

POS-Tag	Anzahl
NN	21
VVIMP	9
NE	3
ADJA	5
ADJD	1
PTKVZ	1
Gesamt	**40**

Beobachtung Durchgang 2:

Gibt es zu einem einfachen Inflektivausdruck eine homonyme Imperativform, so wird die Form in 8 von 14 Fällen als VVIMP klassifiziert.

Interaktive Einheiten in der internetbasierten Kommunikation



Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012b, in press): A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative (jTEI).

Annahme: Emoticons, Aktionswörter, Adressierungsausdrücke und ähnliche Einheiten weisen funktionale Parallelen zur Kategorie der „interaktiven Einheiten“ in der GDS auf – erweitern diese Kategorie aber (IBK-spezifisch) um zusätzliche Ausdrucksmöglichkeiten und Funktionen.

Vorschlag: Eine Einordnung von Emotikons, Aktionswörtern u.Ä. in ein System sprachlicher Formen und Funktionen ist am ehesten als Erweiterung der Kategorie der „interaktiven Einheiten“ möglich. Eine solche Einordnung zeigt dann zugleich, in welchem Bereich IBK sprachlichen Ausbau anstößt: nämlich in dem Bereich, der auf die Organisation und Verstehenssicherung in dialogischer Interaktion spezialisiert ist.

Überlegungen zur Erweiterung von STTS für die Beschreibung IBK-spezifischer interaktiver Einheiten

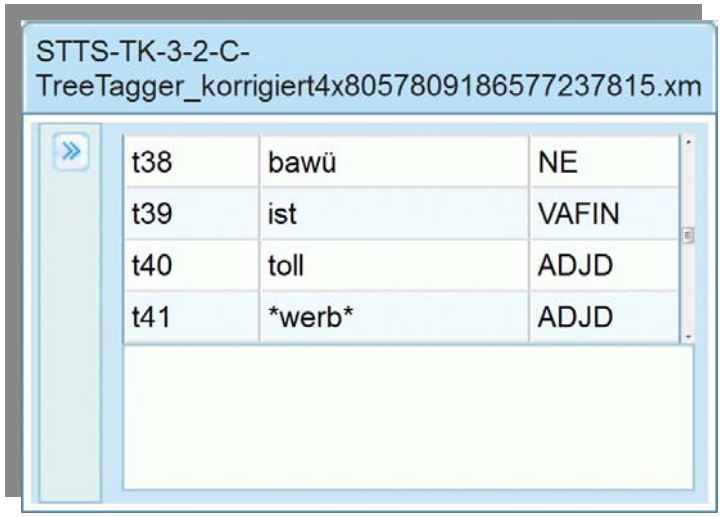
dortmund

STTS:		GDS:	
ITJ (Interjektion)	≈	Interjektion	*interaktive Einheiten*
PTKANT (Antwortpartikel)	≈	Responsiv	

Um den IBK-spezifischen Ausbau im Bereich der interaktiven Einheiten abzubilden, könnte der Bereich „Interjektionen“ und „Antwortpartikeln“ z.B. wie folgt restrukturiert und erweitert werden:

STTS:	„STTS 2.0“:	
	IE...	(Hauptkategorie „interaktive Einheit“)
ITJ (Interjektion)	IEITJ	(Subkat.: Interjektion)
PTKANT (Antwortpartikel)	IERSP	(Subkat.: Responsiv)
	IEEMO	(Subkat.: Emotikon)
	IEAKT	(Subkat.: Aktionswort)
	IEADR	(Subkat.: Adressierungsausdruck)

Überlegungen zur Modifikation und Erweiterung von STTS für das Tagging von Korpora zur internetbasierten Kommunikation


STTS-TK-3-2-C-TreeTagger_korrigiert4x8057809186577237815.xm

t38	bawü	NE
t39	ist	VAFIN
t40	toll	ADJD
t41	*werb*	ADJD

thomas.bartz@tu-dortmund.de
michael.beisswenger@tu-dortmund.de
angelika.storrer@tu-dortmund.de

technische universität dortmund

CLARIN-D-Workshop:

Das STTS-Tagset für Wortartentagging: Stand und Perspektiven

Stuttgart, 24. September 2012