



CLARIN-D-Kurationsprojekt: Linguistische Annotation von Nichtstandardvarietäten Guidelines und „Best Practices“ (F-AG 7)

Anke Lüdeling*, Stefanie Dippert†,
Marc Reznicek+*, Burkhard Dietterle*

† Ruhr-Universität Bochum, * Humboldt-Universität zu Berlin

- Projekt und Projektziele
- Nichtstandardsprachliche Strukturen
- Vorverarbeitung und Annotation
- Output

Clarín-D F-AG7 Kurationsprojekt (KP2)



Linguistische Annotation von Nichtstandard-Varietäten Guidelines und „Best Practices“ (F-AG 7)

- Annotationschemata, Guidelines und automatische Tools basieren auf Zeitungssprache
- **Pilotprojekt: Erweiterung existierender Ressourcen für**
 - **5 Nichtstandard-Varietäten**

Clarín-D F-AG7 Kurationsprojekt (KP2)



Linguistische Annotation von Nichtstandard-Varietäten Guidelines und „Best Practices“ (F-AG 7)

- Annotationschemata, Guidelines und automatische Tools basieren auf Zeitungssprache
- **Pilotprojekt: Erweiterung existierender Ressourcen für**
 - **5 Nichtstandard-Varietäten**
 - **3 Anntotationstypen**
 - Abhängigkeitsanalyse
 - Named Entities
 - Koreferenz

L2-Lerneraufsätze:

Falko

6,762 Tokens

Falko

- Wortstellungsabweichungen
- kreative Wortbildung
- nicht-kanonische syntaktische Argumentstruktur
- abweichende morphologische Markierung

gespr. Map Tasks:

BeMaTac

6,731 Tokens



- Wiederholungen
- Selbstkorrekturen
- Anakoluth
- Online-Argumententwicklung

Chat-Protokolle:

DCK – Plauderchat

6,664 Tokens

tu technische universität dortmund
Dortmunder Chat-Korpus

- Schreibfehler
- Inflektive (V_{end})
- Asterisk-Ausdrücke
- @-Adressierung
- Emoticons
- Verkettungen

Literarische Prosa:

Kafka – Der Prozeß

7,294 Tokens



- mehrfache Argumentbesetzung
- komplexe Parenthesen

Zeitung:

TüBa-DZ

5,000 Tokens



Standard-Varietät

Historische Texte:

DDB & Anselm

2,348 + 4,705 Tokens



- keine Satzsegmentierung
- eher freie Wortstellung
- keine standardisierte Schreibung

(Dipper et al. erscheint) <https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/clarin-d>

Guidelines & Best Practices

- Dokumentation der Datenverarbeitung
- Evaluation bestehender Annotationsrichtlinien
- Erweiterung der Annotationsschemata für NoSta-D-Varietäten

Named Entity

- STTS
(Schiller et al. 1999)
- Tüba D/Z
(Telljohan et al. 2012)
- MUC-6
(Grishman 1995)

Dependenz

- Constraint Dependenz
Grammatik
(Foth 2006)
- TiGer
(Albert et al. 2003)

Koreferenz

- Tüba-D/Z
(Naumann 2007)
- PoCos erweitertes
Schema
(Kaupat et al. 2013)

Interessante Nichtstandard-Phänomene

Named Entities (Chat)

▪ Kreative Namensschreibung

#	Sprecher	Beitrag
25	system	Lantonie betritt den Raum.
26	Lantonie	:)
27	quaki	lantonieeeeeee
28	Lantonie	Hallo. :)
29	zora	LANTOOO :)))
34	marc30	Lantöööö :o)
35	TomcatMJ	hi lanto

Named Entities (Chat)

■ Kreative Namensschreibung

#	Sprecher	Beitrag
25	system	Lantonie betritt den Raum.
26	Lantonie	:)
27	quaki	lantonieeeeeee
28	Lantonie	Hallo. :)
29	zora	LANTOOO :)))
34	marc30	Lantöööö :o)
35	TomcatMJ	hi lanto

■ Kreative Namensgebung

#	Sprecher	Beitrag
434	system	zurück betritt den Raum.

Syntax (Map Task & Chat)

- **Selbstkorrekturen**
- **Satzfragmente**

➤ deswegen würde ich dir vorschlagen dass du bis zum Burger gehst

➤ okay

➤ **auf der rechten** **auf der rechten** in der rechten oberen Ecke

➤ **rechte obere Ecke**

Bematac_2011-12-14-B:108

Syntax (Map Task & Chat)

- **Selbstkorrekturen**
- **Satzfragmente**

➤ deswegen würde ich dir vorschlagen dass du bis zum Burger gehst

➤ okay

➤ **auf der rechten** **auf der rechten** in der rechten oberen Ecke

➤ **rechte obere Ecke**

- **@-Adressierungen**

Bematac_2011-12-14-B:108

#	Sprecher	Beitrag
514	TomcatMJ	ja,mit dem ast in der hand im teich am rumsitzen @stoeps *G*

221006_unicum_21-02-2003_1

- **Fehlerhafte Verwendung von Pronomen**

Wenn **man** sich mit dieser Frage im Rahmen der Ethik beschäftigt, wird **er** fast auf jeden Fall sagen dass Kriminalität sich nicht auszahlt.

Falko_fk002_2006_08

▪ Fehlerhafte Verwendung von Pronomen

Wenn **man** sich mit dieser Frage im Rahmen der Ethik beschäftigt, wird **er** fast auf jeden Fall sagen dass Kriminalität sich nicht auszahlt.

Falko_fk002_2006_08

▪ Nicht-realisierte Referenten

516 Erdbeer\$ bochum ist ne stadt im pott

519 quaki Bochum is dunkelgrün???

524 Erdbeere\$ ne ~~Ø~~ grau

528 Erdbeere\$ ~~Ø~~ im grauen pott

534 Erdbeere\$ ~~Ø~~ hell schwarz?

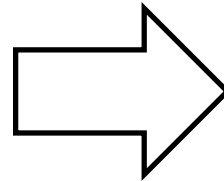
221006_unicum_21-02-2003_1

Vorverarbeitung

- Linearisierung
- Segmentierung
- Tokenisierung
- Normalisierung

Überlappende Redebeiträge in Map Tasks

Sprecher 1	Sprecher 2
äh und	
gehst jetzt nach	warte mal Tür is noch nich ganz zu



Sprecher 1	Sprecher 2
Sprecher 1	äh und
Sprecher 1	gehst jetzt nach
Sprecher 2	warte mal Tür is noch nich ganz zu

Vorverarbeitung

- Linearisierung
- **Segmentierung**
- Tokenisierung
- Normalisierung

Fehlende Satzgrenzen in historischen Texten

■ Pro Segment: 1 Matrixsatz mit Abhängigen

B1_1v,16	bifchhof tete fente anhel(=)	sente anshelmus bat marien manch iar
B1_1v,17	m ⁹ bat marien manch	myt heysen trenen das sy ym offenbarte
B1_1v,18	iar myt heysen trenen·	wy vnser here ih-us cristus syne marter
B1_2r,01	das fy ym offenbarte wy	irleden hatte
B1_2r,02	vnser here ihus cristus	
B1_2r,03	fyne marter irleden hatte	
B1_2r,04	do sprach vnse vrouwe Anf=	do sprach vnse vrouwe Anshelme ich sage
B1_2r,05	helme ich fage dir das	dir das myn here ihesus cristus alzo grose
B1_2r,06	myn here ihesus cristus·	martir irleden hot . das sy nyrkeyn
B1_2r,07	alzo grose martir irleden	mensche usgelegten mak
B1_2r,08	hot· das fy nyrkeyn men=	
B1_2r,09	fche us gelegen mak ¶ Doch	Doch salt u wissen daz ich an sotane
B1_2r,10	faltu wiffen· daz ich an fo ta=	wirdekeit komen byn das ich nvmmermer
B1_2r,11	ne wirdekeit komen byn·	betrubet mak werden
B1_2r,12	das ich nvmmermer be=	
B1_2r,13	trubet mak werden ¶ dar(=)	

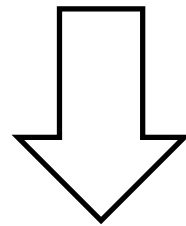
Vorverarbeitung

- Linearisierung
- Segmentierung
- **Tokenisierung**
- Normalisierung

Kontatationen in Chat

- Lexeme werden in einzelne Tokens getrennt
- mit „|“ markiert

165 quaki *nagut50cmlaufleine*



165 quaki * na| gut| 50| cm| laufleine| *

Vorverarbeitung

- Linearisierung
- Segmentierung
- Tokenisierung
- **Normalisierung**

Normalisierung: Motivation

Named Entity:

- **Uneinheitliche Namensschreibung**

Einheitliche Schreibung erleichtert nachvollziehbare Klassifizierung

→ **Normalisierung auf Sprecheralias (Chat), Lexikoneinträge**

Normalisierung: Motivation

Named Entity:

- **Uneinheitliche Namensschreibung**

Einheitliche Schreibung erleichtert nachvollziehbare Klassifizierung

→ **Normalisierung auf Sprecheralias (Chat), Lexikoneinträge**

Dependenz:

- **Satzfragmente**

Nur Verben können grammatische Rolle im Satz verteilen.

→ **Ergänzung von Auslassungen und Ellipsen**

Normalisierung: Motivation

Named Entity:

- **Uneinheitliche Namensschreibung**

Einheitliche Schreibung erleichtert nachvollziehbare Klassifizierung

→ **Normalisierung auf Sprecheralias (Chat), Lexikoneinträge**

Dependenz:

- **Satzfragmente**

Nur Verben können grammatische Rolle im Satz verteilen.

→ **Ergänzung von Auslassungen und Ellipsen**

Koreferenz:

- **Nicht/falsch an der sprachlichen Oberfläche realisierte Referenzen**

Explizierung nicht-realisierten Referenten erlaubt Einbindung in referenzielle Kette

→ **Ergänzung und Korrektur von Referenzausdrücken**

Zwei Perspektiven

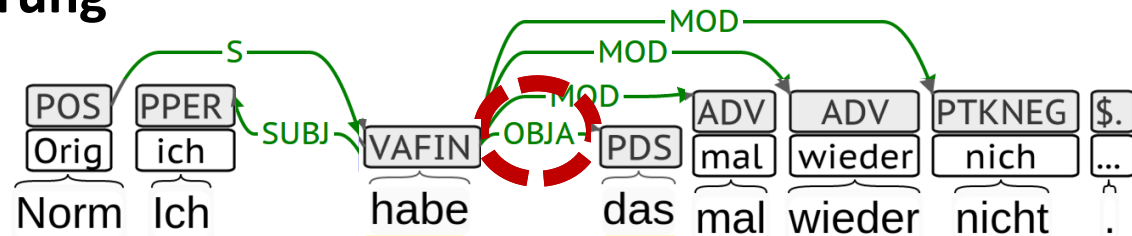
Zwei Repräsentationen

1 Variationistischer Ansatz:

Normalisierung = Index zur Klassifizierung vergleichbarer Phänomene in Korpora

Annotation der Normalisierung

Wie variiert die Realisierung von Subjekten zwischen Varietäten?



Zwei Perspektiven

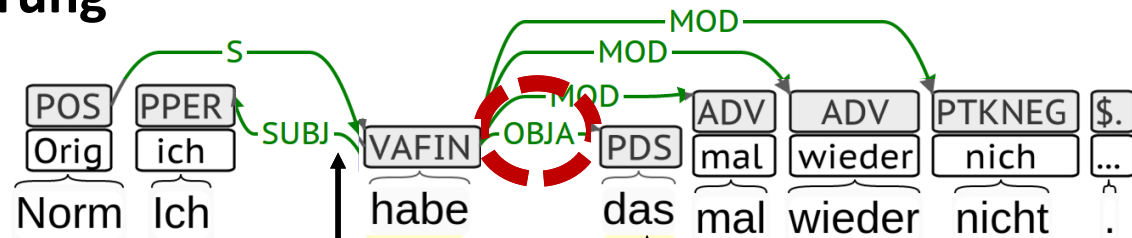
Zwei Repräsentationen

1 Variationistischer Ansatz:

Normalisierung = Index zur Klassifizierung vergleichbarer Phänomene in Korpora

Annotation der Normalisierung

Wie variiert die Realisierung von Subjekten zwischen Varietäten?

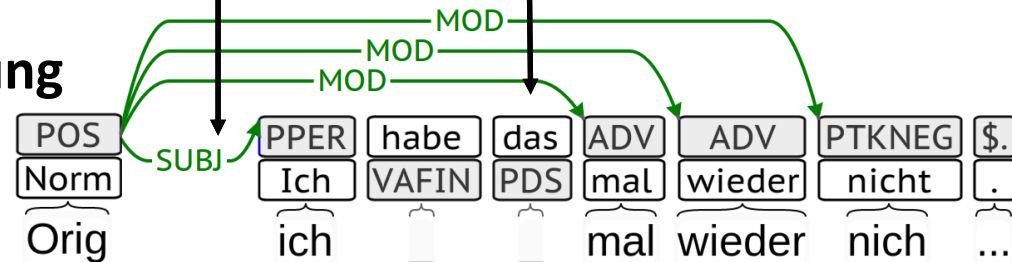


2 Computerlinguistischer Ansatz:

Normalisierung = minimaler Vorverarbeitungsschritt für weitere Verarbeitung

Annotation des Originaltextes auf Grundlage der Normalisierung

Wie annotiert man Fragmente?



Normalisierung: Named Entities

Regel (Chat): Normalisierung = Alias → PER

#	Sprecher	Original	Normalisierung
25	system	[Lantonie]PER betritt den Raum.	[Lantonie]PER betritt den Raum.
26	Lantonie	:)	:)
27	quaki	[lantonieeeee]PER	[Lantonie]PER!
28	Lantonie	Hallo. :)	Hallo :)
29	zora	[LANTOOO]PER :)))	[Lantonie]PER :)))
34	marc30	[Lantöööö]PER :o)	[Lantonie]PER :o)
35	TomcatMJ	hi [lanto]PER	Hi, [Lantonie]PER!

Normalisierung: Named Entities

Normalisierung = Organisation → ORG

#	Sprecher	Original	Normalisierung
429	Emon	boah... ich bekomme echt augenkrebs von bochum...	Boah! Ich bekomme echt Augenkrebs von Bochum.
436	Thor...	ich vom [has]ORG emon *g*	Ich vom [HSV]ORG, Emon *g*
439	Thor...	[hasv]ORG	[HSV]ORG
453	Emon	[hsv]ORG heisst dat	[HSV]ORG heißt das.

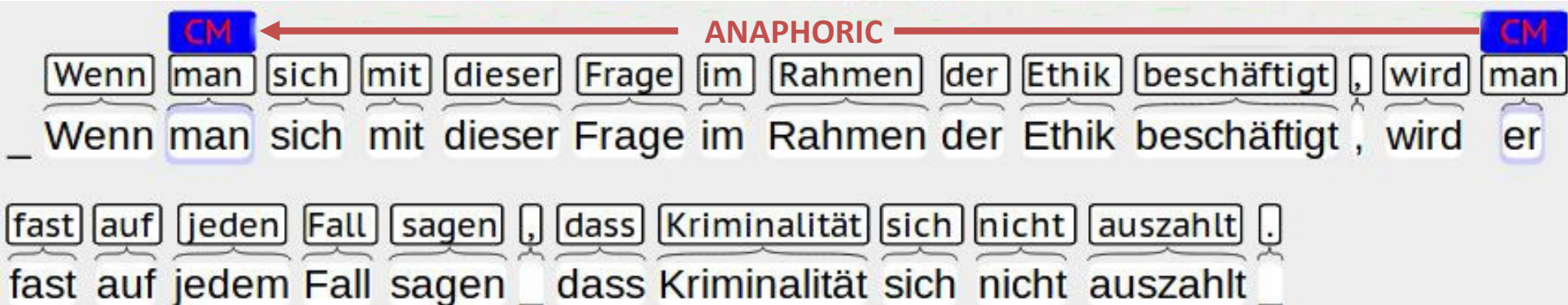
221006_unicum_21-02-2003_1

Normalisierung: Koreferenz

Norm. hat Antezedens → Orig. hat gleichen Antezedens

Norm-Regel: Ersetze in Kombinationen von “man” und “er/sie”
Letzteres durch Ersteres, wenn Letzteres im Kontext keinen
Antezedenten besitzt!

*Wenn **man** sich mit dieser Frage im Rahmen der Ethik beschäftigt, wird
(**er/ man**) fast auf jeden Fall sagen dass Kriminalität sich nicht auszahlt.*

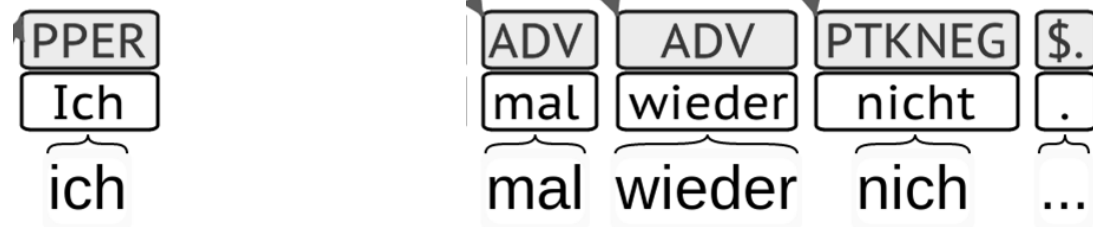


Normalisierung: Dependenz

Explizite Einfügung prototypischer Verben mit passender Argumentstruktur in verblose Sätze.

→ **Motivation der Fragmentfunktionen**

Original

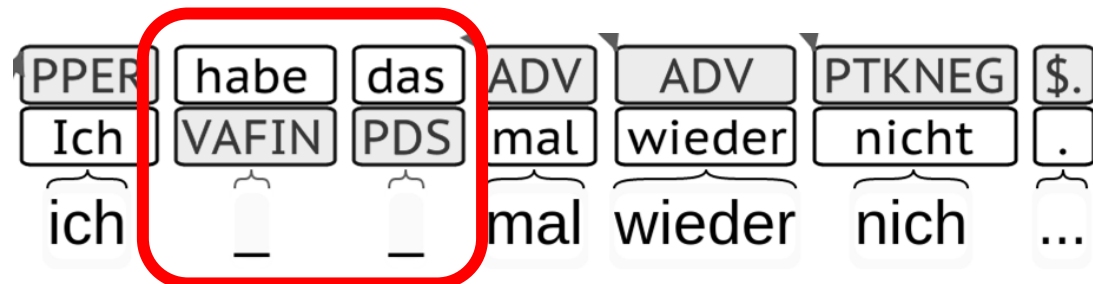


Normalisierung: Dependenz

Explizite Einfügung prototypischer Verben mit passender Argumentstruktur in verblose Sätze.

→ **Motivation der Fragmentfunktionen**

Original

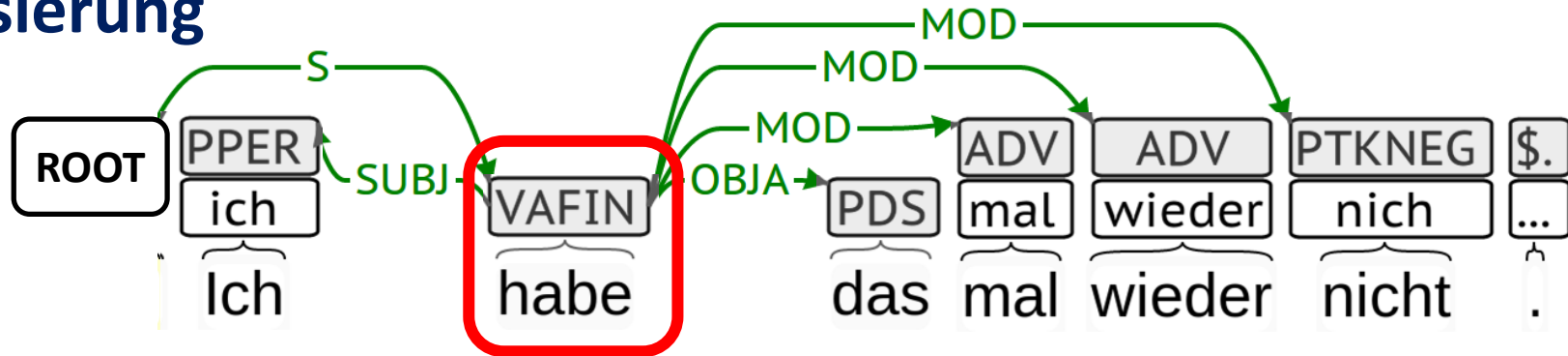


Normalization: Motivation

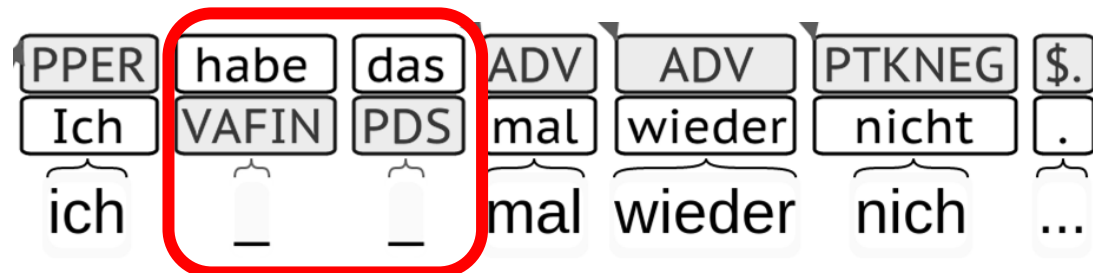
Explizite Einfügung prototypischer Verben mit passender Argumentstruktur in verblose Sätze.

→ **Motivation der Fragmentfunktionen**

Normalisierung



Original

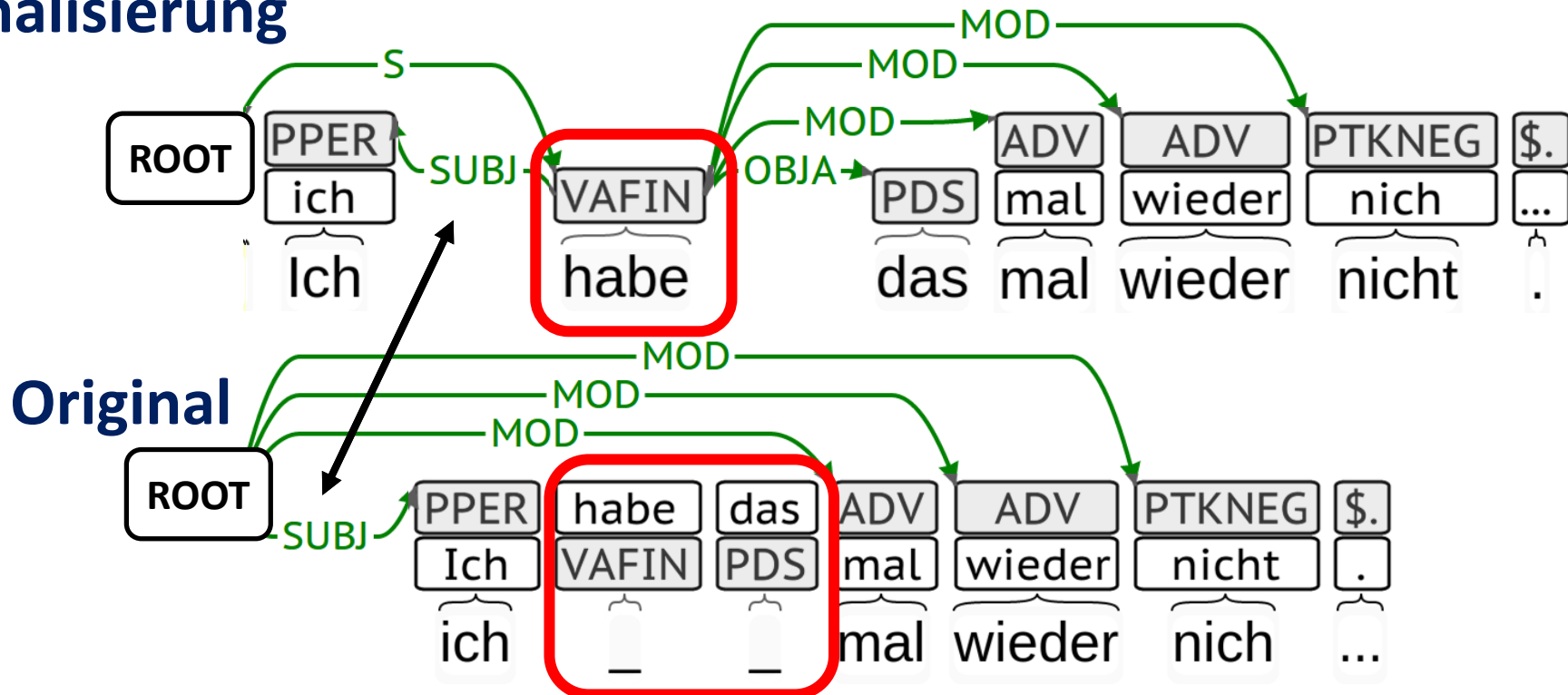


Retokenisierung : @-Adressierung

Explizite Einfügung prototypischer Verben mit passender Argumentstruktur in verblose Sätze.

→ **Motivation der Fragmentfunktionen**

Normalisierung



Retokenisierung: @-Adressierung

mit dem ast in der hand im teich am rumsitzen @stoeps

mit dem ast in der hand im teich am rumsitzen @ stoeps

Retokenisierung : @-Adressierung

mit dem ast in der hand im teich am rumsitzen @ stoeps

mit dem ast in der hand im teich am rumsitzen @ stoeps

XX

PN

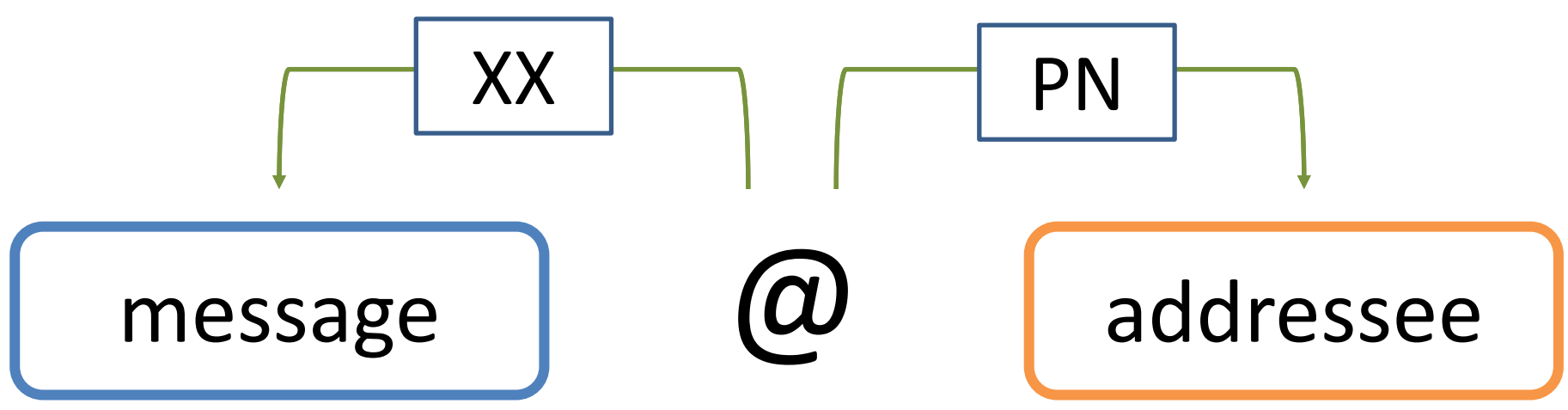
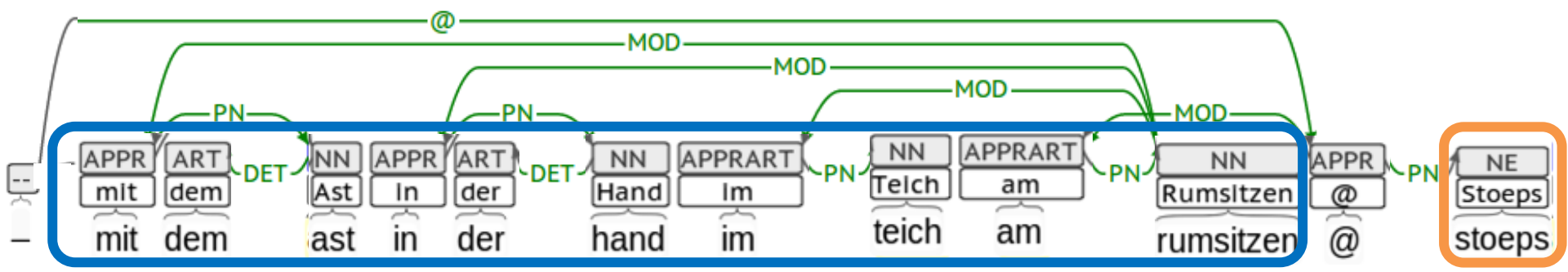
message

@

addressee

Retokenisierung : @-Adressierung

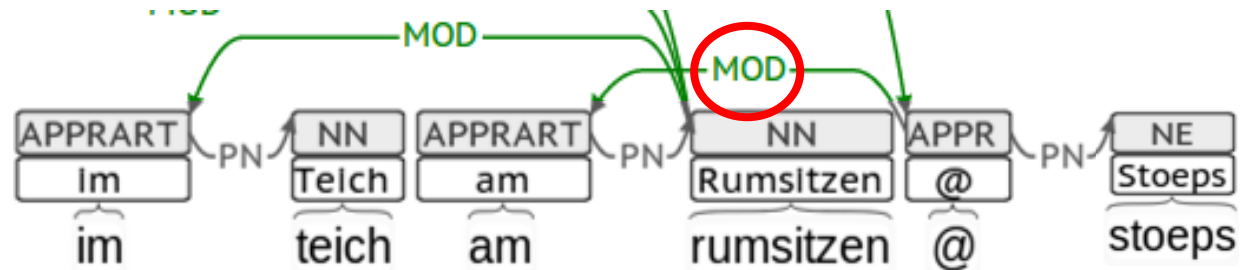
mit dem ast in der hand im teich am rumsitzen @stoeps



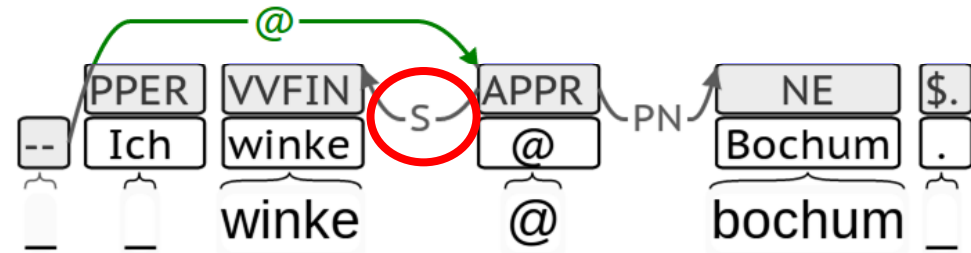
Retokenisierung : @-Adressierung

@-attached arguments are of variable type:

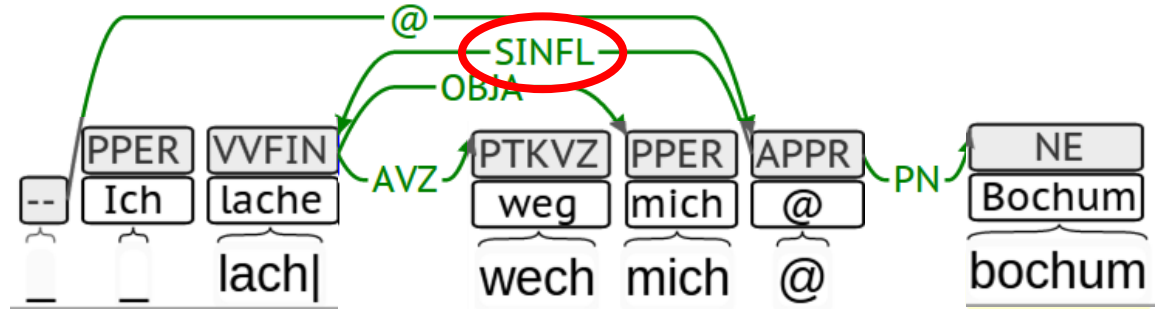
Modification: **MOD**



Sentence root: **S**



Inflective root: **SINFL**



Annotation mit NoSta-D Guidelines

Abhängigkeiten (Guidelines)

NoSta-D ³⁸

Du hast gegessen und geraucht

Abhängigkeiten (Guidelines)

TiGer Annotationsschema (Albert et al. 2003)

Koordination

- 9.1 Grundstruktur der NP-, AP-, PP-Koordination
 - 9.1.1 Koordinierende Konjunktionen
 - 9.1.2 Binäre koordinierende Konjunktionen
- 9.2 Koordination von satzeinleitenden Konjunktionen (CPs)
- 9.3 Koordination von Nominal- und Präpositionalphrasen
- 9.4 Koordinierte Adjektive **Du**
- 9.5 Koordinierte Präpositionen
- 9.6 Koordination von Verbalphrasen und Sätzen

hast gegessen und geraucht



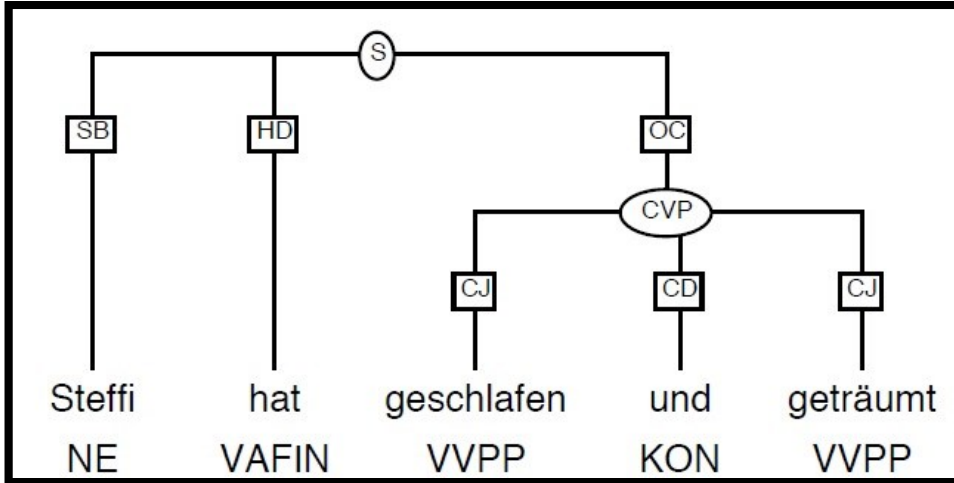
Abhängigkeiten (Guidelines)

TiGer AS 2003

S.117, Bsp. 2

Koordination von Verbalphrasen

NoSta-D 40



Du hast geschlafen und geträumt

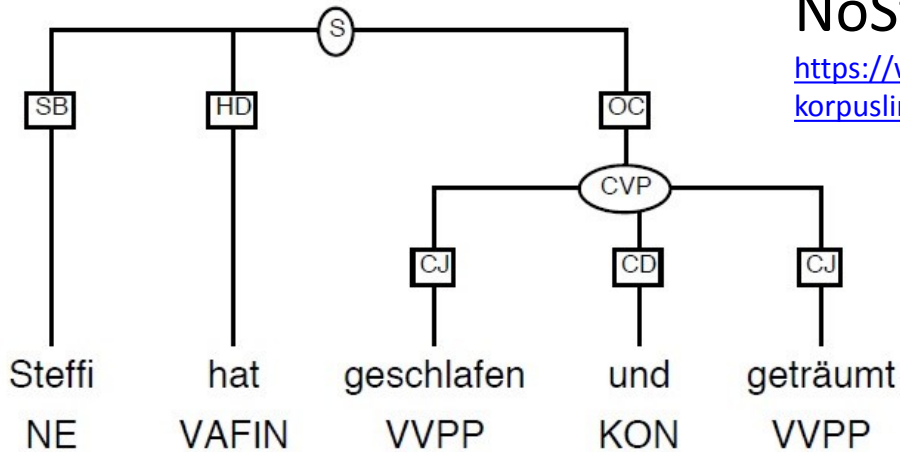
Abhängigkeiten (Richtlinien)

TiGer 2003

S.117, Bsp. 2

NoSta-D Richtlinien für Abhängigkeiten

<https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/clarin-d>



Du hast gegessen **und** geraucht

Dependens

(C...)-[CD]

Was

Abhängigkeiten (Richtlinien)

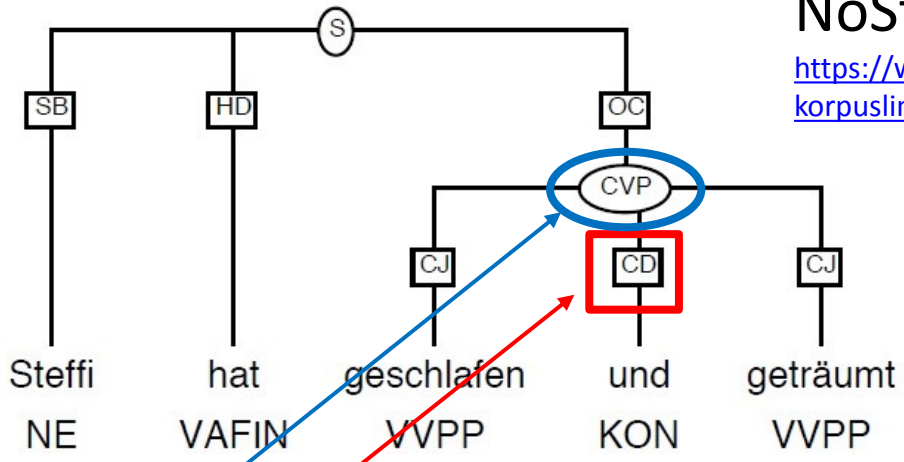
NoSta-D

TiGer 2003

S.117, Bsp. 2

NoSta-D Richtlinien für Abhängigkeiten

<https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/clarin-d>

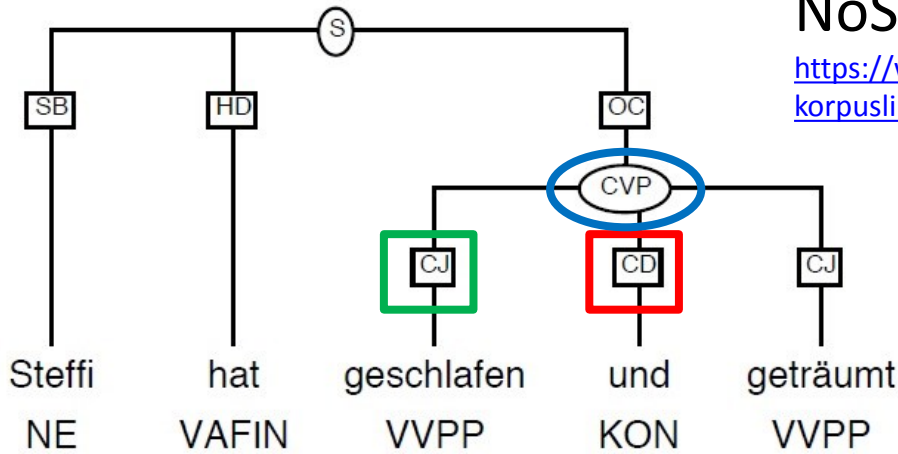


Abhängens
(C...)-[CD]
Was

Abhängigkeiten (Richtlinien)

TiGer 2003

S.117, Bsp. 2



NoSta-D Richtlinien für Abhängigkeiten

<https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/clarin-d>



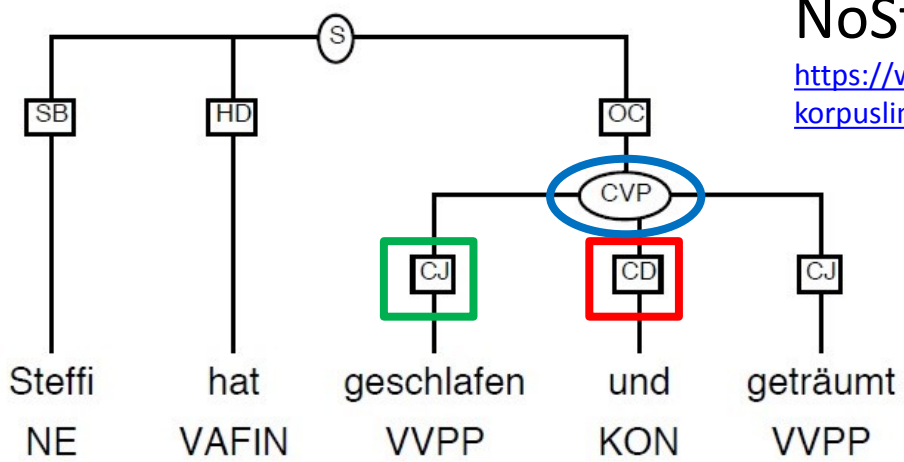
Abhängens	Regens
(C...)-[CD]	das innerhalb derselben (C...) unmittelbar vorangehende (C...)-[CJ], es sei denn ...
Was	ist Tochter von

Abhängigkeiten (Guidelines)

TiGer 2003

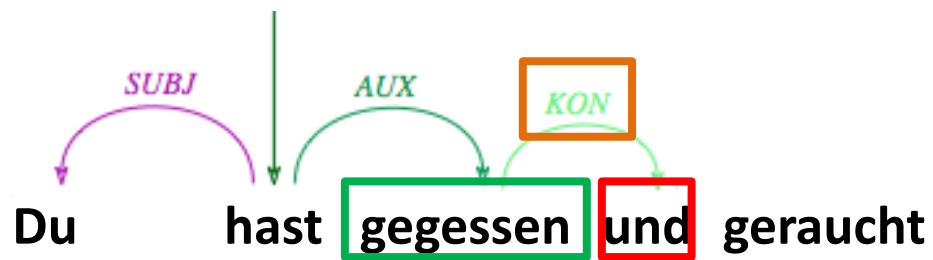
NoSta-D

S.117, Bsp. 2



NoSta-D Guidelines für Abhängigkeiten

<https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/clarin-d>



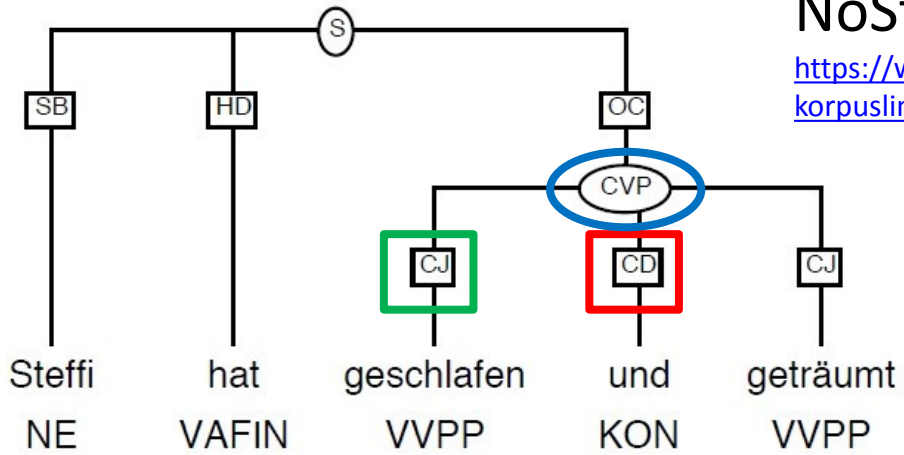
Dependens	Regens	Label
(C...)-[CD]	das innerhalb derselben (C...) unmittelbar vorangehende (C...)-[CJ], es sei denn ...	KON
Was	ist Tochter von	wie

Abhängigkeiten (Guidelines)

TiGer 2003

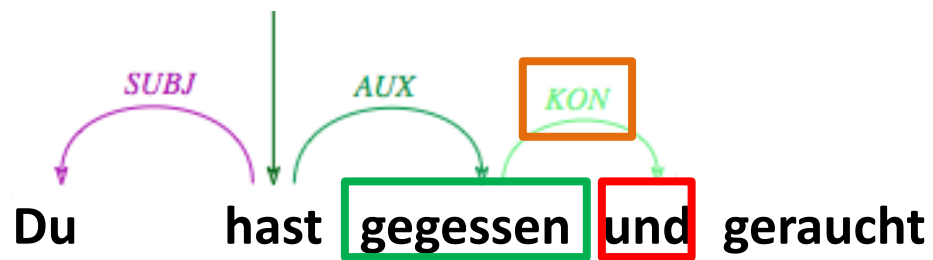
NoSta-D

S.117, Bsp. 2



NoSta-D Guidelines für Abhängigkeiten

<https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/clarin-d>



Dependens	Regens	Label	Beispiel
(C...)-[CD]	das innerhalb derselben (C...) unmittelbar vorangehende (C...)-[CJ], es sei denn ...	KON	S.117, Bsp. 2
Was	ist Tochter von	wie	?

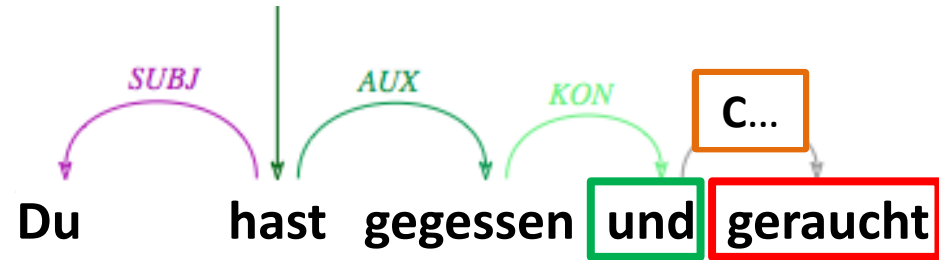
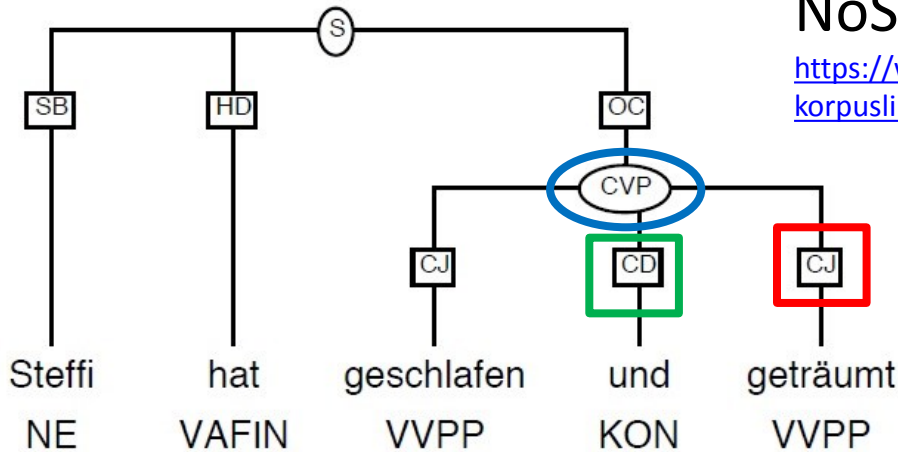
Abhängigkeiten (Richtlinien)

TiGer 2003

S.117, Bsp. 2

NoSta-D Richtlinien für Abhängigkeiten

<https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/clarin-d>



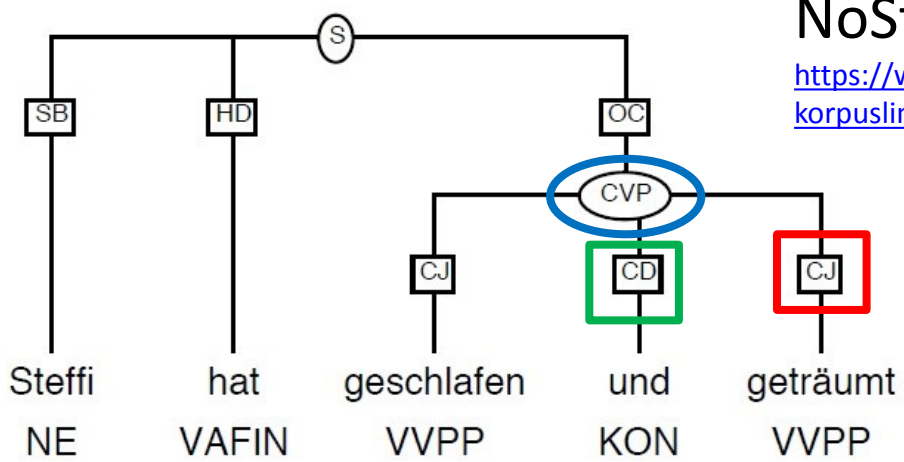
Abhängens	Regens	Label	Beispiel
(C...)-[CJ]	das innerhalb derselben (C...) unmittelbar links stehende (C...)-[C...], sofern ...	C...	S.117, Bsp. 2
Was	ist Tochter von	wie	

Abhängigkeiten (Guidelines)

TiGer 2003

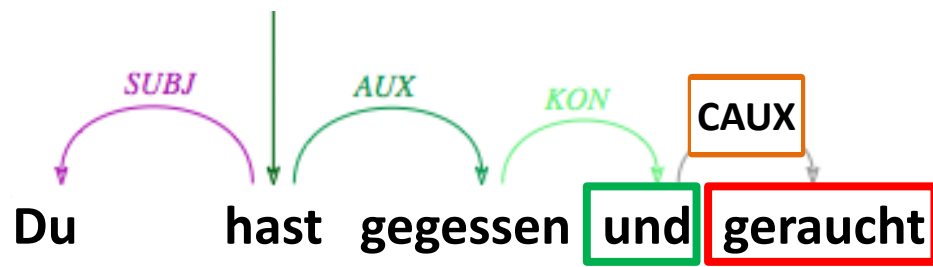
NoSta-D

S.117, Bsp. 2



NoSta-D Guidelines für Abhängigkeiten

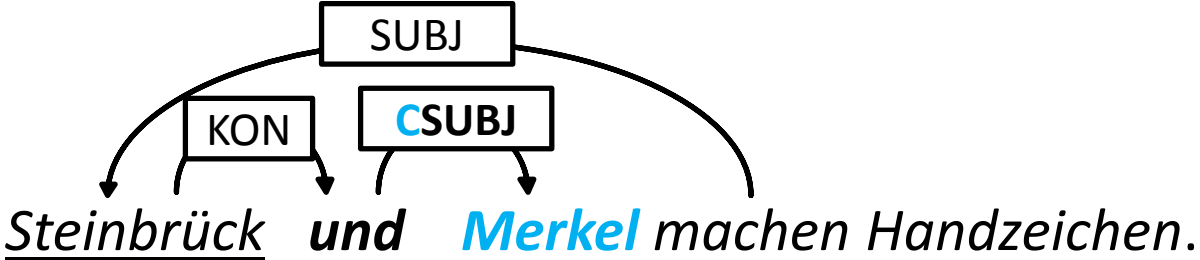
<https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/clarin-d>

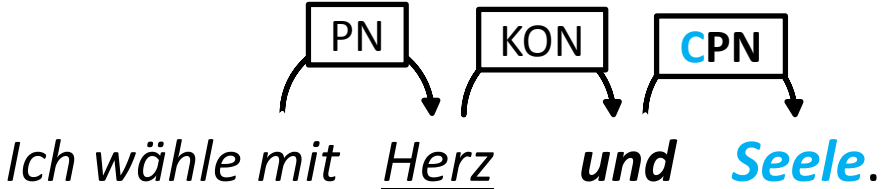


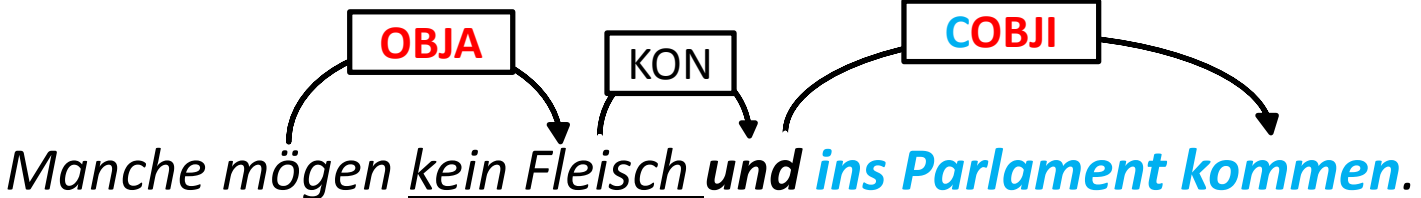
Dependens	Regens	Label	Beispiel
(C...)-[CJ]	das innerhalb derselben (C...) unmittelbar links stehende (C...)-[C...], sofern ... Das Restlabel ergibt sich aus der direkten Relation der Mutter des ersten koordinierten Elements zum Dependens.	C...	S.117, Bsp. 2
Was	ist Tochter von	wie	

Kreuzklasse C

Koordination

- **CSUBJ**


Steinbrück und Merkel machen Handzeichen.
- **CPN**


Ich wähle mit Herz und Seele.
- **COBJI**


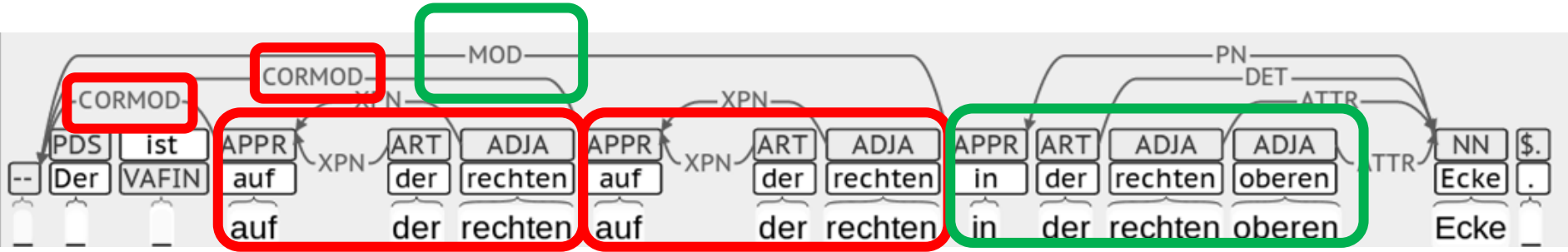
Manche mögen kein Fleisch und ins Parlament kommen.

Abhängigkeiten (Guidelines)

Kreuzklasse **COR** Korrektur

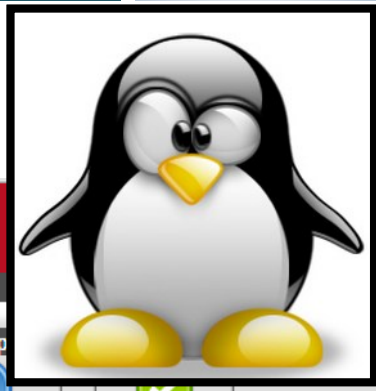
Die höchste Mutter, ab der ein Korrektur beginnt.

- **CORMOD** → MOD-Kante, ab der korrigiert wird
- [auf der rechten]_{CORMOD} ... [in der rechten oberen Ecke]_{MOD}



Dependency annotation

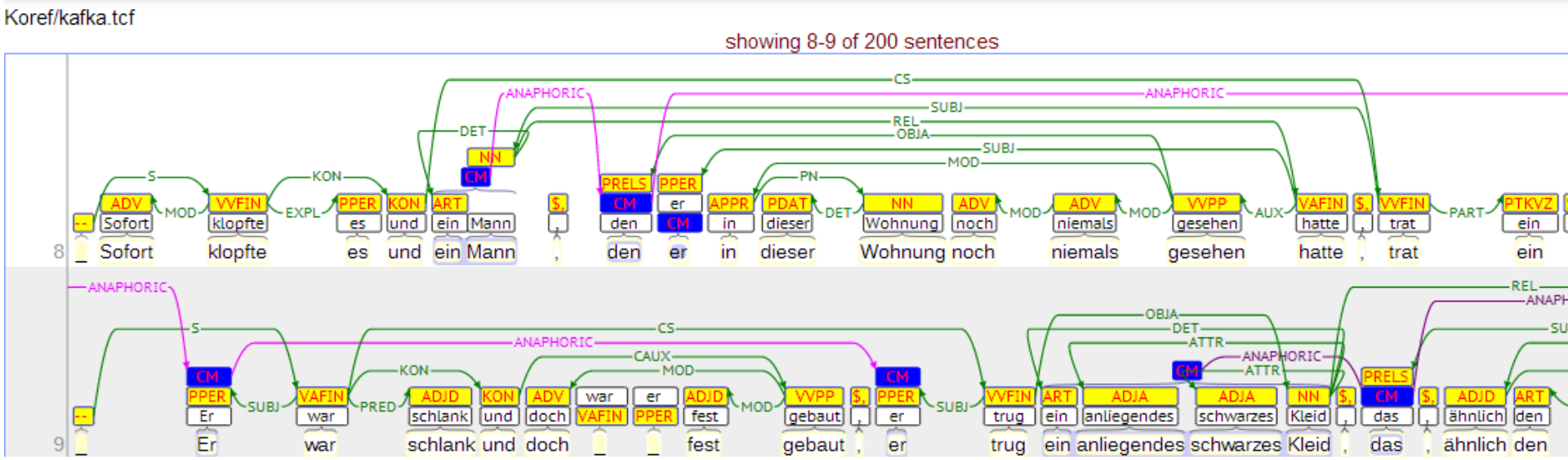
Annotation with WebAnno



Annotation
WebAnno | Home

Document | **Page** | **Help**

Open Prev Next Export Settings | First Prev Go to Next Last | Guidelines Done



<https://clarin.ukp.informatik.tu-darmstadt.de>

Daten und Guidelines

Bild ändern



HUMBOLDT-UNIVERSITÄT ZU BERLIN



1

Deutsch English Links Sitemap Impressum

CLARIN-D-Kurationsprojekt: Linguistische Annotation von Nichtstandardvarietäten — Guidelines und „Best Practices“ (F-AG 7) Korpora NoSta-D

Die Daten stehen in Kürze zur Verfügung

- | | | |
|---|-----|-----|
| ■ NoSta-D: Zeitung (TüBa-D/Z) | TSV | TCF |
| ■ NoSta-D: Literarische Prosa (Kafka - Der Prozeß) | TSV | TCF |
| ■ NoSta-D: gedrochene Map Tasks (BeMaTaC) | TSV | TCF |
| ■ NoSta-D: L2-Lerneraufsätze (Falko) | TSV | TCF |
| ■ NoSta-D: Chat-Protokolle (Dortmunder Chat Korpus) | TSV | TCF |

NoSta-D Guidelines

Die Guidelines stehen in Kürze zur Verfügung

- Vorverarbeitung
- Named Entity Recognition
- Abhängigkeiten
- Koreferenz



<https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/clarin-d/>

- **NoSta-D → Deutsche Pilot-Ressource für Training und Entwicklung für Nichtstandard-Varietät verfügbar**
- **Ergänzungen zu gängigen Richtlinien (NER, Abhängigkeiten und Korreferenzen) → gute Abdeckung von NoSta-D**
- **Normalisierung explizit in die Korpora integrieren → Nachvollziehbarkeit strittiger Annotationen erhöhen**
- **Parallele Annotation von Normalisierung und abgeleitet daraus der Originaldaten → Vereinbarkeit komplementärer Untersuchungsansätze**

- **Aufbau größerer Ressourcen und weiterer Varietäten für Training**
- **Untersuchung zusätzlicher Annotationstypen**
- **Grundsätzliche Kritik an der linguistischen Adäquatheit dependenzgrammatischer Modelle auch für den deutschen Standard**
 - u.a. Koordination finiter Verben
- **Generell stärkerer Fokus auf den Einsatz von Mehrebenenarchitekturen in Annotations-, Such- und Analysetools**
 - u.a. Gleichzeitige Darstellung und Verarbeitung konkurrierender Normalisierungsebenen

Danke!

Projektseite:

<https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/clarin-d>

Kontakt:

marc.reznicek@staff.hu-berlin.de

- Albert, Stefanie; Anderssen, Jan; Bader, Regine; Becker, Stefanie; Bracht, Tobias; Brants, Thorsten et al. (2003):** TiGer Annotationsschema.
- Dipper, Stefanie; Lüdeling, Anke; Reznicek, Marc (erscheint):** *NoSta-D. A Corpus of German Non-Standard Varieties*. In: Zampieri, Marcos; Diwersy, Sascha (Hgg.): *Non-Standard Data Sources in Corpus-Based Research (ZMS-Studien - Schriften des Zentrums Sprachenvielfalt und Mehrsprachigkeit der Universität zu Köln 5)*, Hamburg: Shaker.
- Foth, Kilian A. (2006):** Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Technischer Report. Universität Hamburg. Hamburg.
- Kaupat, David; Warzecha, Saskia; Stede, Manfred (2013):** Koreferenz. Chapter 5. Erweiterung des PoCoS-Kernschemas.
- Naumann, Karin (2007):** Manual for the Annotation of In-Document Referential Relations. Seminar für Sprachwissenschaft, Abt. Computerlinguistik Universität Tübingen, http://www.sfb441.uni-tuebingen.de/a1/Publikationen/tuebadz_relations_man.pdf
- Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine (1999):** Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical Report. University of Stuttgart; University of Tübingen, <http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf>.
- Telljohann, Heike; Hinrichs, Erhard W.; Kübler, Sandra; Zinsmeister, Heike; Beck (2012):** Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Seminar für Sprachwissenschaft. Universität Tübingen, <http://www.sfs.uni-tuebingen.de/resources/tuebadz-stylebook-1201.pdf>.

Stand: 26.09.2013