# Annotating Dependency Relations in Non-standard Varieties

Marc Reznicek
Stefanie Dipper
Anke Lüdeling
Burkhard Dietterle
Clarin-D F-AG 7 Curation Project II

empirikom

5. Arbeitstagung
25.04.2013, Hamburg

# Overview

- Annotation of non-standard varieties
- NoSta-D corpus
- Dependencies
- Normalisation
- Chat-specific linguistic structures
- Coordination (generell)
- Outlook

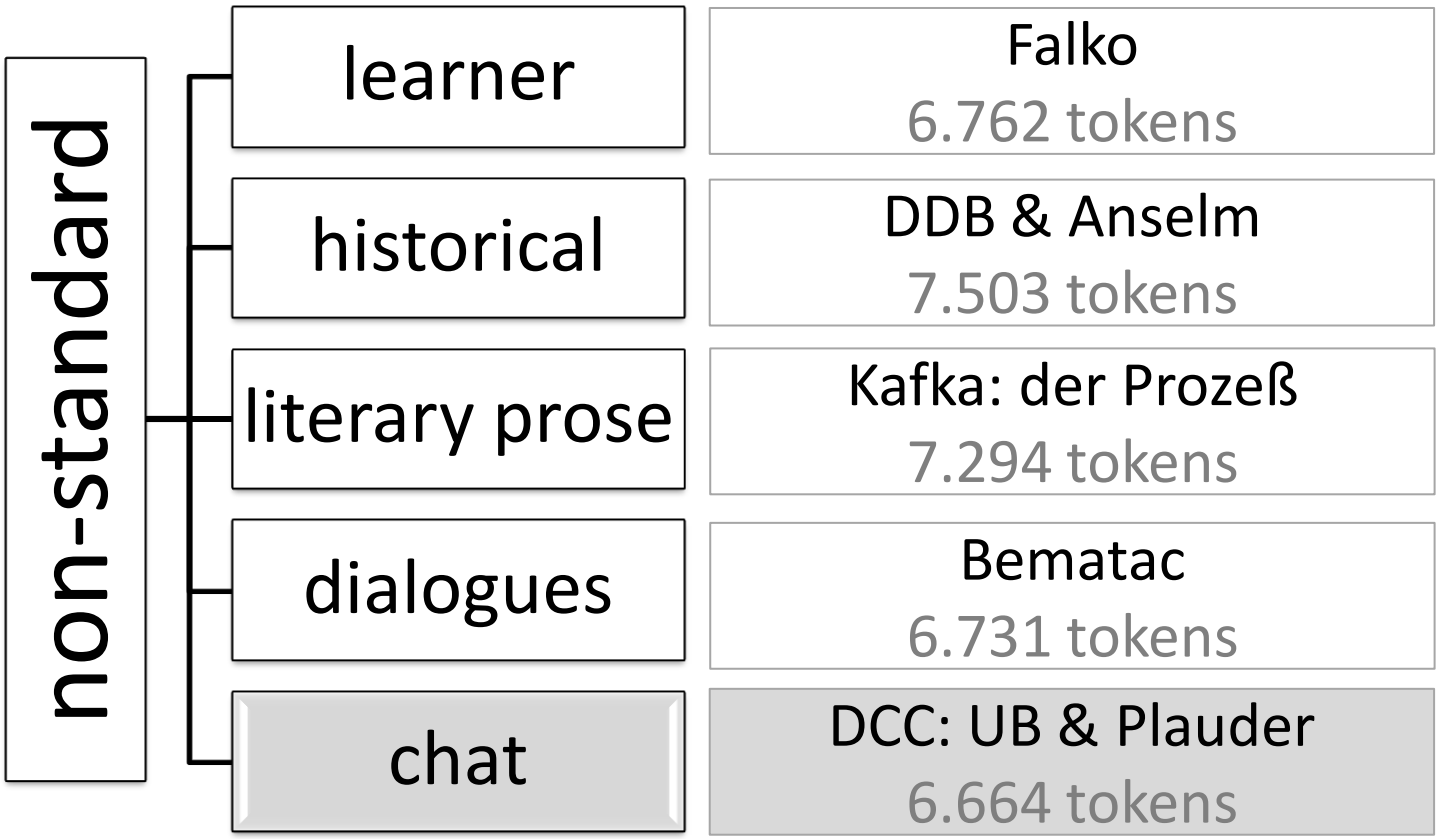
# Clarin-D Curation Project II

**Clarin F-AG 7 - Curation project (KP2):**
Linguistic annotation of non-standard varieties — guidelines and "best practices"

- Annotation categories , guidelines and automatic tools are based on newspaper texts
- Growing demand for the description of other (= non-standard) varieties.
- Pilot project: Extension of given resources for 5 non-standard varieties

# NoSta-D corpus

- Creation of a non-standard variety pilot corpus of German (Dipper et al. to appear)



# 3 types of annotations

- NER (spans)
- Coreference (pointing relations)
- Dependencies (trees)

Dependency parsing for German *"reaches an accuracy* […] *better than the best constituent analysis including grammatical functions."*

(Kübler & Prokic 2006)

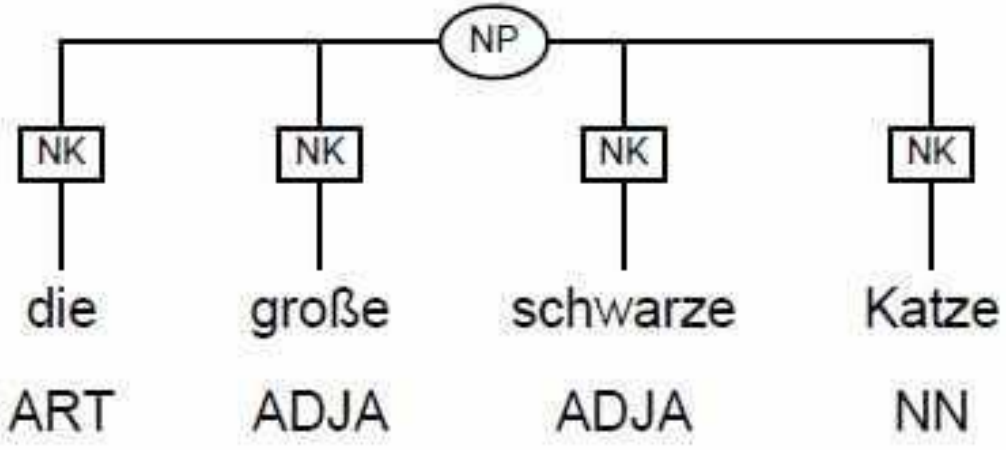
# Non-standard dependencies

- **How to define non-standard dependency structures?**

1) Take guidelines that fully describe structures in a large newspaper corpus of German:

→ **TiGer** (Alberts et al. 2003)

**problem: Constituents**



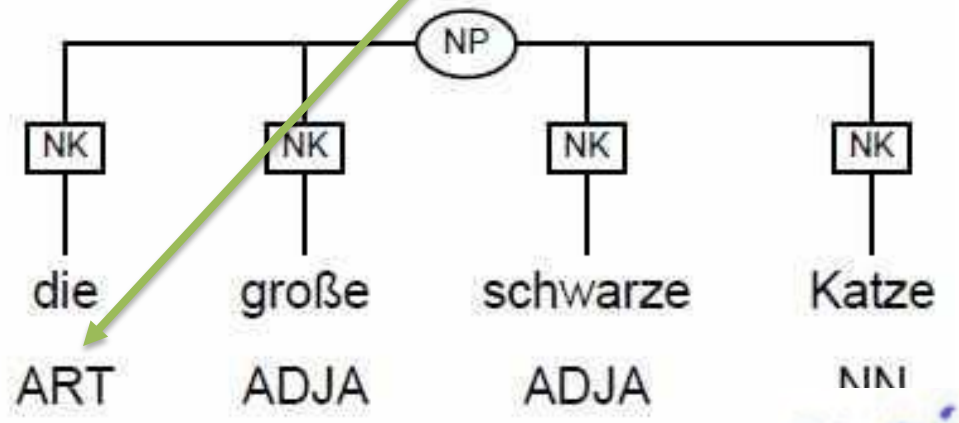

(Alberts et al. 2003:9)

# Non-standard dependencies

- **How to define non-standard dependency structures?**

2) Give human annotators a translation of TiGer-constituent trees into dependencies

PIS → HEAD
NN → HEAD if not head is PIS
ADJA → HEAD if not head is PIS or NN. If not HEAD, then ATTR
ART → DET

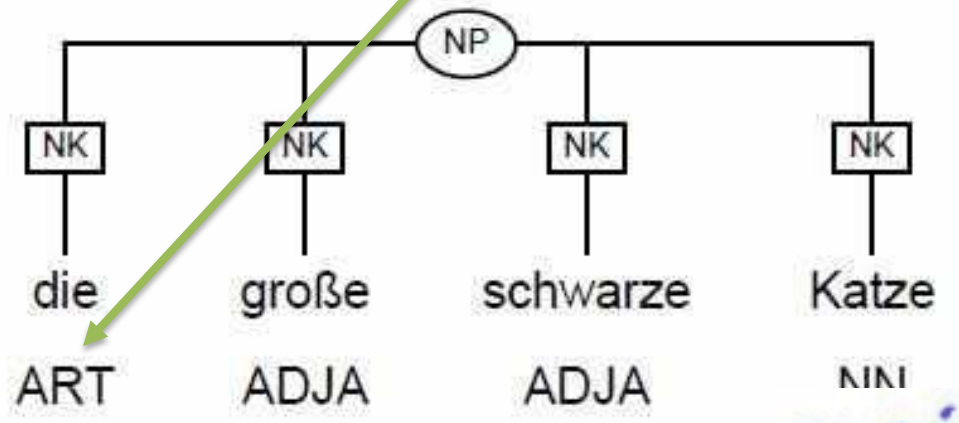


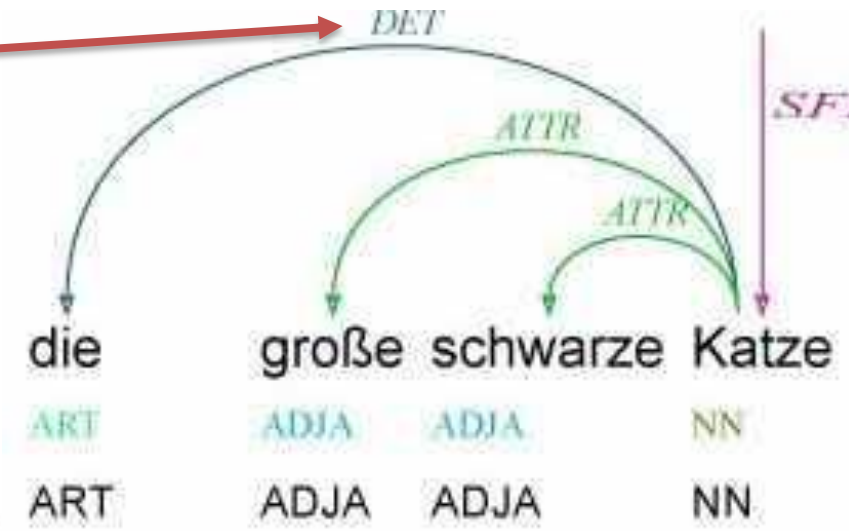
# Non-standard dependencies

- **How to define non-standard dependency structures?**

2) Give human annotators a translation of TiGer-constituent trees into dependencies

PIS → HEAD
NN → HEAD if not head is PIS
ADJA → HEAD if not head is PIS or NN. If not HEAD, then ATTR
ART → DET



# Non-standard dependencies

- **How to define non-standard dependency structures?**

3) Structures that aren't covered by the TiGer guidelines are considered non-standard.

# Dependencies in chat

- Plauderchat (Beißwenger 2013): very heterogeneous

| | | |
|---|---|---|
| 1 | **system** | JustChat 4.0r0.204 (55.204) developed by Medium.net. |
| 2 | **system** | Du betrittst den Raum. |
| 3 | **quaki** | was echt zori? |
| 4 | **system** | little15 betritt den Raum. |
| 5 | **quaki** | das küssen?? |
| 6 | **Pharao** | na gut marc. kein servicepaket nr.1 für dich :) |
| 7 | **zora** | was? |
| 8 | **system** | TomcatMJ kommt aus dem Raum Go-Rin-No-Sho herein. |
| 9 | **TomcatMJ** | hi |
| 10 | **system** | TomcatMJ ist wieder da. |

http://www.chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/unicum_21-02-2003_1.html

# Dependencies in chat

- Plauderchat (Beißwenger 2013): very heterogeneous
  - system messages (close to standard)

```
 2   system Du betrittst den Raum.
 3    quaki was echt zori?
 4   system little15 betritt den Raum.
 5    quaki das küssen??
 6   Pharao na gut marc. kein servicepaket nr.1 für dich :)
 7    zora was?
 8   system TomcatMJ kommt aus dem Raum Go-Rin-
            No-Sho herein.
 9 TomcatMJ hi
10   system TomcatMJ ist wieder da.
```

http://www.chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/unicum_21-02-2003_1.html

# Dependencies in chat

- Plauderchat: very heterogeneous
  - human postings (far from standard)

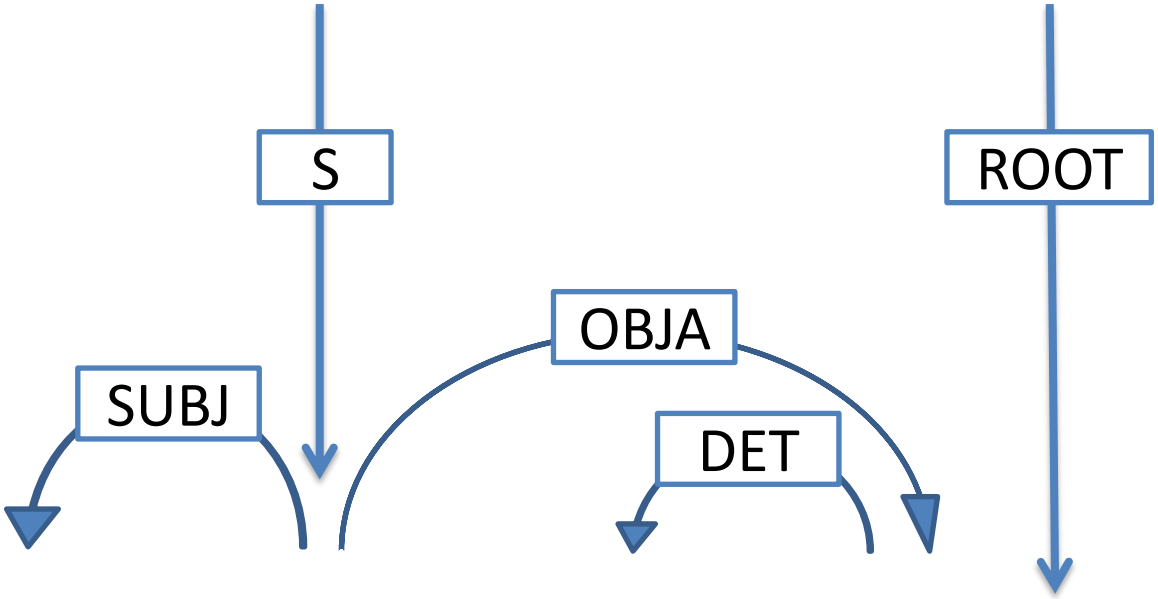
```
 1     system   JustChat 4.0r0.204 (55.204) developed by Medium.net.
 2     system   Du betrittst den Raum.
 3      quaki   was echt zori?
 4     system   little15 betritt den Raum.
 5      quaki   das küssen??
 6     Pharao   na gut marc. kein servicepaket nr.1
                für dich :)
 7       zora   was?
 8     system   TomcatMJ kommt aus dem Raum Go-Rin-No-Sho herein.
 9   TomcatMJ   hi
10     system   TomcatMJ ist wieder da.
```

http://www.chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/unicum_21-02-2003_1.html

# Dependency annotation

- most **system messages** reflect standard variety structures.

| grammatical functions |
|---|
| subject (SUBJ) |
| accusative object (OBJA) |
| dative object (OBJD) |
| determiner (DET) |
| sentence (S) |



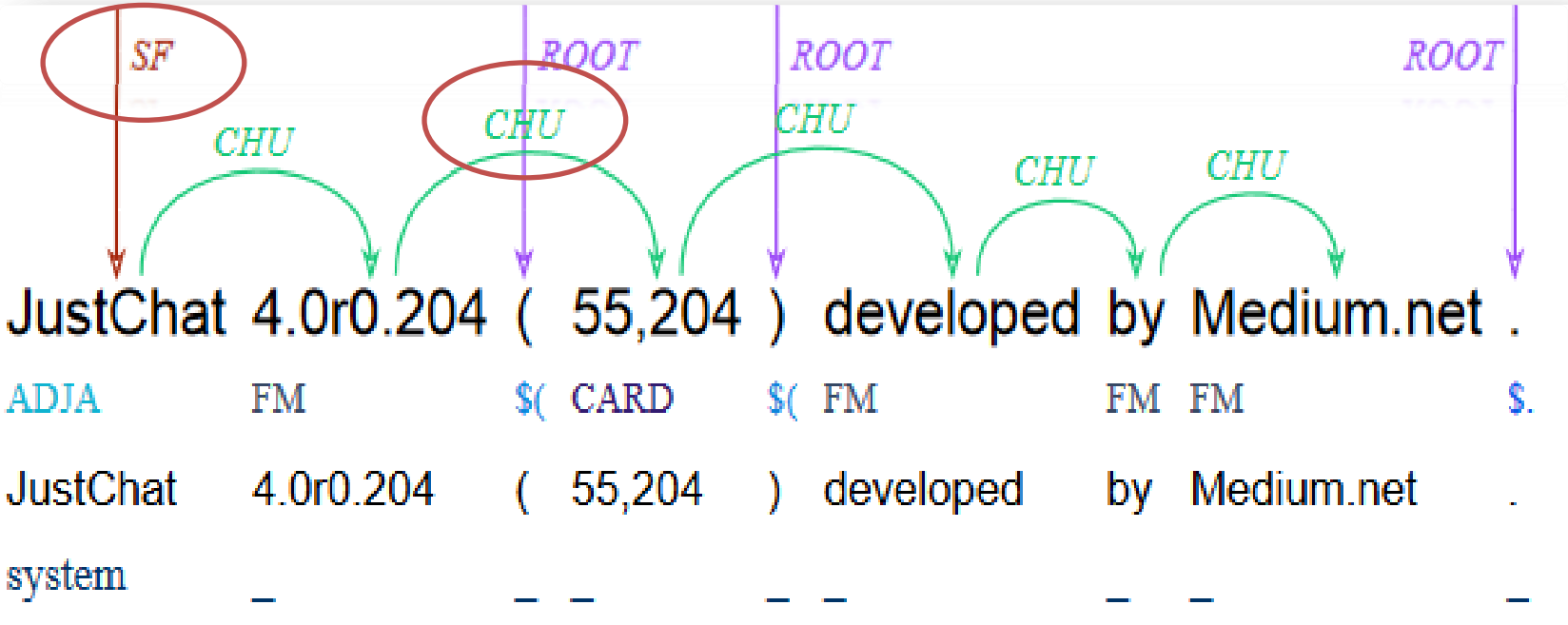

**system** Du betrittst den Raum .

PPER VVFIN ART NN $.

http://www.chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/unicum_21-02-2003_1.html

# Extension of label set

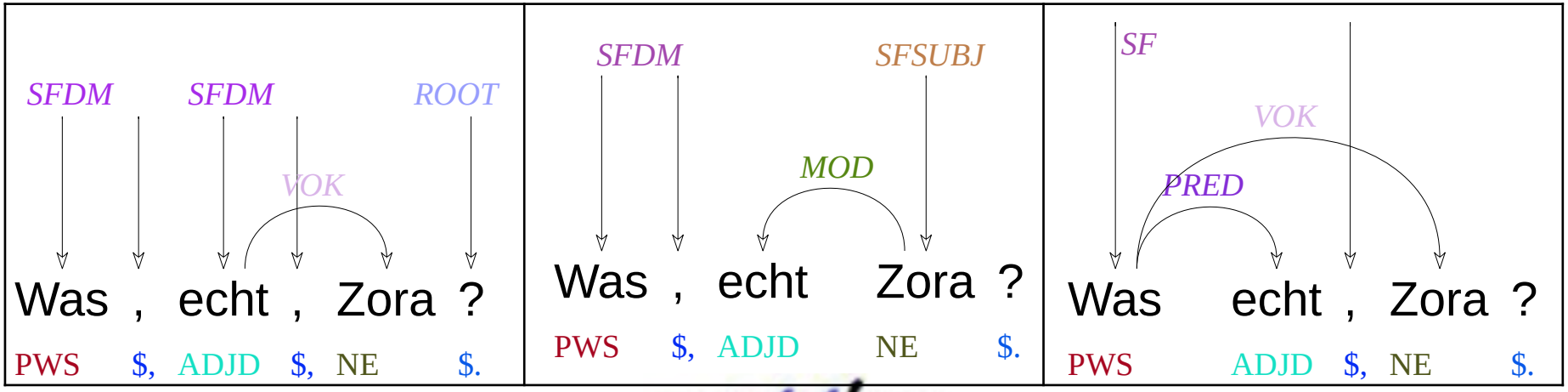
- some need new labels
  - SF → sentence fragment
  - CHU → chunk (non-hierachical multi-word unit)



# Normalisation & Context

- start of chat : missing context → ambiguous parse

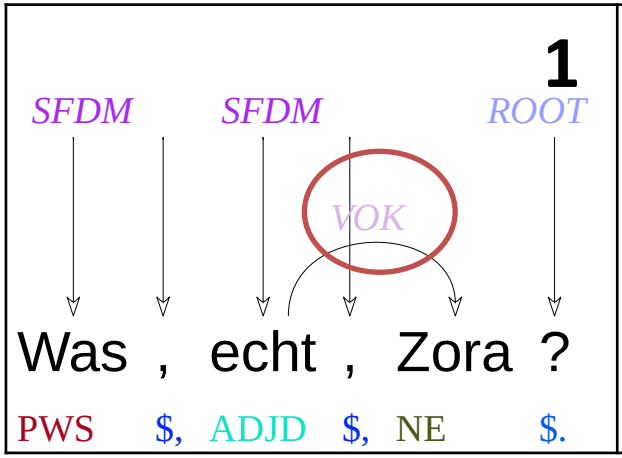
```
1    system   JustChat 4.0r0.204 (55.204) developed by
              Medium.net.
2    system   Du betrittst den Raum.
3    quaki    was echt zori?
4    system   little15 betritt den Raum.
5    quaki    das küssen??
```



# Normalisation & Context

- reconstructing context

| | | | |
|---|---|---|---|
| **1** | 2 | Zora: | Yesterday I kissed someone. |
| | 3 | **quaki:** | Was, ECHT, Zora? |

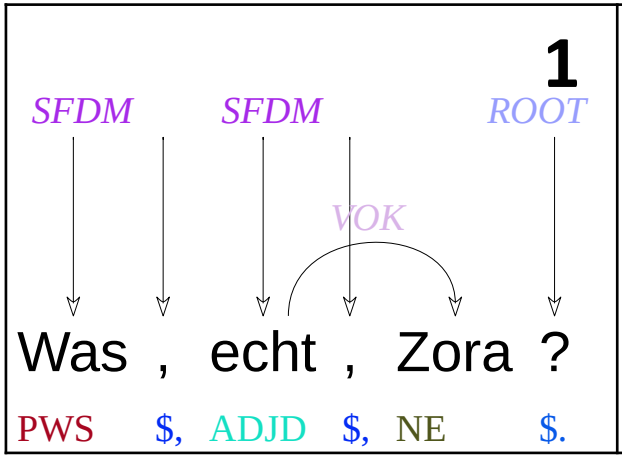


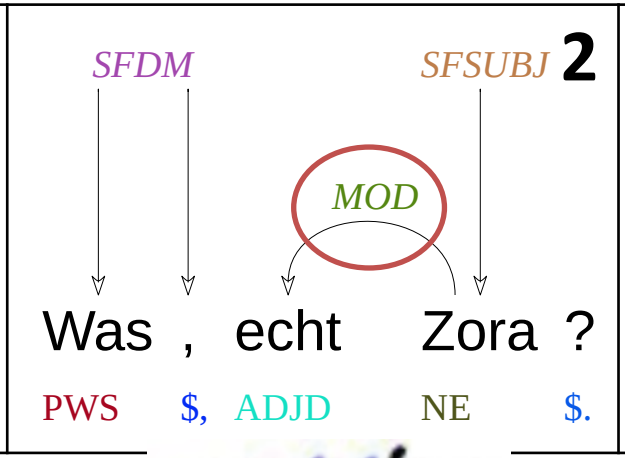
# Normalisation & Context

- reconstructing context

| | | | |
|---|---|---|---|
| **1** | 2 | Zora: | Yesterday I kissed someone. |
| | 3 | **quaki:** | Was, ECHT, Zora? |

| | | | |
|---|---|---|---|
| **2** | 2 | Pharao: | Did you know that Zora kissed someone yesterday? |
| | 3 | **quaki:** | Was, ECHT Zora (hat das gemacht)? |



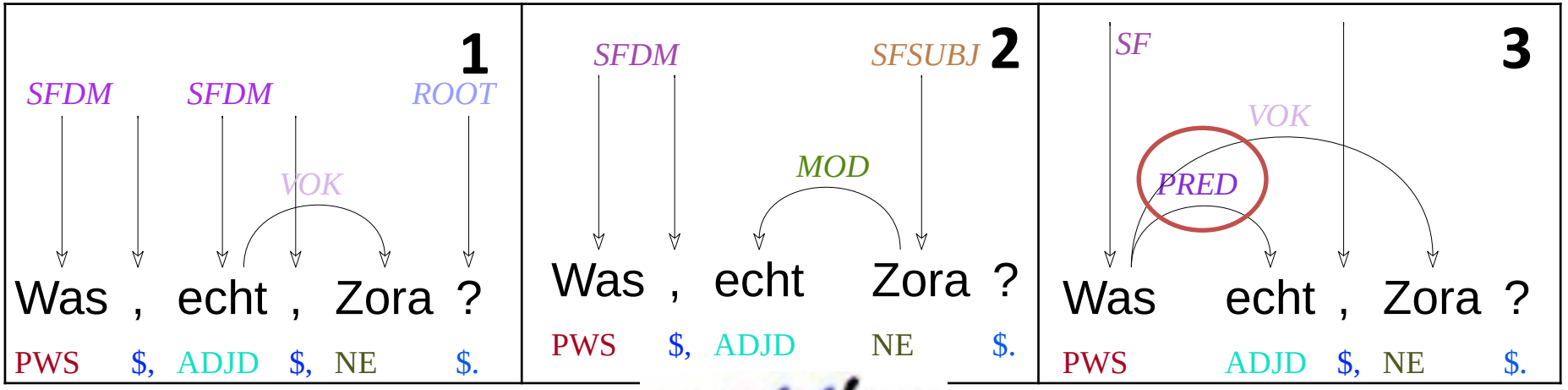
# Normalisation & Context

- reconstructing context

**1**
2 Zora: Yesterday I kissed someone.
3 **quaki:** Was, ECHT, Zora?

**2**
2 Pharao: Did you know that Zora kissed someone yesterday?
3 **quaki:** Was, ECHT Zora (hat das gemacht)?

**3**
2 Zora: Did you really (*echt*) kiss someone yesterday?
3 **quaki:** Was (heißt) ECHT, Zora?



# Normalisation & Context

- alternative: Don't annotate first n postings!

```
 1    system  JustChat 4.0r0.204 (55.204) developed by
              Medium.net.
 2    system  Du betrittst den Raum.
 3     quaki  was echt zori?
 4    system  little15 betritt den Raum.
 5     quaki  das küssen??
 6    Pharao  na gut marc. kein servicepaket nr.1 für
              dich :)
 7      zora  was?
 8    system  TomcatMJ kommt aus dem Raum Go-Rin-No-
              Sho herein.
 9  TomcatMJ  hi
10    system  TomcatMJ ist wieder da.
```

# Fragments and dependencies

The root of a (German) dependency structure is the verb.

→ Fragments are difficult to model.

TiGer Guidelines:

*Bei verblosen Sätzen, die v.a. in Überschriften und Titeln erscheinen, sollte man den Satz **in Gedanken sinnvoll ergänzen** und ihn dann ganz normal annotieren.*

(Albert et al 2003:72)

Verbless sentences as in newpaper titles **should be completed in a sensible way** and then be annotated as usually.

# Fragments and normalisation

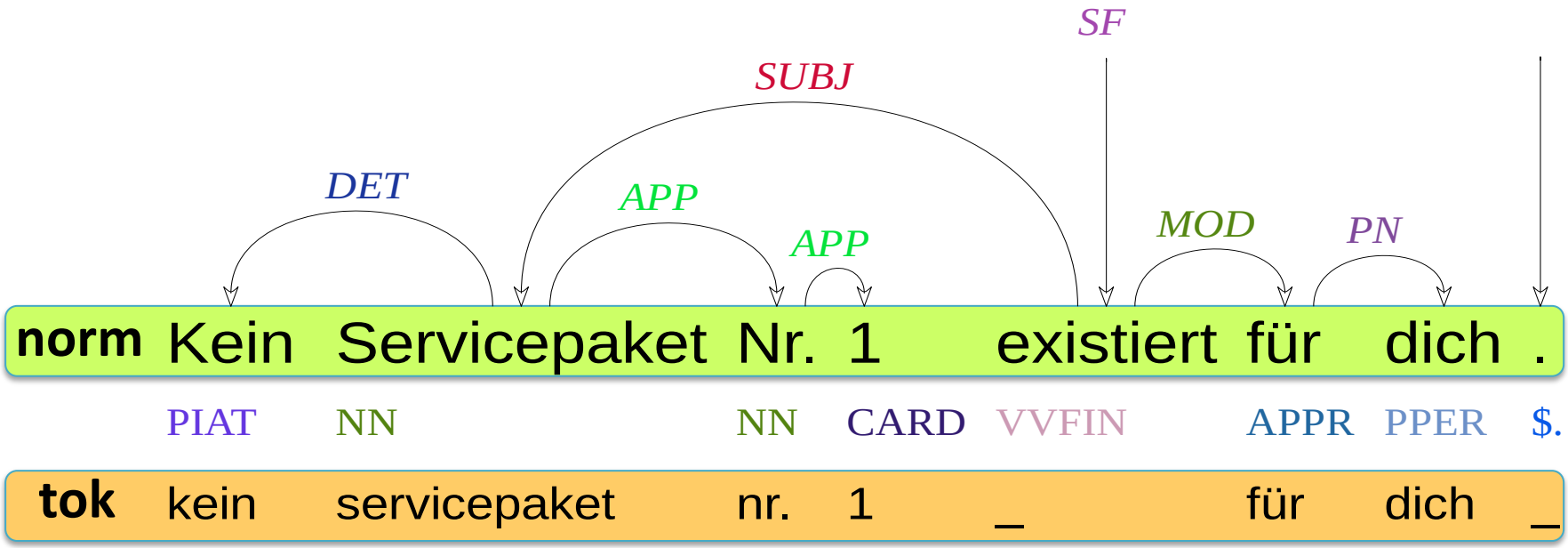
- Normalisations in NoSta-D
  - **are made explicit in the corpus.**
  - **are documented in the manual.**

    (detailed discussion, e.g. Lüdeling et al. 2005, Lüdeling 2008, Reznicek et al. to appear)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **norm** | Kein | Servicepaket | Nr. | 1 | existiert | für | dich | . |
| | PIAT | NN | NN | CARD | VVFIN | APPR | PPER | $. |
| **tok** | kein | servicepaket | nr. | 1 | _ | für | dich | _ |

# Fragments

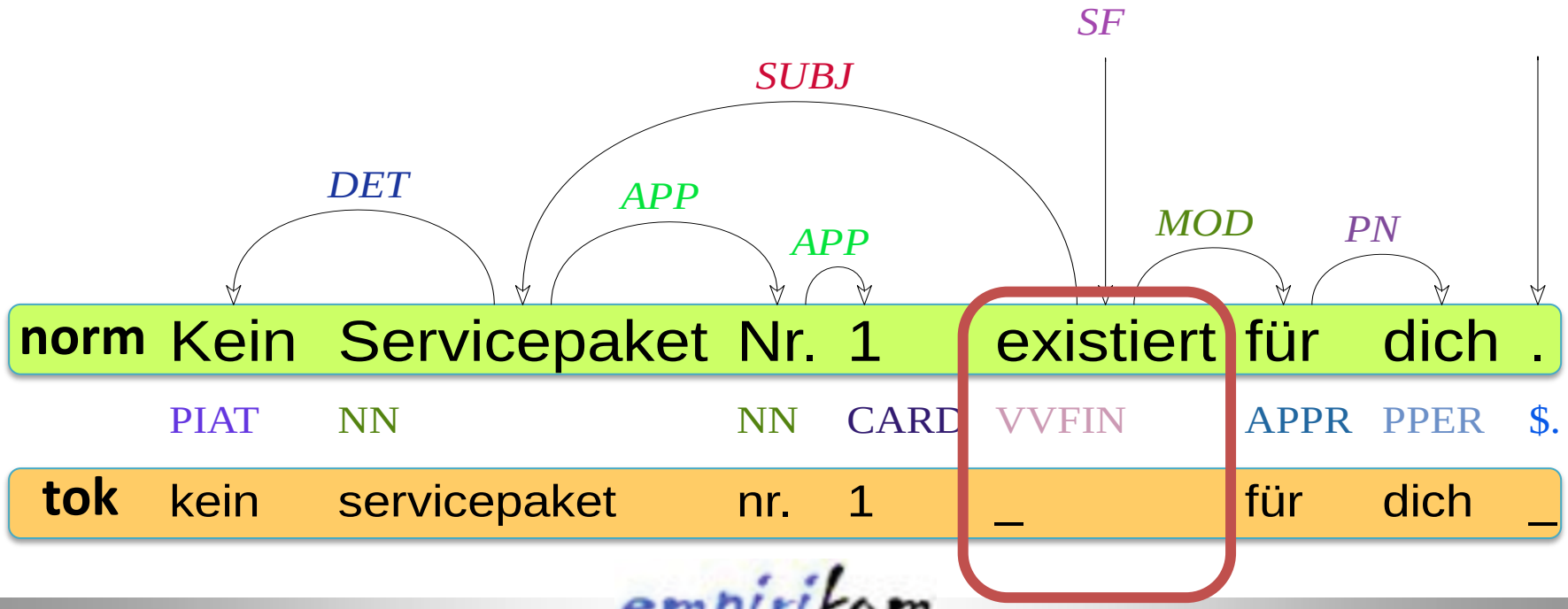
- For **normalisation** a verb is reconstructed (**norm**)
- Grammatical functions are annotated
  - If possible: subject > acc obj. > dat obj.



# Fragments

- For **normalisation** a verb is reconstructed (**norm**)
- Grammatical functions are annotated
  - If possible: subject > acc obj. > dat obj.

**How to deal with missing verbs in the original data?**



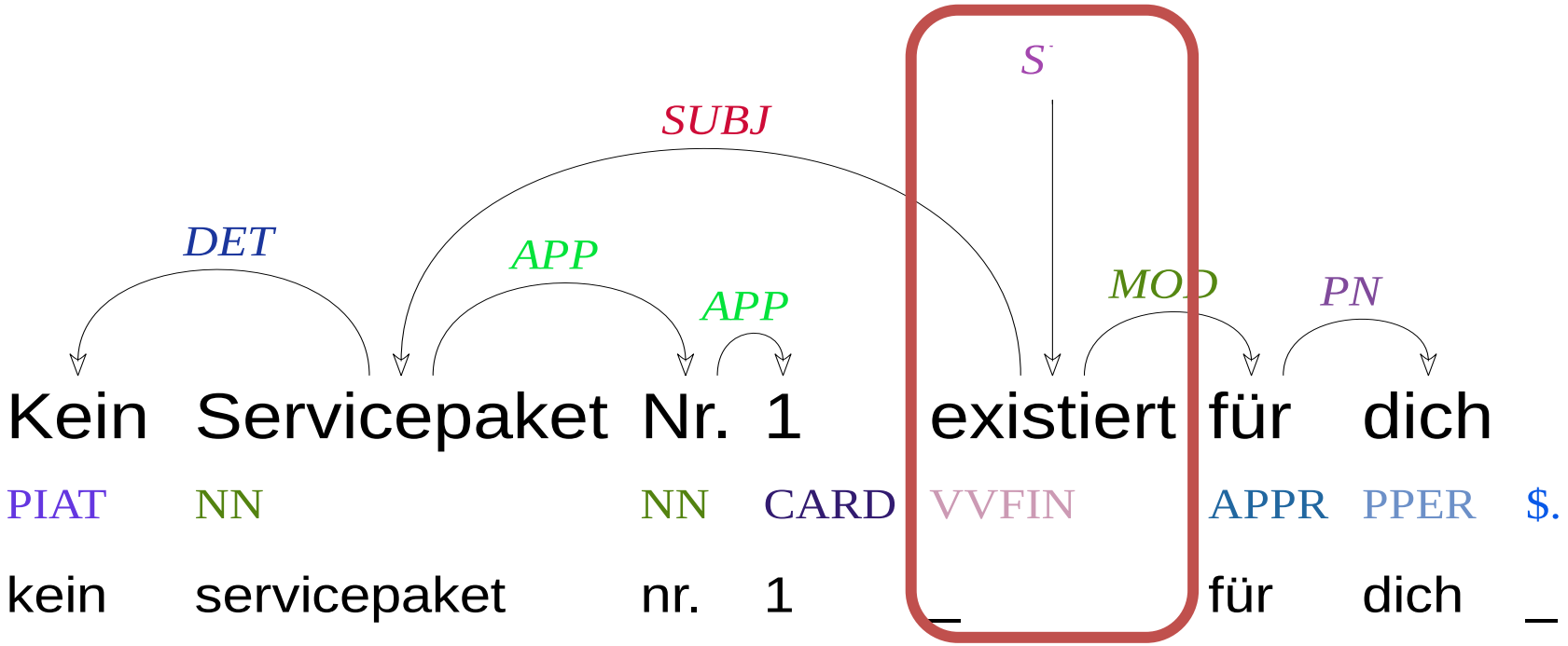
# Fragments

- **3 possible ways modeling the attachement**

# Fragments

- **3 possible ways modeling the attachement**

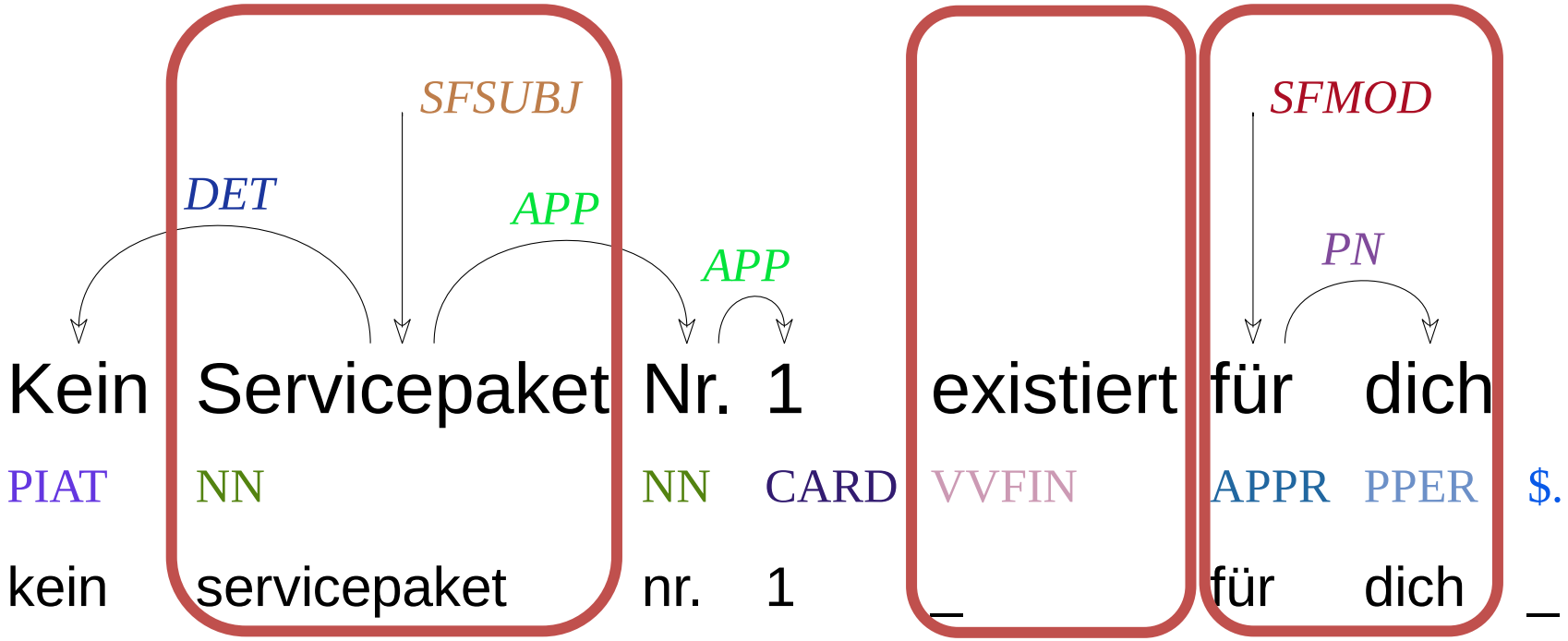
**a) Dummy verb is highest head** (Seeker & Kuhn 2012)



# Fragments

- **3 possible ways modeling the attachement**

**b) Verb is not annotated.** (Foth 2006)
**Fragments are not linked.**

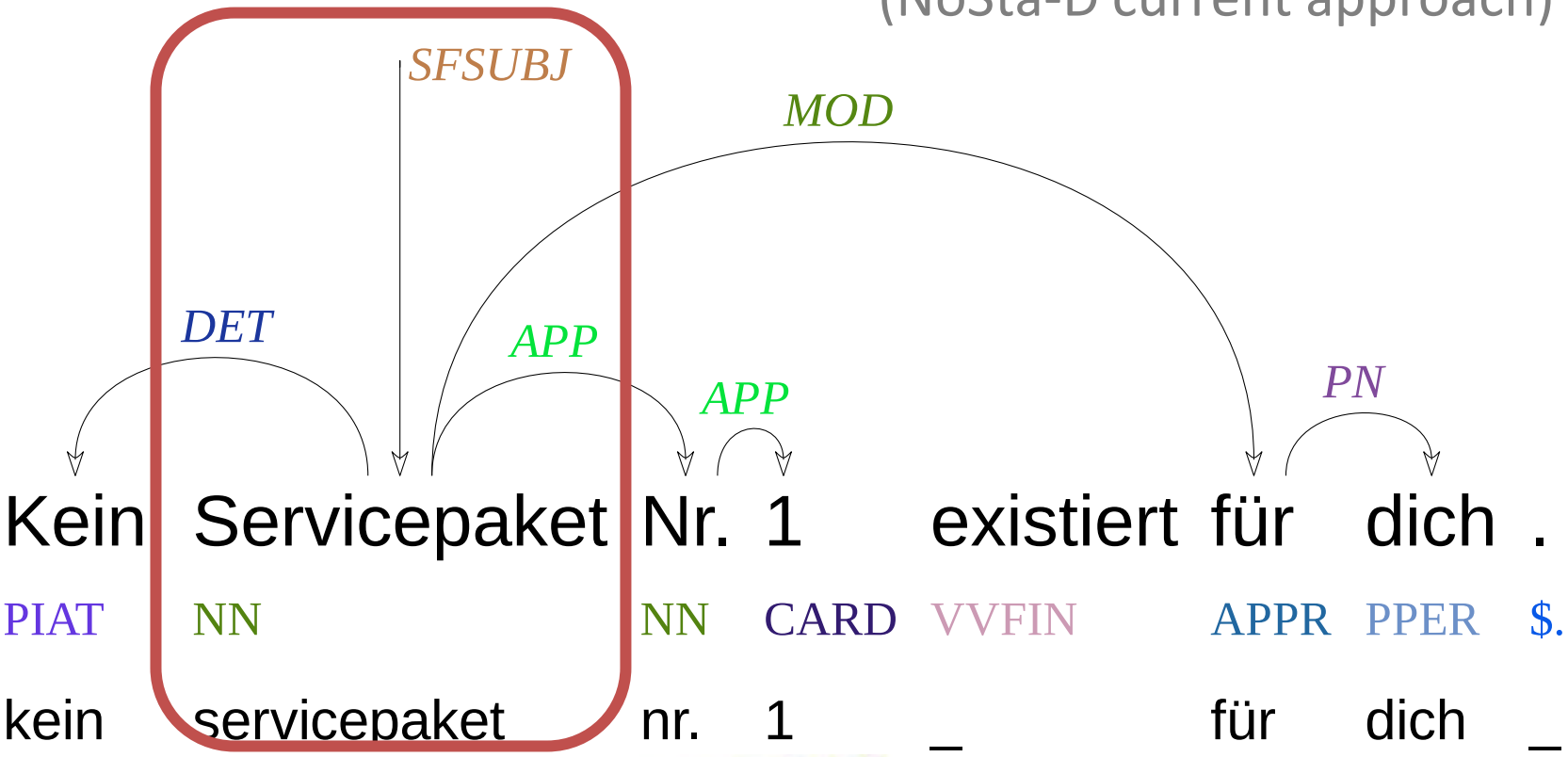


# Fragments

- **3 possible ways modeling the attachement**

**c) All fragments are linked to the highest head.**

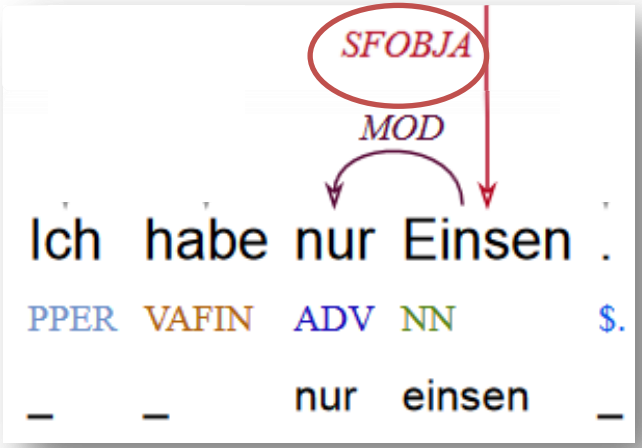
(NoSta-D current approach)



# Fragments

- Fragment roots are assigned grammatical functions where possible

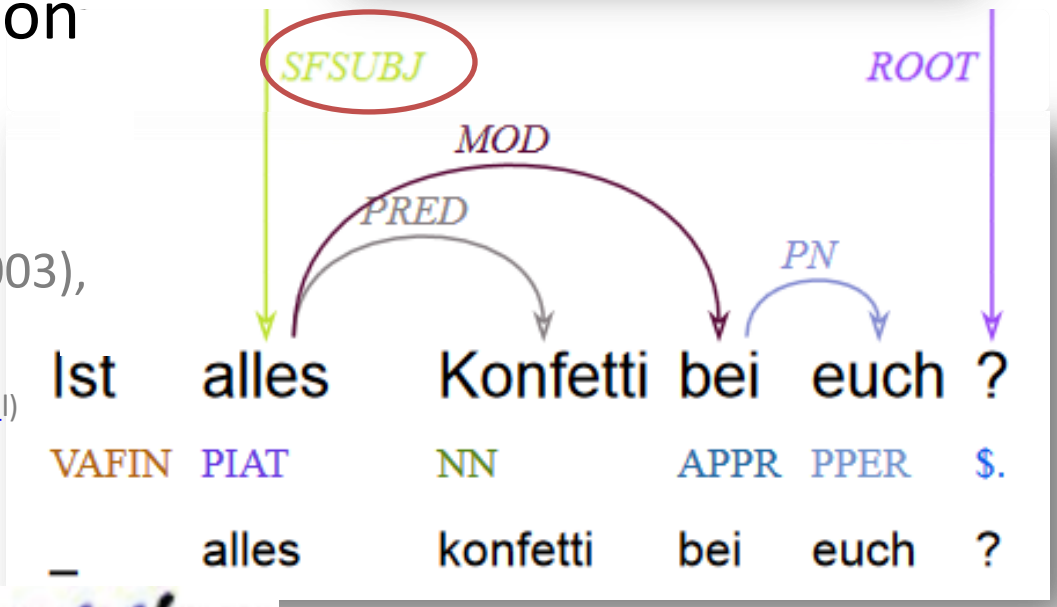
  SF + gram. funct.




- Redundant SF-annotation helpful in up-to-date query tools

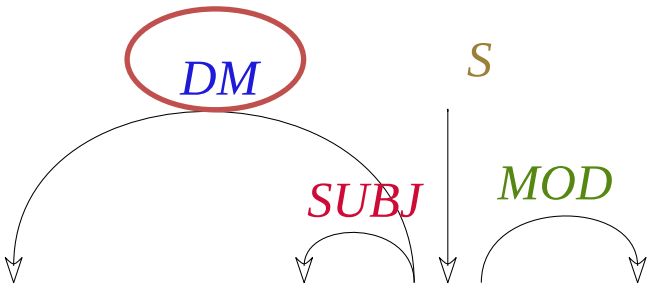
  e.g. TiGer-Search (König et al. 2003), ANNIS3 (Zeldes et al. 2009)

  http://www.sfb632.uni-potsdam.de/annis/annis3.html)

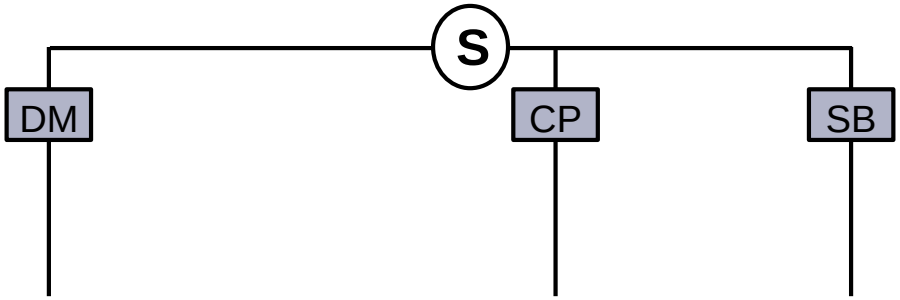


# Interjections and friends

- Linked classic interjections
  → Discourse marker (**DM**)




| Juhu | , | Tom | ist | da | . |
|---|---|---|---|---|---|
| ITJ | $, | NE | VAFIN | ADV | $. |
| juhuuu | _ | tom | is | dada | _ |

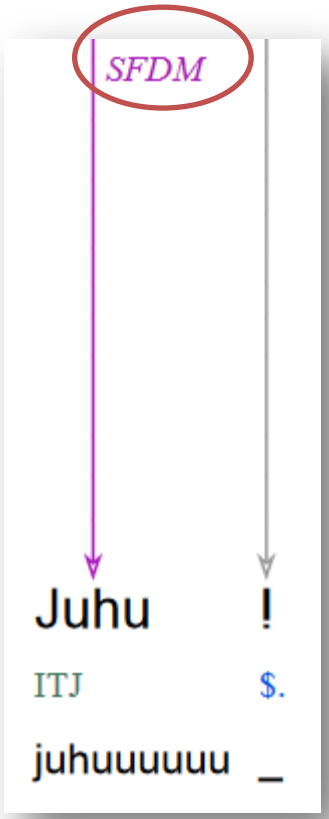



| Ach | , | wenn | ich |
|---|---|---|---|
| ach | | wenn | ich |
| ITJ | $, | KOUS | PPER |

TiGer: s31718

# Interjections and friends

- Non-linked classic interjections
  → Discourse marker fragments (**SFDM**)




```
28 Lantonie   Hallo. :)
29 zora       LANTOOO :)))
30 TomcatMJ   *mal guck wo quaki sich
              nu hinstelt*G*
31 quaki      freu
32 zora       juhuuu
33 Lantonie   Hallo quaki.
34 marc30     Lantööööö :o)
35 TomcatMJ   hi lanto
```

# Interjections and friends

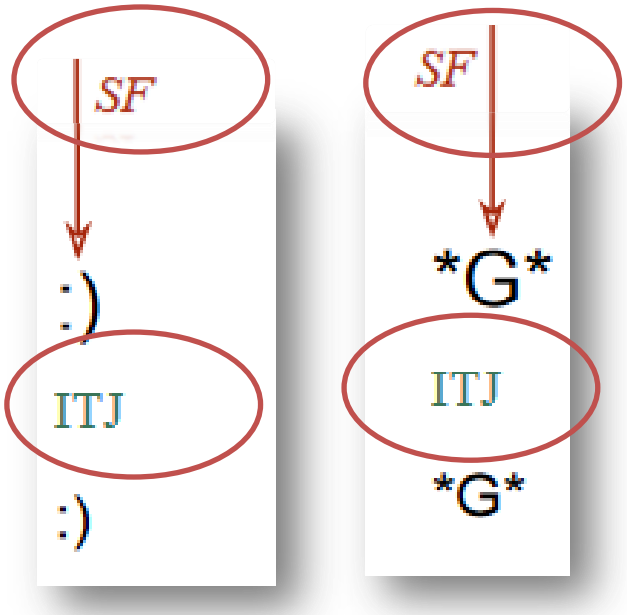
- Emoticons and *-expressions are …
  … tagged as interjections.

  … always considered non-linked fragments when peripheral.

  … of an underspecified kind.

111 Emon mann, habe tatsächlich was verdauliches gegessen... :)

112 TomcatMJ da is sone etwa 25 m hohe pappel wo die drinsitzen und rumzetern*G*



# Underspecification

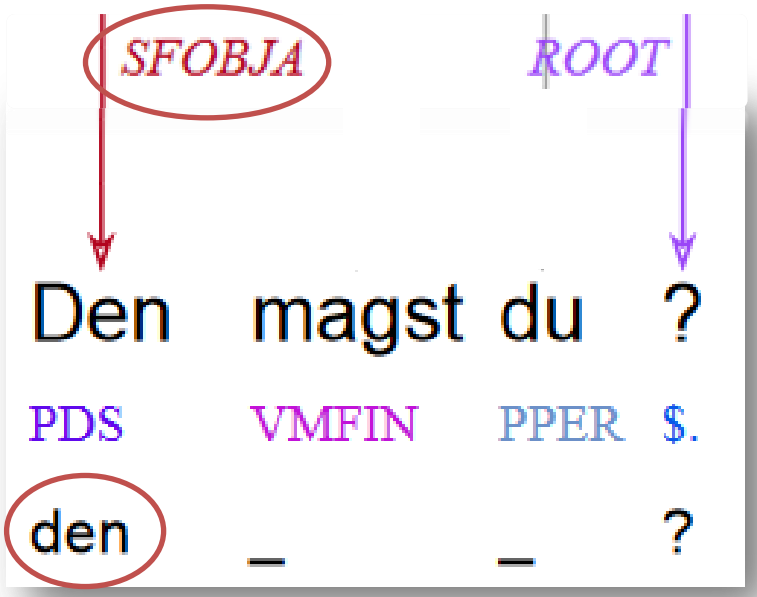
- Grammatical functions are only assigned if unambiguous.

44 Lantonie Ich **finde den quaki klasse**, ein toller Neuzugang, der sich echt bewährt.

45 Lantonie :)))

46 Lantonie Na, zori? :))

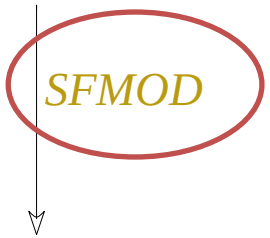
47 marc30 **den?**



# Underspecification

- Grammatical functions are only assigned if unambiguous.

```
12 quaki     juhuuu tom is dada
13 zora      echt?
```




| Meinst | du | das | echt | ? |
|---|---|---|---|---|
| VAFIN | PPER | PDS | ADJD | $. |
| _ | _ | _ | echt | ? |

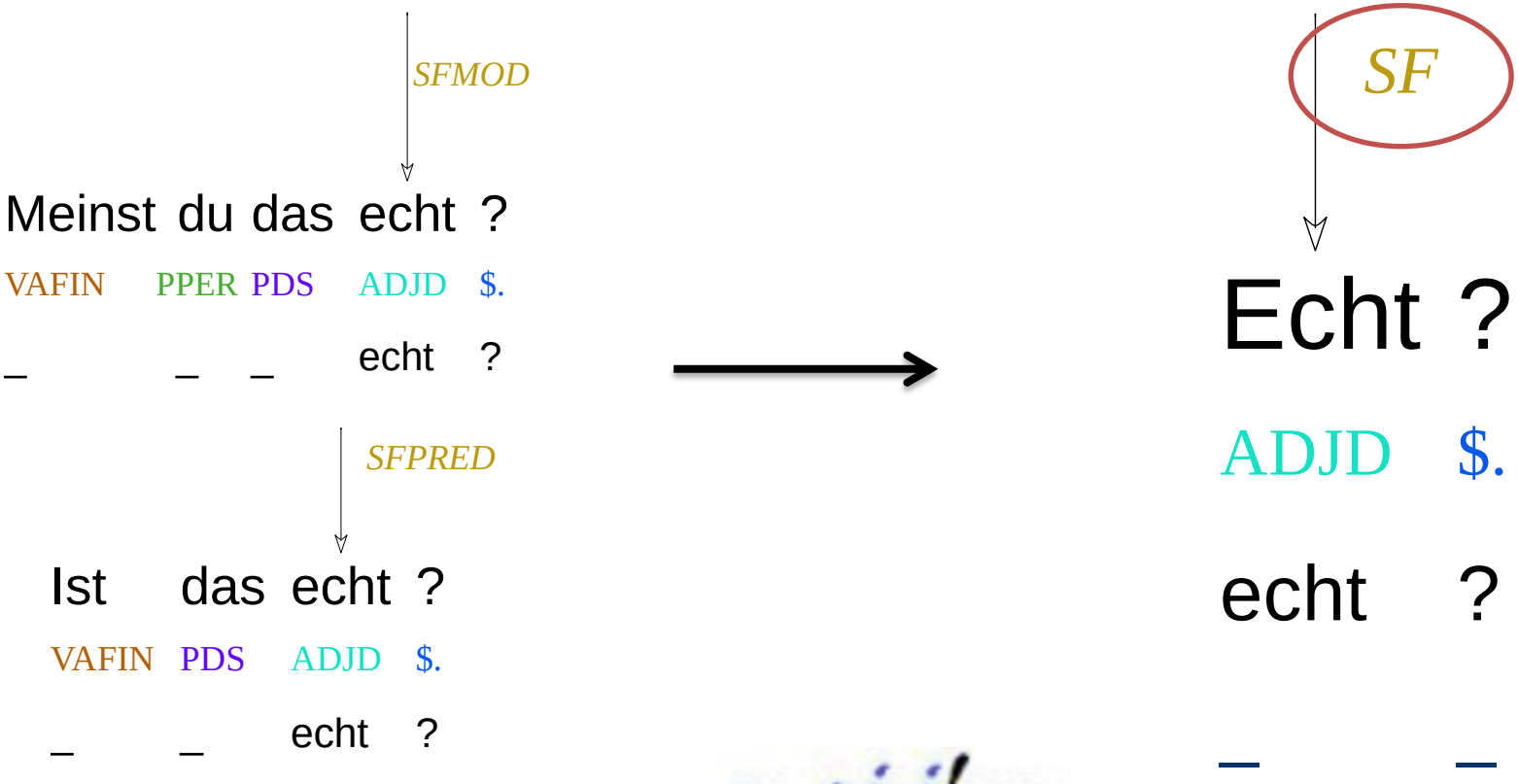

| Ist | das | echt | ? |
|---|---|---|---|
| VAFIN | PDS | ADJD | $. |
| _ | _ | echt | ? |

# Underspecification

- Grammatical functions are only assigned if unambiguous.

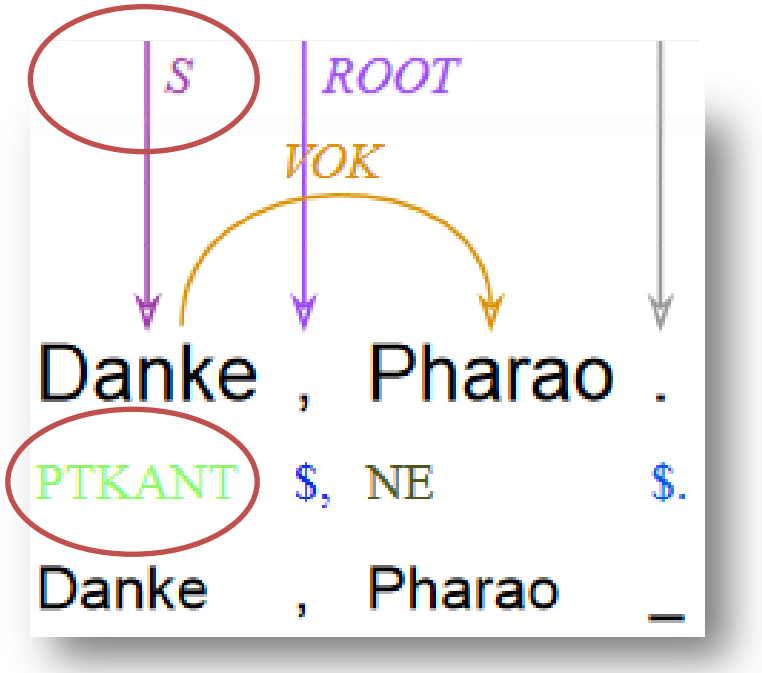
12 **quaki** juhuuu tom is dada

13 **zora** **echt?**



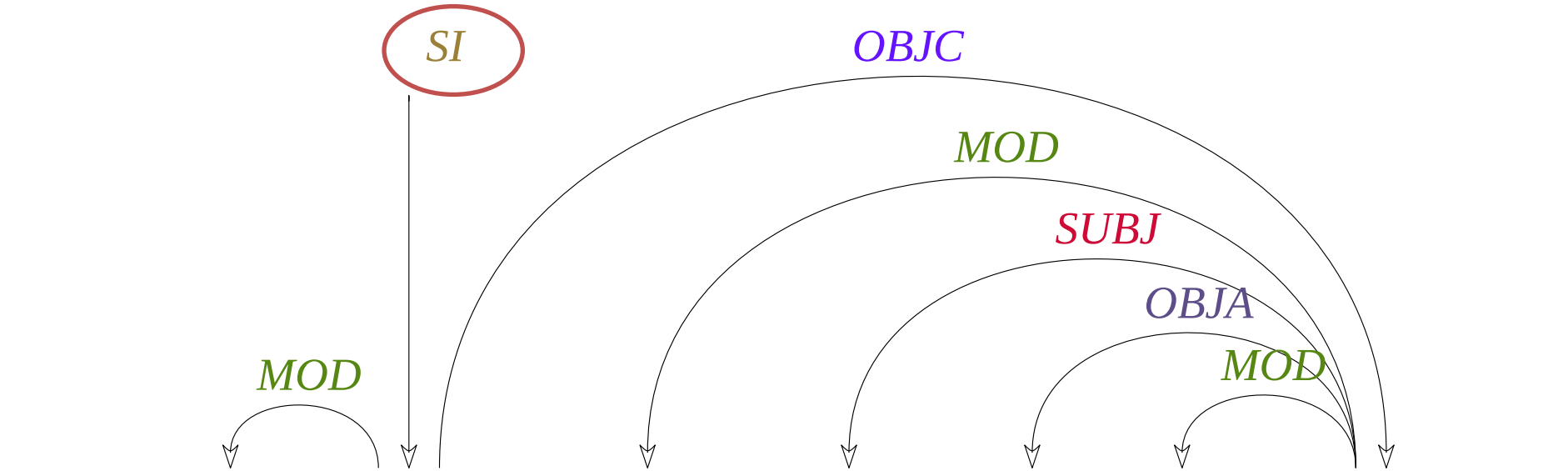
# sentence quivalences

- Responsive particles are treated as full sentences.

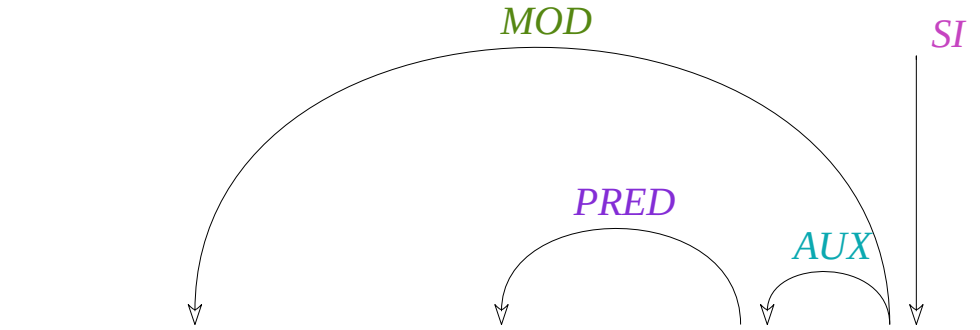


# Inflectives I

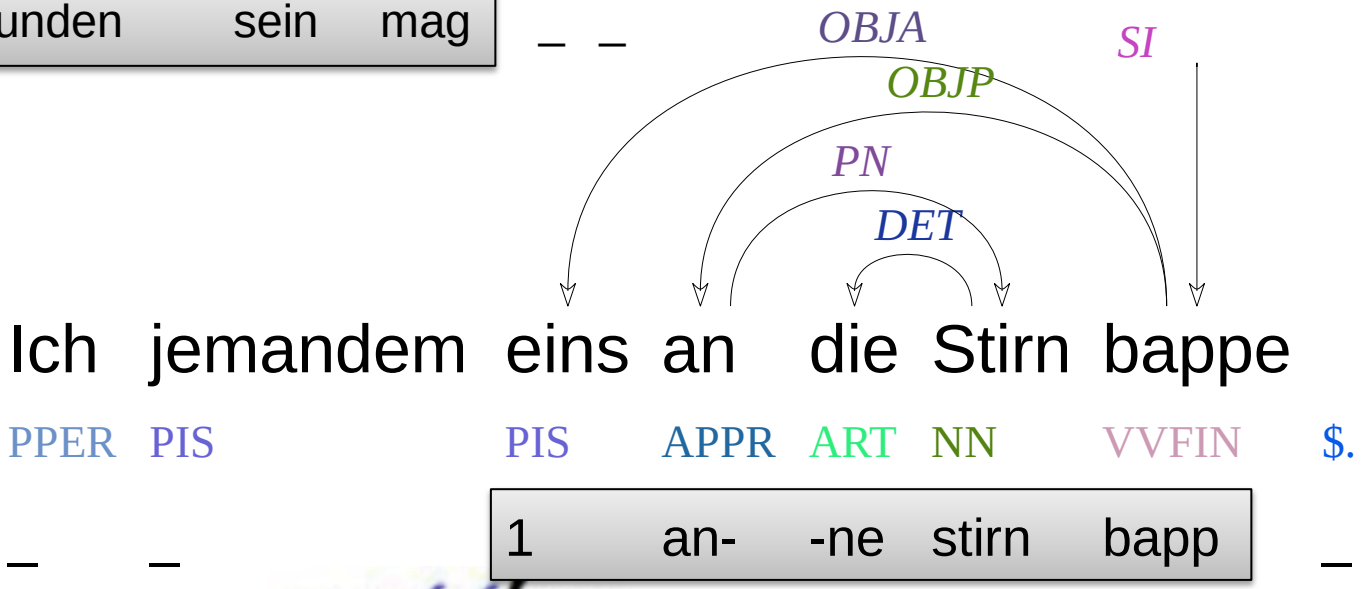
- Inflectives are labeled "**SI**".



# Inflectives II




- For normalisation inflectives are **$V_{end}$ sentences.**



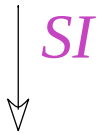
# Inflectives II


Ich mich aufplustere

PPER PPER VVFIN $.

_ _ aufpluster _

- For normalisation inflectives are **$V_{end}$ sentences.**



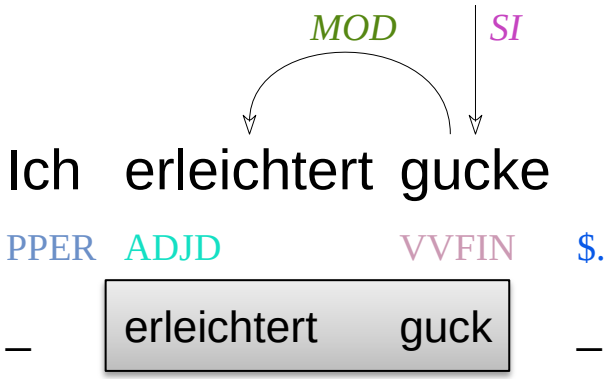

Ich mich freue

PPER PPER VVFIN $.

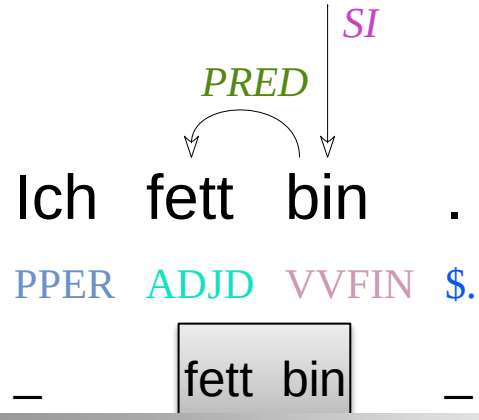
_ _ freu _

# Inflectives III retokenization

11 marc30 Danke Pharao ***erleichtertguck***




299 marc30 ***fettbin***

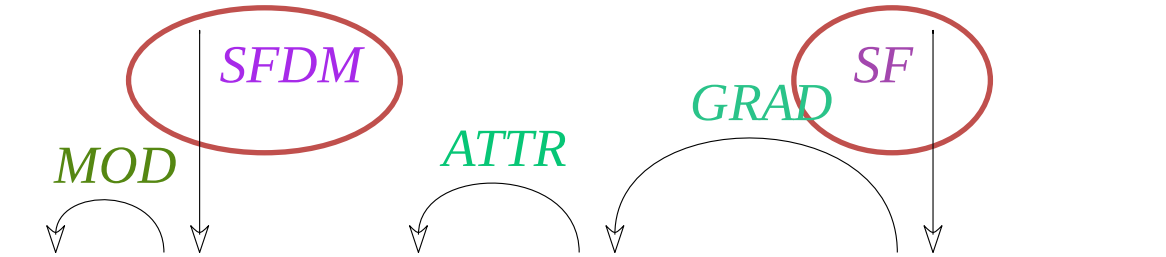



- Concatenations are retokenized into separate words.
- In this annotation pilot we do not worry about automatic performance.

# Asterisk expressions (retokenized)

- Not all concatinated tokens are inflectives.

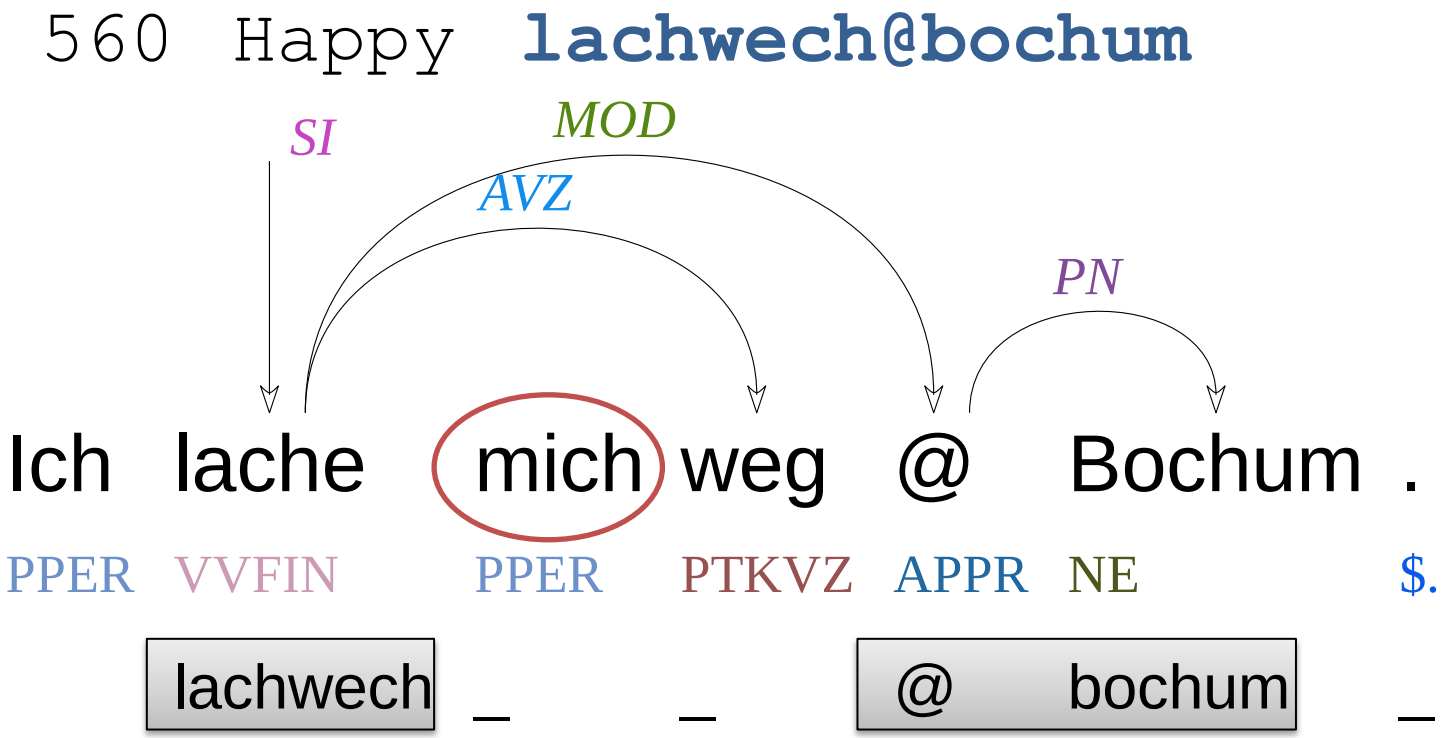
165 quaki ***nagut50cmlauflaufleine***




| na | gut | 50 | cm | lauflaufleine |
|---|---|---|---|---|

# V2-derived inflectives

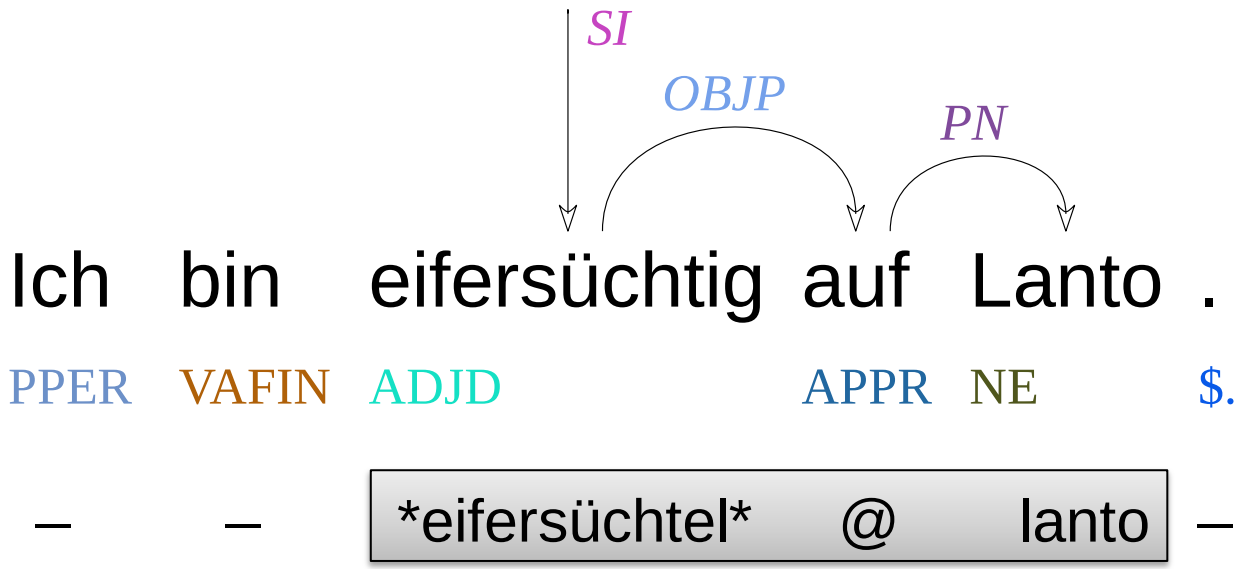
- Not all inflectives are Vend.
  - **Normalisation expects an inserted object here.**



# @ expressions as prepositions

- **@** may replace subcategorized prepositions.

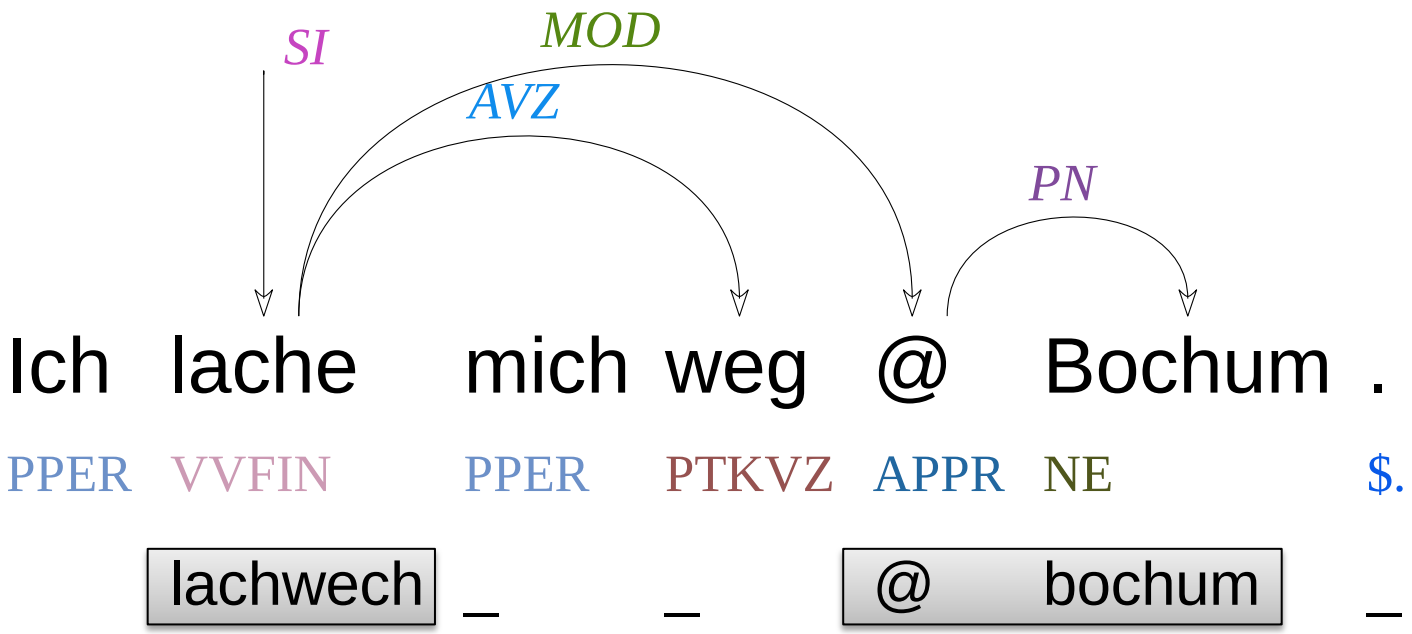
```
323   zora      wos? *eifersüchtel*@lanto
```



# @ expressions as address

- **@** may direct the content of a whole utterance at some adressee.

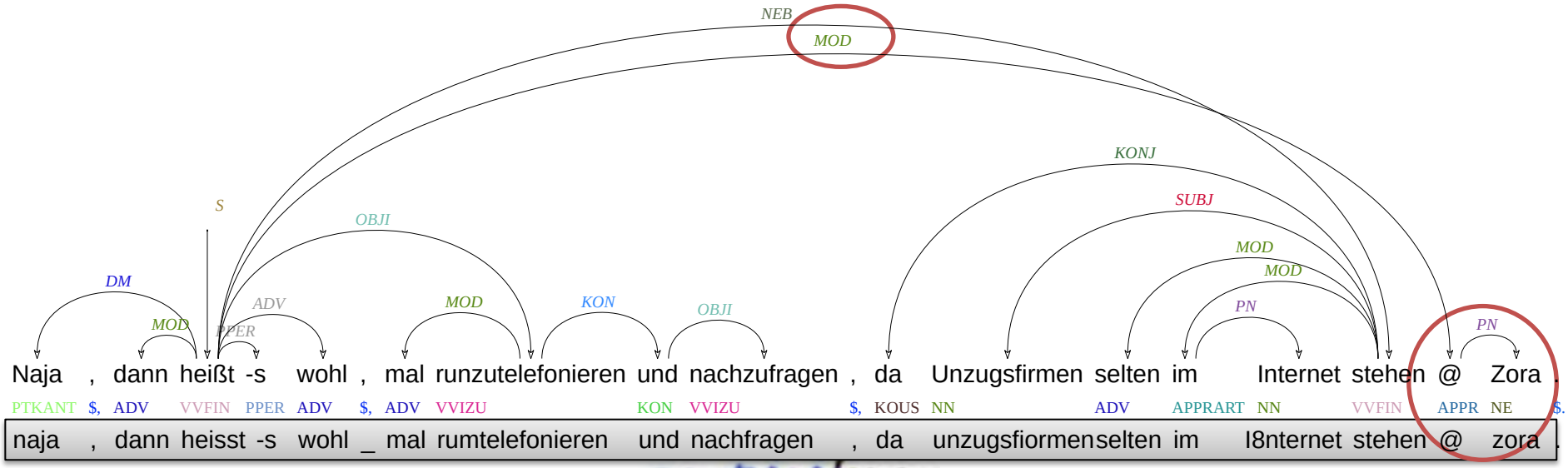
```
560  Happy  lachwech@bochum
```



# @ expressions as address

- **@** may direct the content of a whole utterance at some adressee.

269 TomcatMJ **naja,dann heissts wohl mal rumtelefonieren und nachfragen da umzugsfiormen selten im i8nternet stehen@zora**

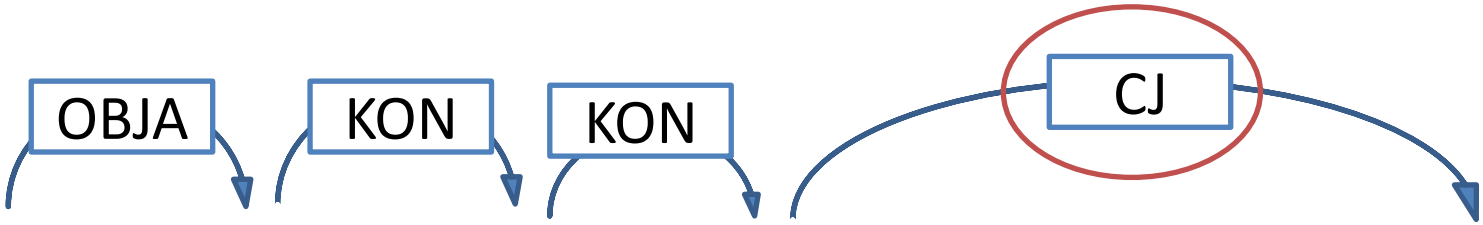


# extra: Coordination

- Classical problem for dependencies

Solution 1) **KON & CJ** (Foth 2006)

**Problem: What category does CJ have?**



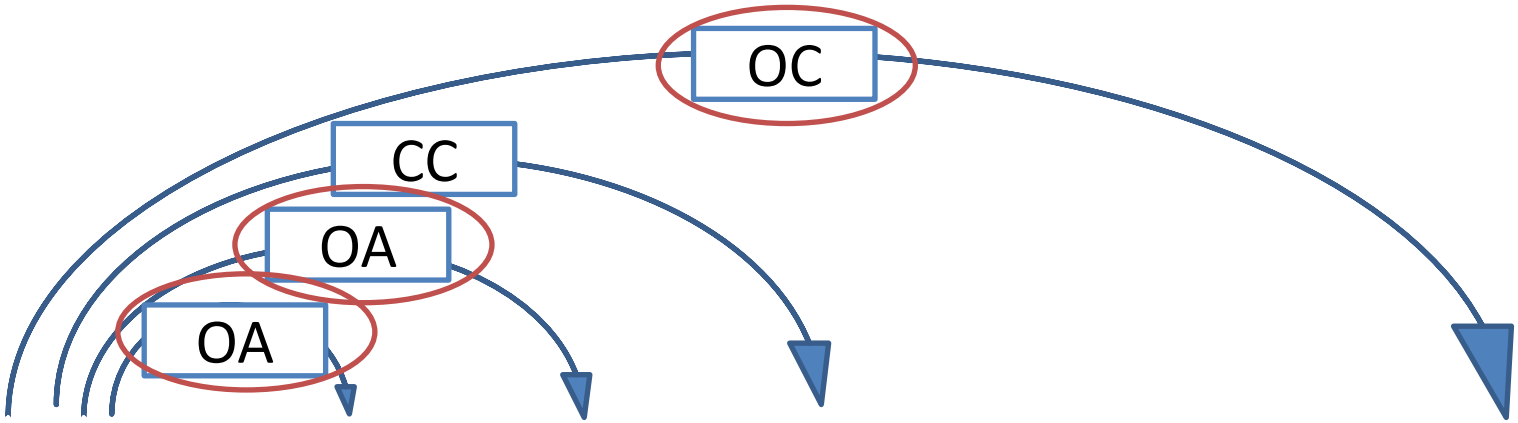

| Ich | liebe | Kaffee | , | Zucker | und | in | der | Sonne | sitzen |
|---|---|---|---|---|---|---|---|---|---|
| I | love | coffee | , | suggar | and | in | the | sun | sitting |

# extra: Coordination

- Classical problem for dependencies

Solution 2) **loose CC & gram. funct.** (Kübler et al. 2012)

**Problem: Which daughters are coordinated?**



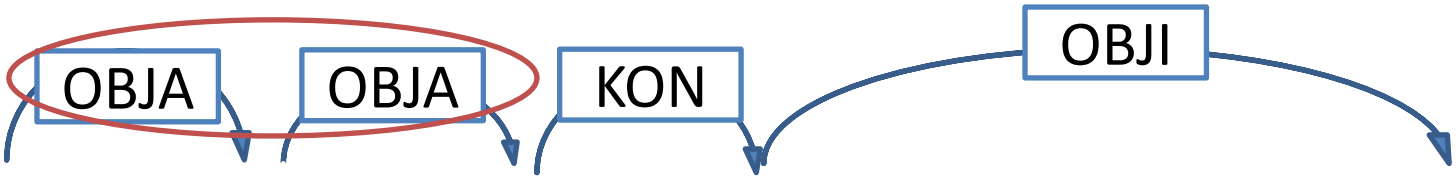

| Ich | liebe | Kaffee | , | Zucker | und | in | der | Sonne | sitzen |
|---|---|---|---|---|---|---|---|---|---|
| I | love | coffee | , | suggar | and | in | the | sun | sitting |

# extra: Coordination

- Classical problem for dependencies

Solution 3) **KON & GF** (NoSta-D)

**Problem: Obj-Obj chains**




| Ich | liebe | Kaffee | , | Zucker | und | in | der | Sonne | sitzen |
|---|---|---|---|---|---|---|---|---|---|
| I | love | coffee | , | suggar | and | in | the | sun | sitting |

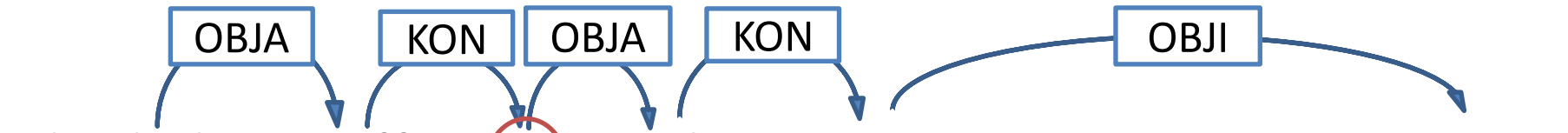
# extra: Coordination

- Classical problem for dependencies

Solution 4) **KON & GF with ',' as KON**

**Problem: Comma is annotated only in coordinations**

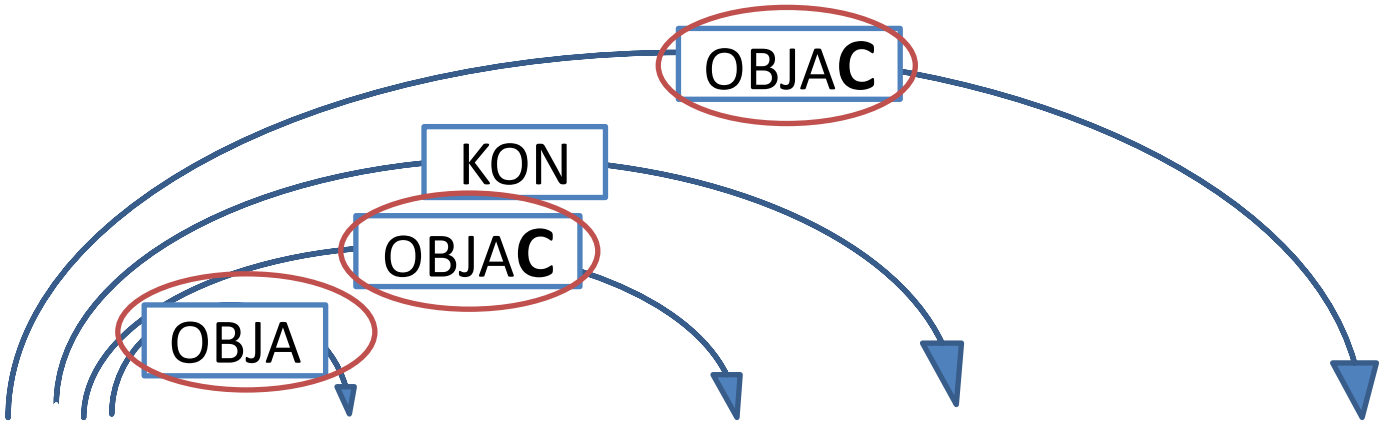
**Only works for annotation of normalisation**



# extra: Coordination

- Classical problem for dependencies

Solution 4) **loose CC &**
**new gram. funct.**




| Ich | nenne | ihn | einen | Hund | und | Ehebrecher |
|---|---|---|---|---|---|---|
| I | call | him | a | dog | and | adulterer |

# Summary

**Tasks:**

- We need guidelines to decide how to deal with the first postings of a chat that don't come with context.
- We need a new POS-tag for inflectives.

# Summary

**Tasks:**

- We need guidelines to decide how to deal with the first postings of a chat that don't come with context.
- We need a new POS-tag for inflectives.

**Questions:**

- What would be a preferable attachment of fragments?
- Are response particles full sentences?
- Is retokenizing of concatenations a helpful analysis?
- Should we analyse concatenations as $V_{end}$?

# Outlook

- Annotation of
  - coreference
  - named entity recognition
    - including crowd-sourcing reference

- Publication of the pilot corpus NoSta-D
  - end of summer

82 stoeps :-)
83 TomcatMJ hi stoeps
84 Emon unf tom : (
85 Emon *g*
86 Thor... über meine
87 TomcatMJ hi emon
88 quaki 200 krähen?
89 TomcatMJ jo..
90 Lantonie Die eins Minus
91 marc30 was ist Benehme
92 quaki die vögel
93 system B67 betritt der
94 system Lantonie ist jetzt weg Pissen

# Thanks ;0)

# Bibliography


**Albert, Stefanie; Anderssen, Jan; Bader, Regine; Becker, Stefanie; Bracht, Tobias; Brants, Thorsten et al. (2003):** TIGER Annotationsschema.. Online http://www.ims.uni-stuttgart.de/projekte/TIGER/ .

**Beißwenger, Michael (2013):** Das Dortmunder Chat-Korpus: ein annotiertes Korpus zur Sprachverwendung und sprachlichen Variation in der deutschsprachigen Chat-Kommunikation. Online-Publikation, LINSE - Linguistik Server Essen http://www.linse.uni-due.de/tl_files/PDFs/Publikationen-Rezensionen/Chatkorpus_Beisswenger_2013.pdf

**Dipper, Stefanie; Lüdeling, Anke; Reznicek, Marc (to appear):** NoSta-D. A Corpus of German Non-Standard Varieties. In: Marcos Zampieri (ed.): Non-Standard Data Sources in Corpus-Based Research: Shaker.

**Foth, Kilian A. (2006):** Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Technischer Report. Universität Hamburg.

**König, Esther; Lezius, Wolfgang; Voormann, Holger (2003):** TIGERSearch User's Manual. Stuttgart. Online verfügbar unter http://www.tigersearch.de.

**Lüdeling, Anke (2008):** Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora. In: Maik Walter und Patrick Grommes (Hg.): Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung. Tübingen: Max Niemeyer Verlag (Linguistische Arbeiten, 520), S. 119–140.

**Lüdeling, Anke; Walter, Maik; Kroymann, Emil; Adolphs, Peter (2005):** Multi-level Error Annotation in Learner Corpora. In: Proceedings of Corpus Linguistics 2005. Birmingham.

**Kübler, Sandra; Prokic, Jelena (2006):** Why is German Dependency Parsing more Reliable than Constituent Parsing? In: Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories (TLT 2006). Prague, Czech Republic. Prague, Czech Republic.

**Reznicek, Marc; Lüdeling, Anke; Hirschmann, Hagen (to appear):** Competing Target Hypotheses in the Falko Corpus. A Flexible Multi-Layer Corpus Architecture. In: Ana Díaz-Negrillo (Hg.): Automatic Treatment and Analysis of Learner Corpus Data: John Benjamins.

**Seeker, Wolfgang; Kuhn, Jonas (2012):** Making Ellipses Explicit in Dependency Conversion for a German Treebank. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey: European Language Resources Association (ELRA), S. 3132–3139. Online http://www.lrec-conf.org/proceedings/lrec2012/pdf/235_Paper.pdf.

**Tenfjord, Kari; Hagen, Jon Erik; Johansen, Hilde (2006):** The "Hows" and the "Whys" of Coding Categories in a Learner Corpus. (or "How and Why an Error-Tagged Learner Corpus is not 'ipso facto' One Big Comparative Fallacy"). In: *Rivista di psicolinguistica applicata* (3), S. 93–108.

**Zeldes, Amir; Ritz, Julia; Lüdeling, Anke; Chiarcos, Christian (2009):** ANNIS. A Search Tool for Multi-Layer Annotated Corpora. In: Michaela Mahlberg, Victorina González-Díaz und Catherine Smith (Hg.): Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009. Corpus Linguistics. Liverpool, 20-23 July 2009. University of Liverpool.