DIRNDL
A Discourse Information Radio News Database for Linguistic Analysis

Kerstin Eckart, Arndt Riester, Katrin Schweitzer

SFB 732, Projects A1 and B3
Universität Stuttgart
Institut für maschinelle Sprachverarbeitung

DGfS 2012, AG2: Linked Data in Linguistics
March 7-9, 2012, Frankfurt/Main

A German radio news corpus

- 3 days of hourly broadcast German radio news (Deutschlandfunk)
- Recorded on March 25-27, 2007
- Primary data – provided with the corpus
  - recordings of the broadcasts (ca. 5 hrs of speech)
  - transcripts of the news text (ca. 3000 sentences)

Two annotated data sets

Speech:

- Based on the recordings
- Manually annotated so far ca. 17,000 GToBI(S) labels (Mayer, 1995) for pitch accents and prosodic phrase boundaries

Written:

- Based on the written news text
- Sentences parsed with XLE:
  German LFG grammar (Rohrer & Forst, 2006)
- Constituent trees manually annotated
  with about 10,000 information status labels (Riester et al. 2010)

Relating information status and prosody: studies

Disambiguation by prosody / infelicitous prosody

(1) {So, is John going to help us?}

Don't ask me ...

a. I haven't SEEN your brother ALL DAY.
b. #I haven't seen your BROTHER all day.

- Context usually imposes strict constraints on the assignment of prosody, but sometimes allows for some flexibility.
- Radio news allows us to study prosodic variation on **repeated** news items.

Relating information status and prosody: studies

(2)

der	EU-Außenbeauftragte	Solana	betonte,	die Tür	zu
the	EU High Representative	Solana	stressed,	the door	for
Verhandlungen	mit	Teheran	bleibe	offen	
negotiations	with	Teheran	remains	open	

Relating information status and prosody: studies

(2) der EU-Außenbeauftragte Solana betonte, die Tür zu
H*L H*L H*
Verhandlungen mit Teheran bleibe offen
!H* !H*L

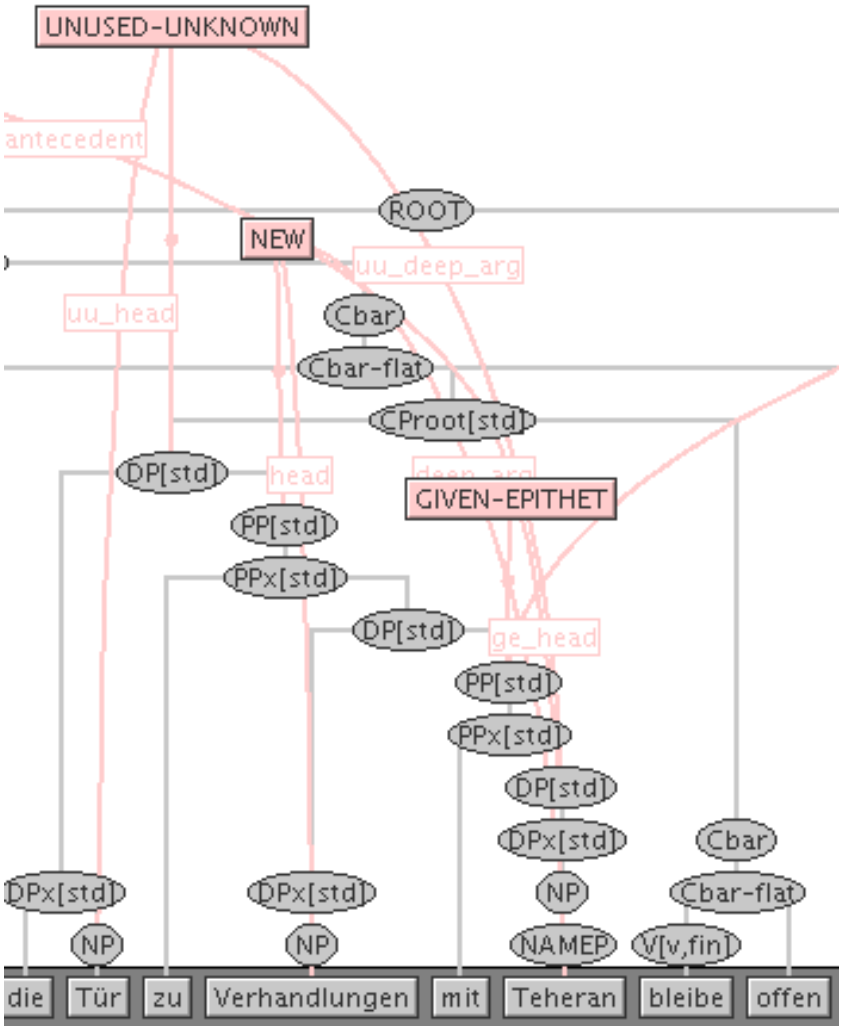
UNUSED-KNOWN

Annotating referential information status

(Riester et al. 2010), (Baumann and Riester, to appear)

Units: definite DPs	
given	anaphor corefers with antecedent in previous discourse
given-sit	symbolic deixis, e.g. discourse participants
bridging	non-coreferring, context-dependent anaphor
unused-known	discourse-new item which is generally known
unused-unknown	discourse-new, context-free item which is not known
Units: definite or indefinite DPs	
cataphor	item whose referent is established later on
generic	generic / abstract / hypothetical item
Units: indefinite DPs	
new	specific indefinite introducing a new referent

Annotation with SALTO



- Enables annotations on syntactic trees and nested embeddings
- Syntactic information can be integrated into queries

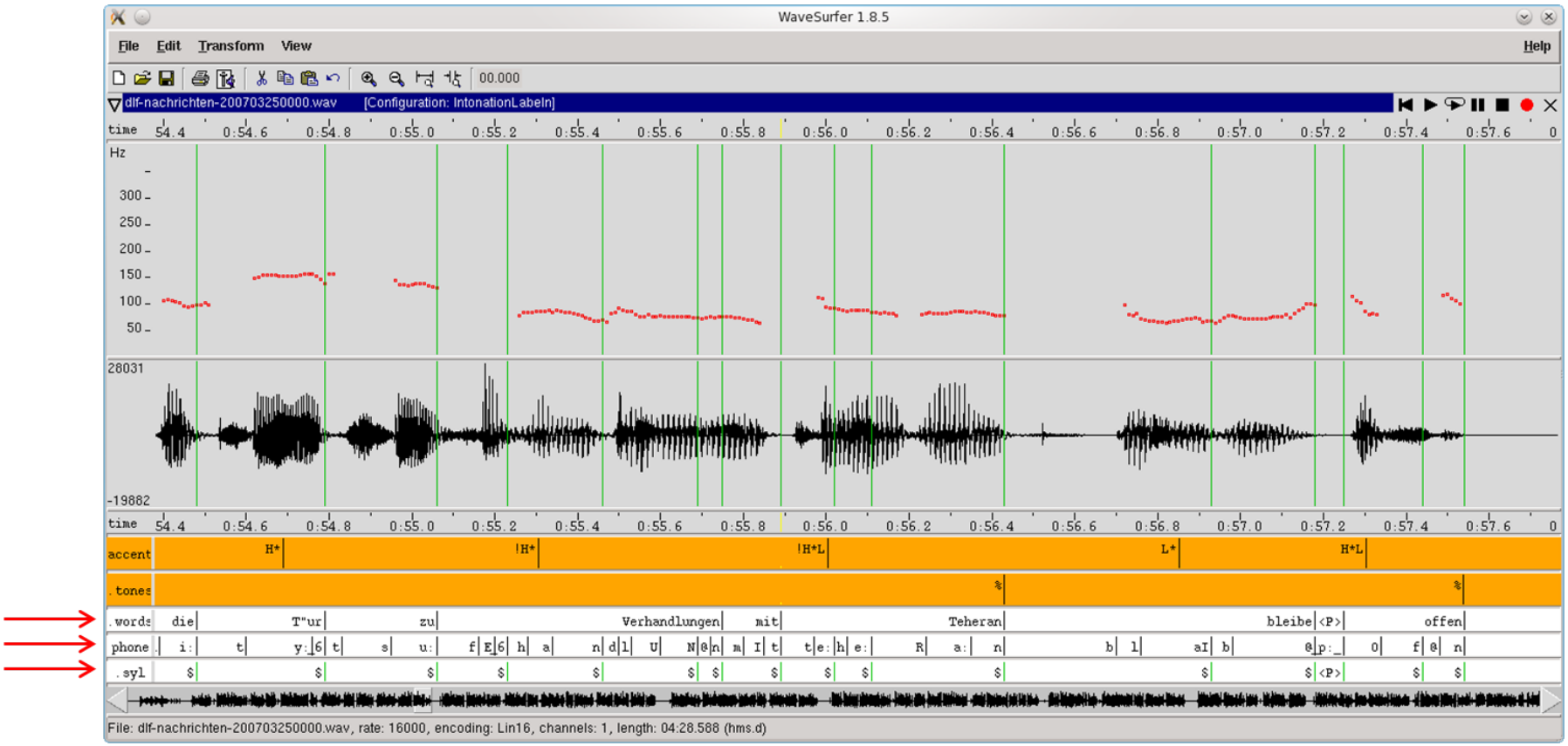
Annotation with SALTO

```
<s id="s7">
 <graph root="s7_500">
  <terminals>
    ...
   <t id="s7_6" word="die" pos="D[std]"/>
   <t id="s7_7" word="Tuer" pos="N[comm]"/>
   <t id="s7_8" word="zu" pos="P[pre]"/>
   <t id="s7_9" word="Verhandlungen" pos="N[comm]"/>
   <t id="s7_10" word="mit" pos="P[pre]"/>
   <t id="s7_11" word="Teheran" pos="NAME"/>
    ...
  </terminals>
  <nonterminals>
    ...
   <nt id="s7_511" cat="DPx[std]">
    <edge label="--" idref="s7_6"/>
    <edge label="--" idref="s7_512"/>
   </nt>
   <nt id="s7_517" cat="NP">
    <edge label="--" idref="s7_9"/>
   </nt>
    ...
  </nonterminals>
 </graph>
  ...
</s>
```

- Enables annotations on syntactic trees and nested embeddings
- Syntactic information can be integrated into queries
- Representation of annotations in Tiger-/Salsa-XML

Annotating speech using WaveSurfer

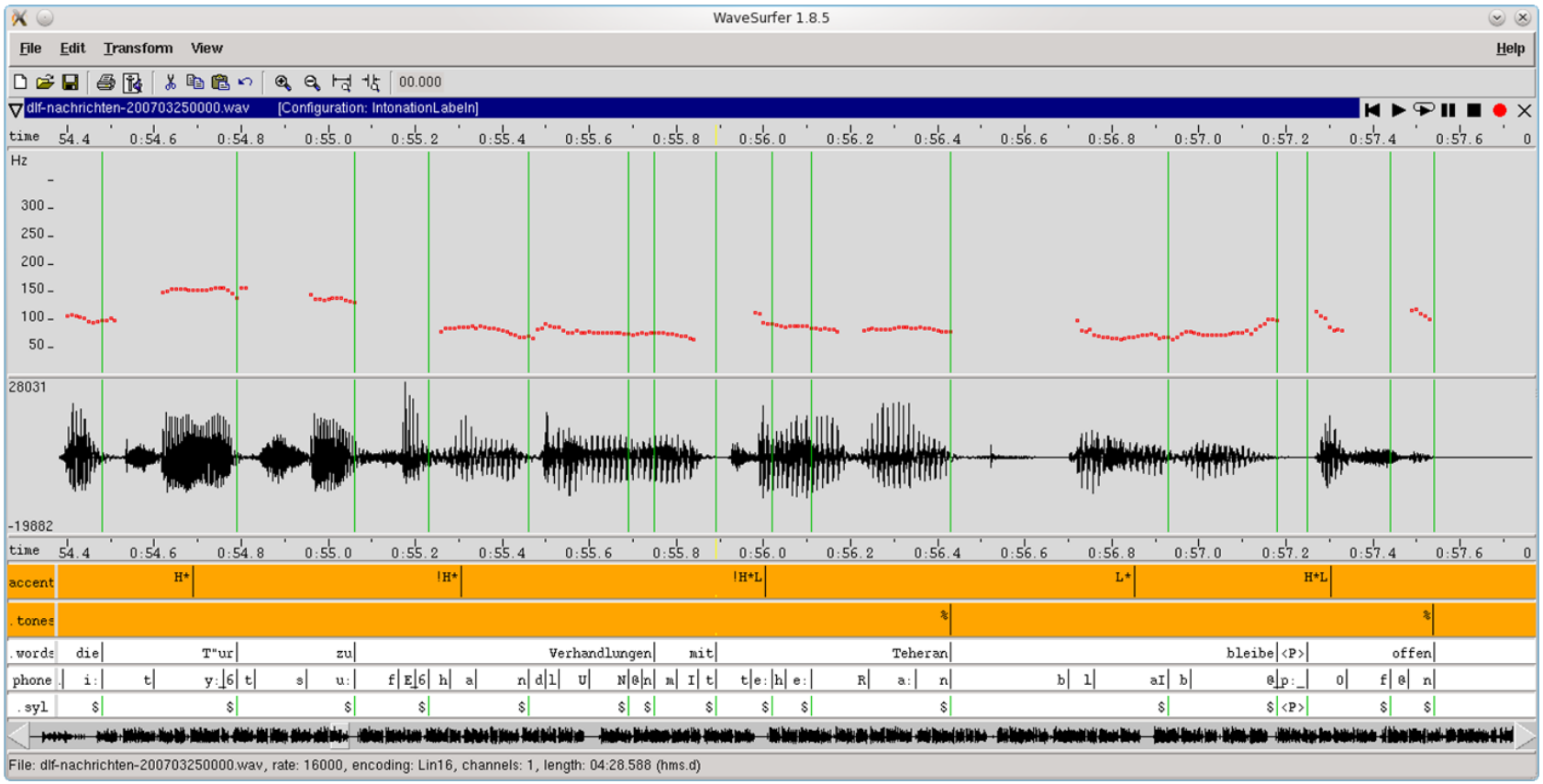
Prosody: GToBI(S) labels (Mayer, 1995)



- Preprocessing with speech analysis software:
  alignment with words, segmentation into syllables and phonemes

Annotating speech using WaveSurfer

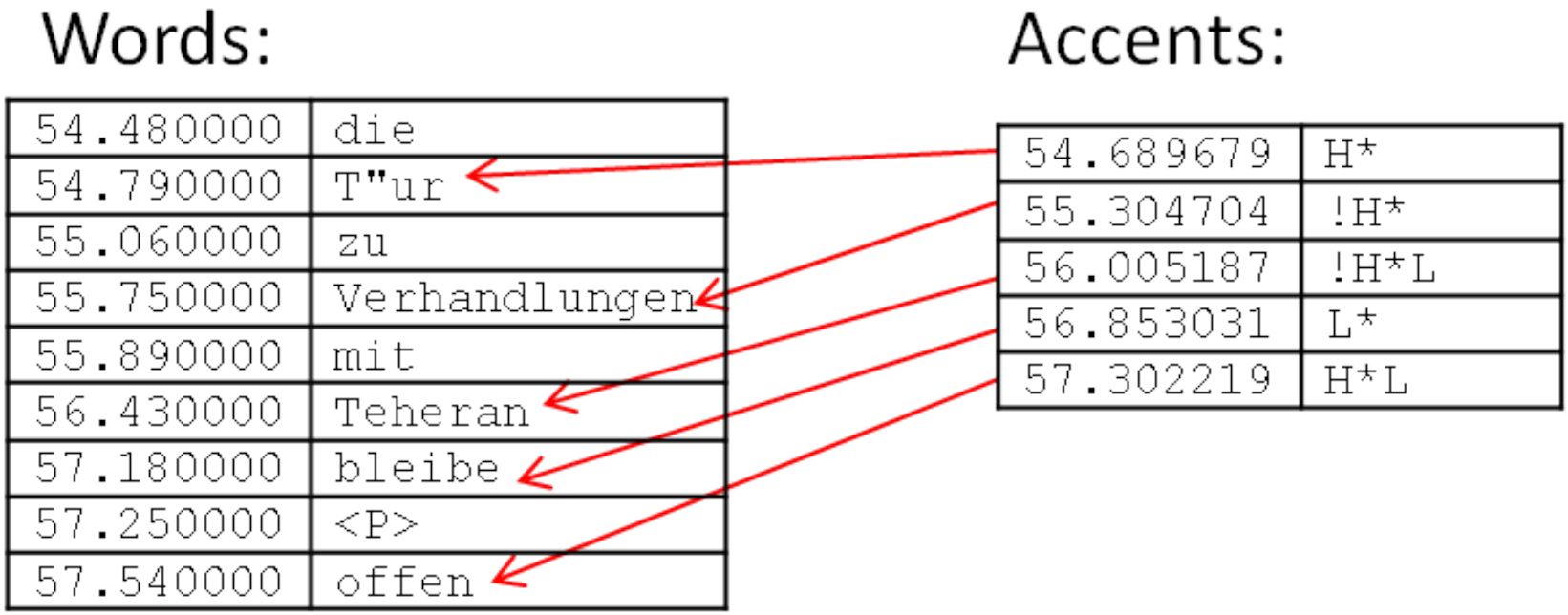
Prosody: GToBI(S) labels (Mayer, 1995)



- Preprocessing with speech analysis software:
  alignment with words, segmentation into syllables and phonemes
- Manual prosodic annotation: pitch accents, prosodic boundaries

Label files

Annotations encoded in simple tables



- Time stamps represent end of label
- Some words are unaccented
- Some words carry several pitch accents

Problems of speech/text alignment

Differences between spoken and written language

- Speech is temporally preassigned – written language isn't
- Prosodic events are usually related to (sub-)word level
- Information status can be hierarchically organized
  and is usually related to (syntactic) phrases

$\Rightarrow$ Annotating information status in a speech tool?

  - possible, if speech data is relatively simple
    and mainly contains short expressions (e.g. spontaneous speech)
  - very problematic in the case of news language,
    which contains many complex expressions

Problems of speech/text alignment

Differences between spoken text (S) and underlying manuscript (W)

(3) (S) Wie das Unternehmen in einer Pflichtme mitteilung [...] bekannt gab
(W) wie das Unternehmen in einer Pflichtmitteilung [...] bekannt gab

As the company stated in a mandatory notification [...]

(4) (S) Bundeskanzler Köhler hat das ich korrigiere Bundespräsident Köhler
Chancellor Köhler – correction, Federal President Köhler
hat das Gesetz zur Gesundheitsreform unterschrieben

signed the bill on the health care reform

(5) (S) Kanzlerin Merkel setzte aber auf eine einvernehmliche Verständigung
(W) Kanzlerin Merkel setze aber auf eine einvernehmliche Verständigung

But Chancellor Merkel relied/(allegedly) relies on a mutual agreement

⇒ Direction of change depends on task
Changes for an exact match probably restrict further explorations

Problems of speech/text alignment

More differences are added during processing

Transcription conventions

(6)

(S) mit einer Stärke von 6 Komma 9 auf der Richterskala

(W) mit einer Stärke von 6,9 auf der Richterskala

with a magnitude of 6 point 9 / 6.9 on the Richter scale

Tokenization

(7)

(S) |EU |Staaten |

(W) |EU-Staaten |

EU member states

Representing the two datasets in the database

Mapping to data structures

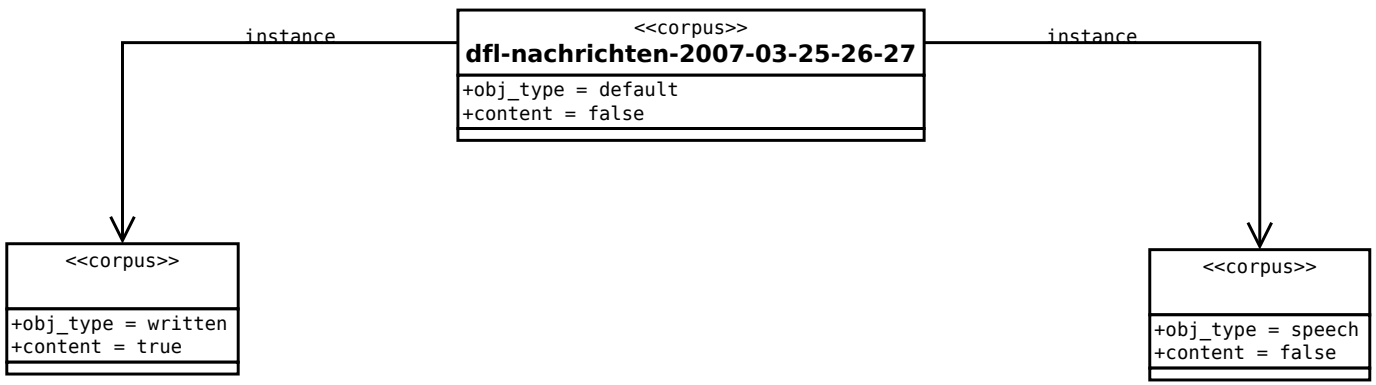
- Represent each annotation layer as graph
  $\Rightarrow$ Stand-off approach
- Choose most common level as a flexible interface
  $\Rightarrow$ Link representations at word level
- All information is annotation
  $\Rightarrow$ Timestamps are treated as annotations

Generic database approach

- Relational database management system (PostgreSQL)
- Representing different types of data objects
  accumulating in a corpus-based research project:
  primary (corpus) data, information about tools for linguistic analysis, annotations, ...
- Managing analysis results produced with different analysis processes
- Theory-independence:
  data structures
  designed without preference for a particular linguistic theory
- Interoperability
  - with existing infrastructure and formats (ANNIS/PAULA)
  - with upcoming ISO-standards (LAF/GrAF framework)

Macro layer

Model project workflow



Micro layer

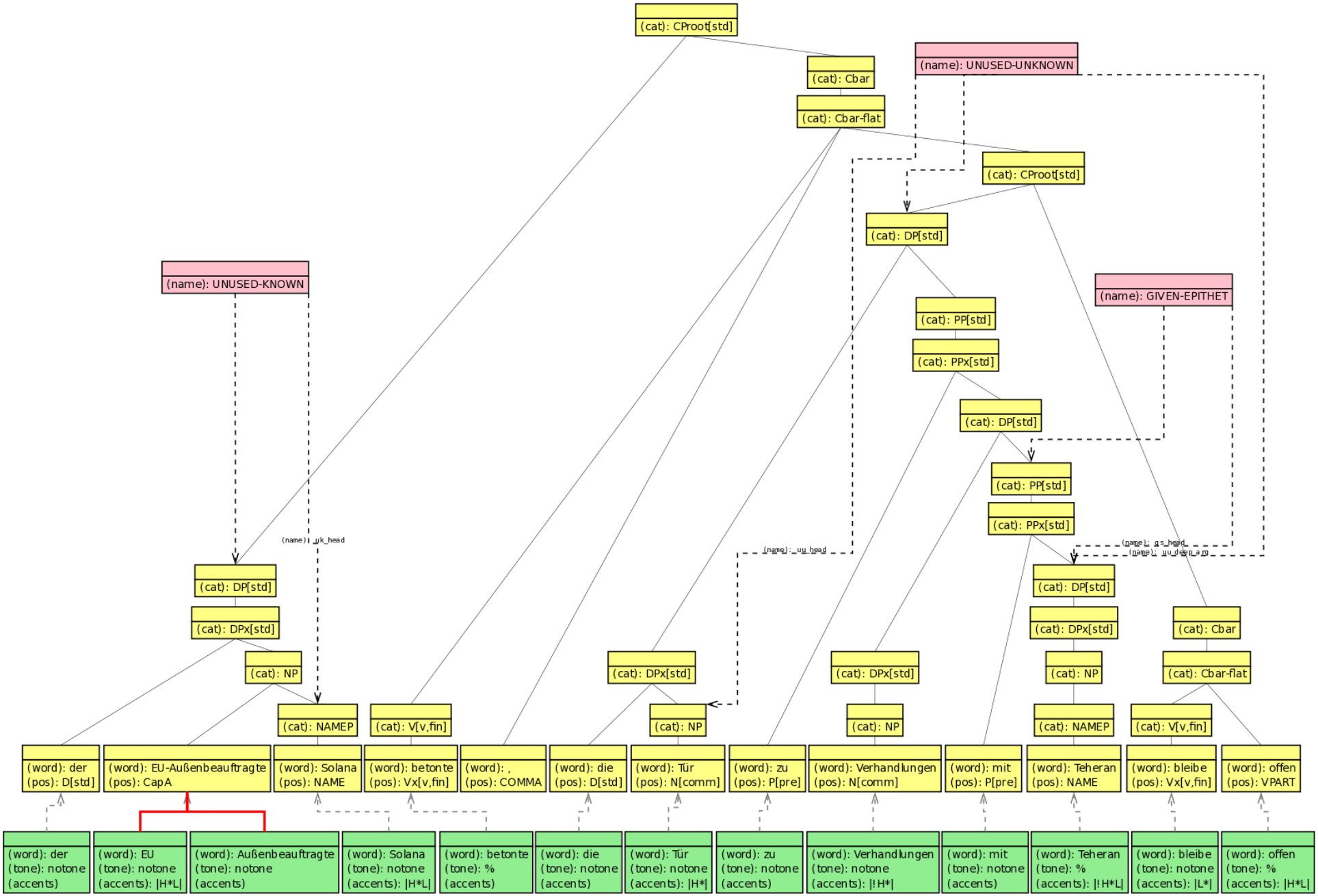
GrAF-based data structures for complex annotations

- Implementation of data structures defined in the serialization of LAF (Linguistic Annotation Framework – ISO/FDIS 24612)
- Different layers of stand-off annotations represented by
  - nodes
  - edges
  - simple annotations (labels) or
  - complex annotations (feature structures)

$\longrightarrow$ Graph-based

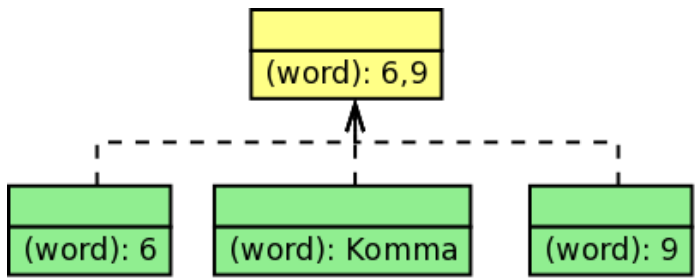
$\longrightarrow$ Theory-independent

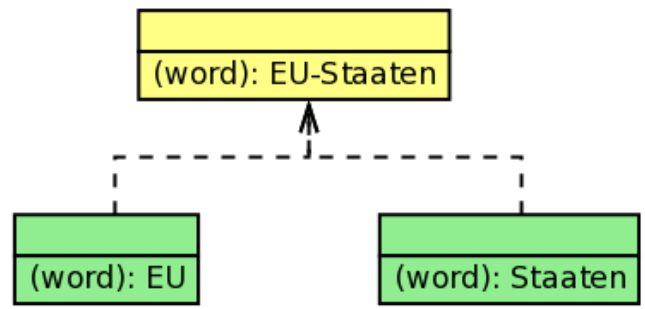
$\longrightarrow$ Exchange format – supports interoperability for data export


(cat): CProot[std]
(cat): Cbar
(name): UNUSED-UNKNOWN
(cat): Cbar-flat
(cat): CProot[std]
(cat): DP[std]
(name): UNUSED-KNOWN
(name): GIVEN-EPITHET
(cat): PP[std]
(cat): PPx[std]
(cat): DP[std]
(cat): PP[std]
(cat): PPx[std]
(name): uk_head
(name): uu_head
(name): gs_head
(name): uu_deep_arg
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): V[v,fin]
(cat): DPx[std]
(cat): NP
(cat): DPx[std]
(cat): NP
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): Cbar
(cat): Cbar-flat
(cat): V[v,fin]
(word): der (pos): D[std]
(word): EU-Außenbeauftragte (pos): CapA
(word): Solana (pos): NAME
(word): betonte (pos): Vx[v,fin]
(word): , (pos): COMMA
(word): die (pos): D[std]
(word): Tür (pos): N[comm]
(word): zu (pos): P[pre]
(word): Verhandlungen (pos): N[comm]
(word): mit (pos): P[pre]
(word): Teheran (pos): NAME
(word): bleibe (pos): Vx[v,fin]
(word): offen (pos): VPART
(word): der (tone): notone (accents)
(word): EU (tone): notone (accents): |H*L|
(word): Außenbeauftragte (tone): notone (accents)
(word): Solana (tone): notone (accents): |H*L|
(word): betonte (tone): % (accents)
(word): die (tone): notone (accents)
(word): Tür (tone): notone (accents): |H*|
(word): zu (tone): notone (accents)
(word): Verhandlungen (tone): notone (accents): |!H*|
(word): mit (tone): notone (accents)
(word): Teheran (tone): % (accents): |!H*L|
(word): bleibe (tone): notone (accents): |L*|
(word): offen (tone): % (accents): |H*L|

Linking of data and annotations

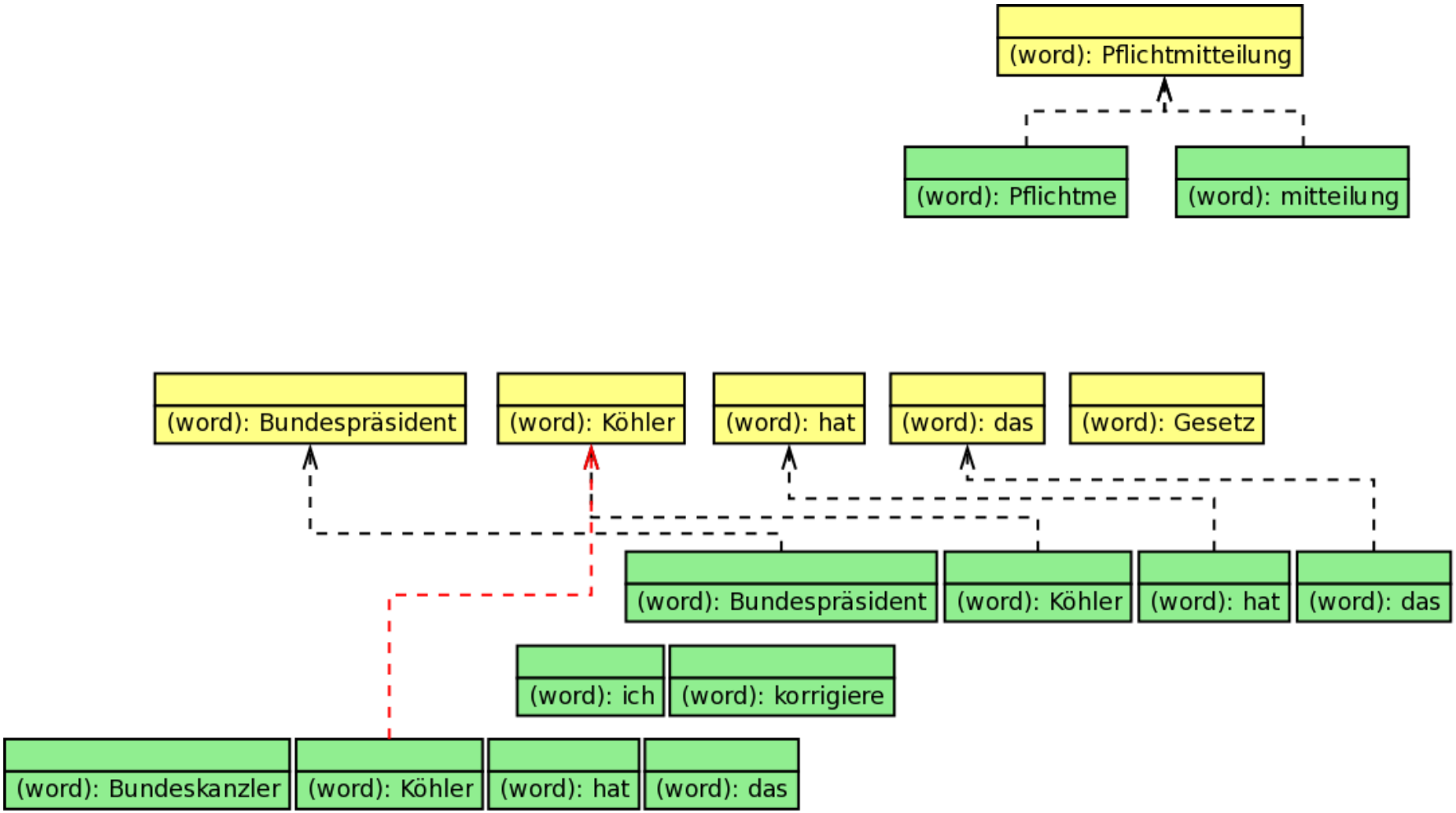
Automatic linking





Linking of data and annotations

Linking: difficult cases



Queries

Complexity trade-off

- Generic structures
  allow for a representation of different annotation layers
  and for connecting these layers in a flexible way
- Queries need to specify in detail which (type of) data to select

Queries

Example: two-word phrases

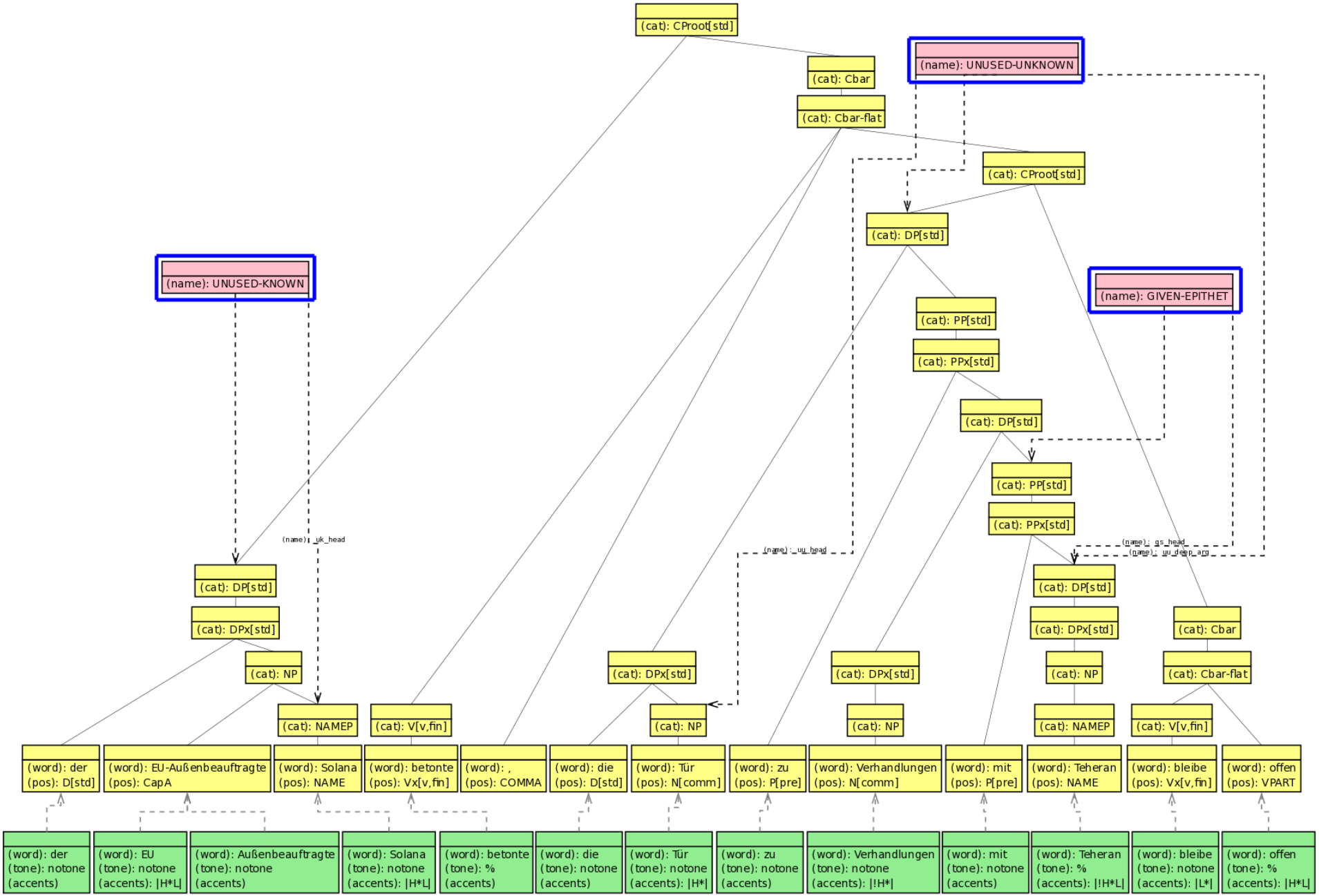
"Find all IS-labelled phrases of length 2/consisting of two words. Display the phrase, the IS-label and the prosodic realization."

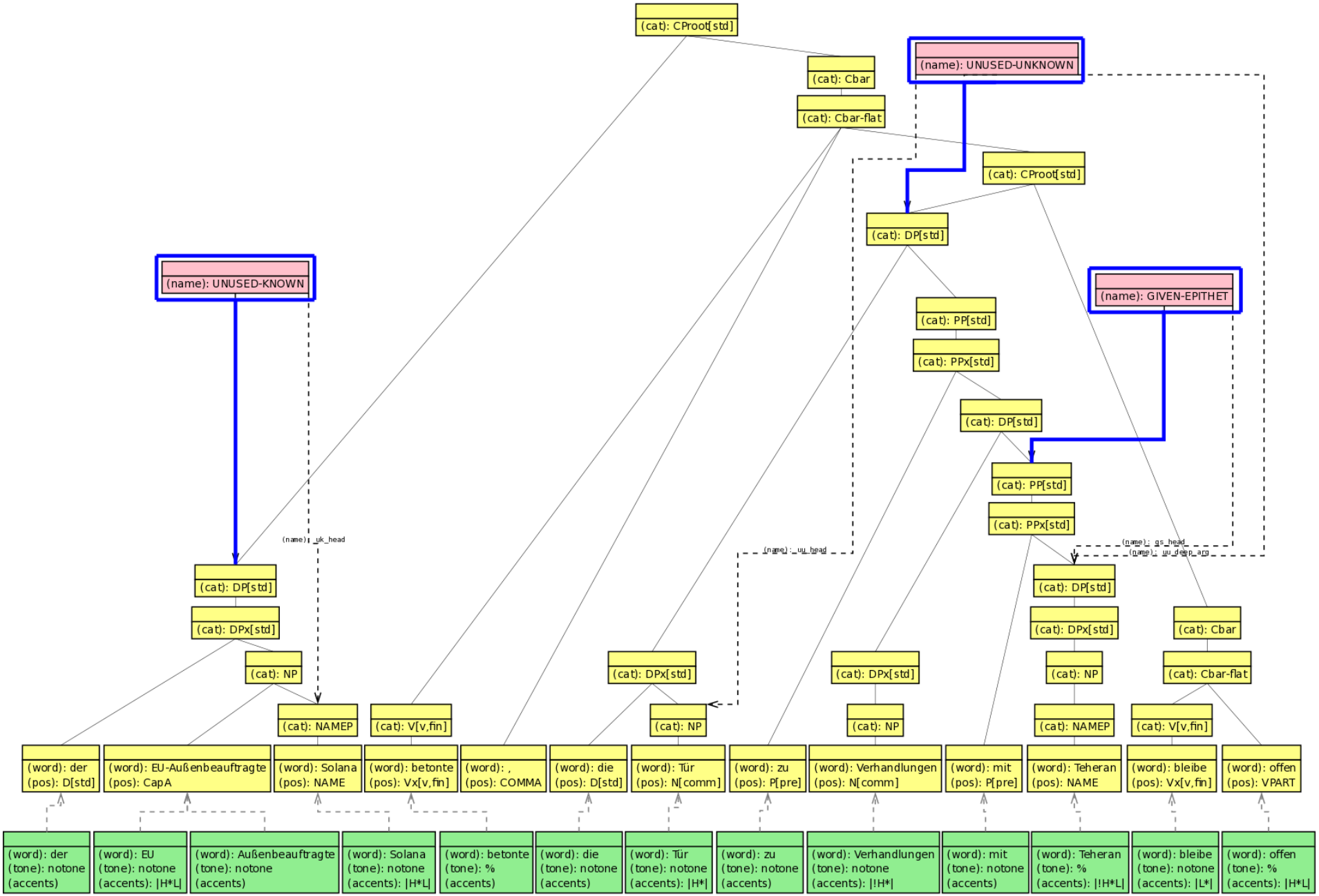
Excerpt of example query:

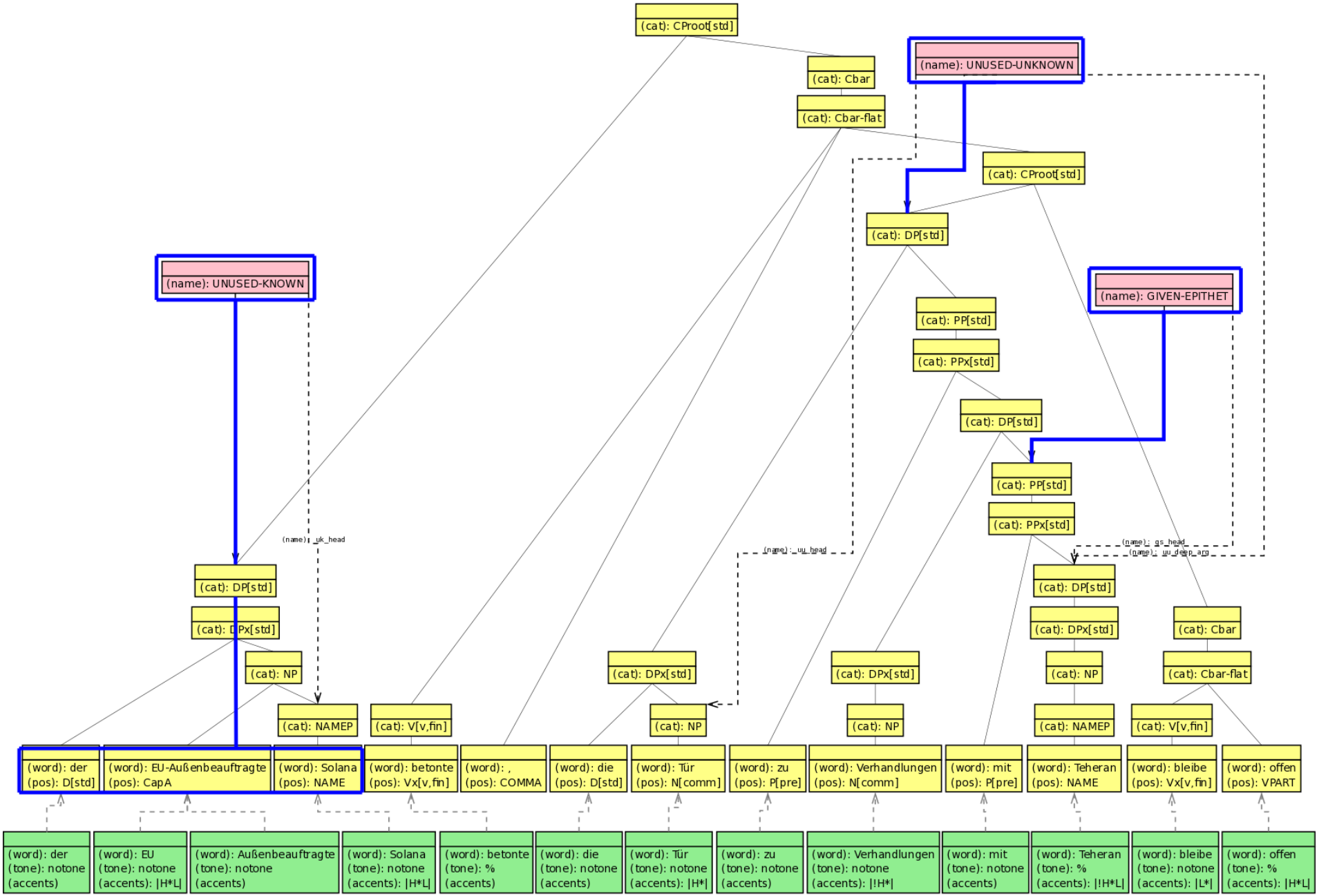
`is_syn_p`

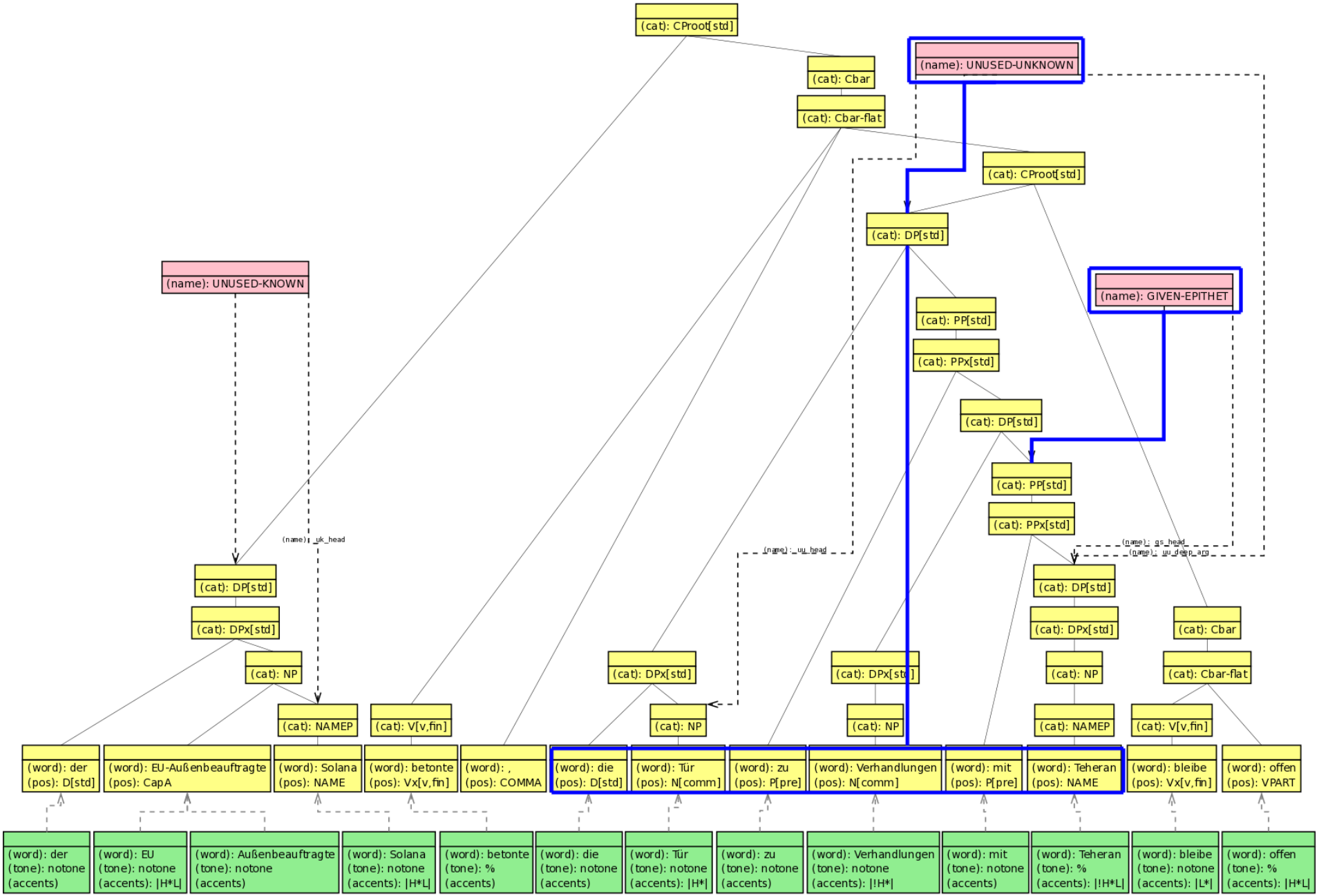
Temporary table relating each IS label with its corresponding syntactic phrase, and each of the phrases with the corresponding accent contour.

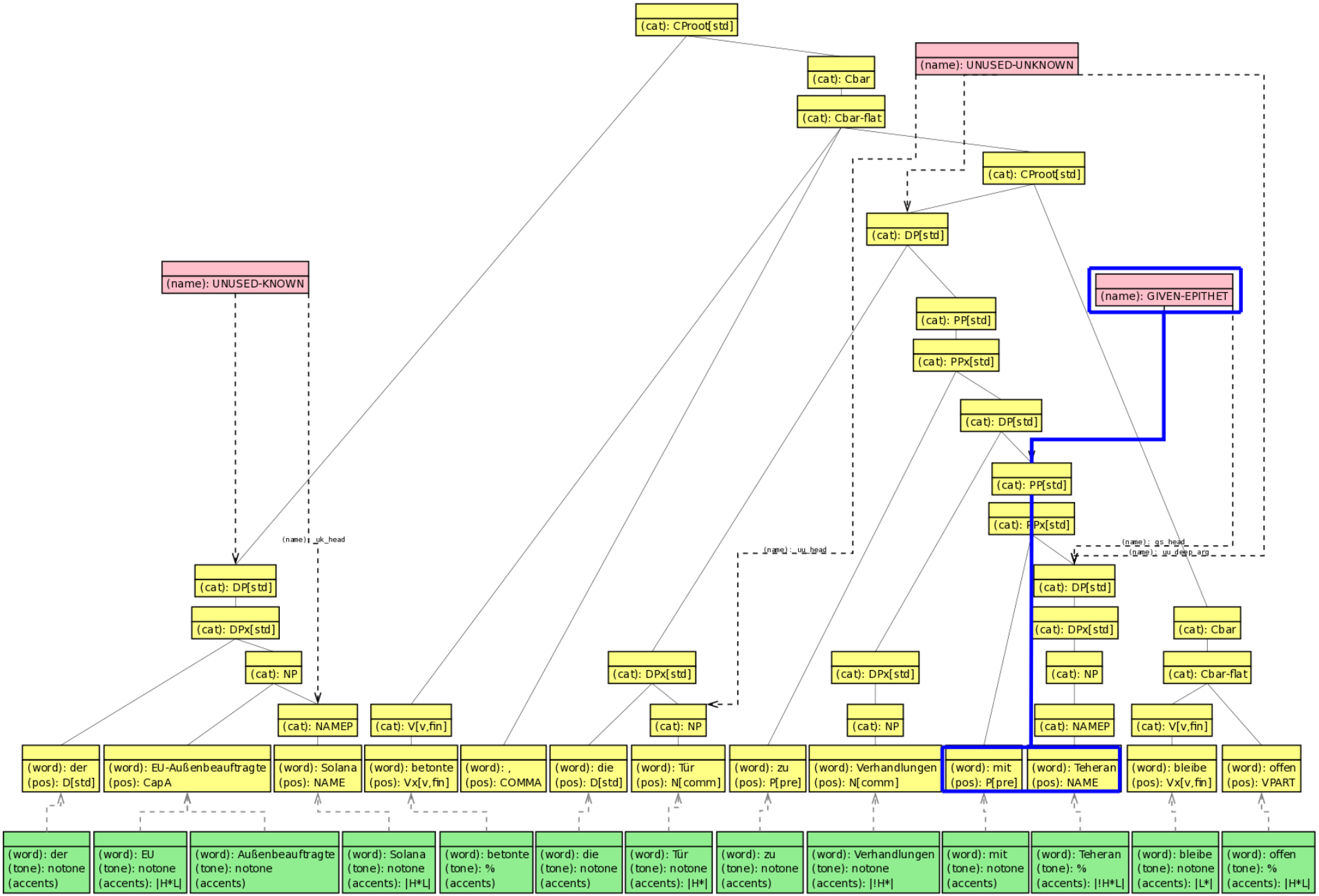
```
SELECT
    is_syn_p.syn_s_num,
    is_syn_p.is_label,
    is_syn_p.phrase,
    is_syn_p.accent_sequence
FROM
    is_syn_p,
    sentences
WHERE
    is_syn_p.syn_s_num=sentences.s_num
   AND
    is_syn_p.syn_phrase_length=2;
```

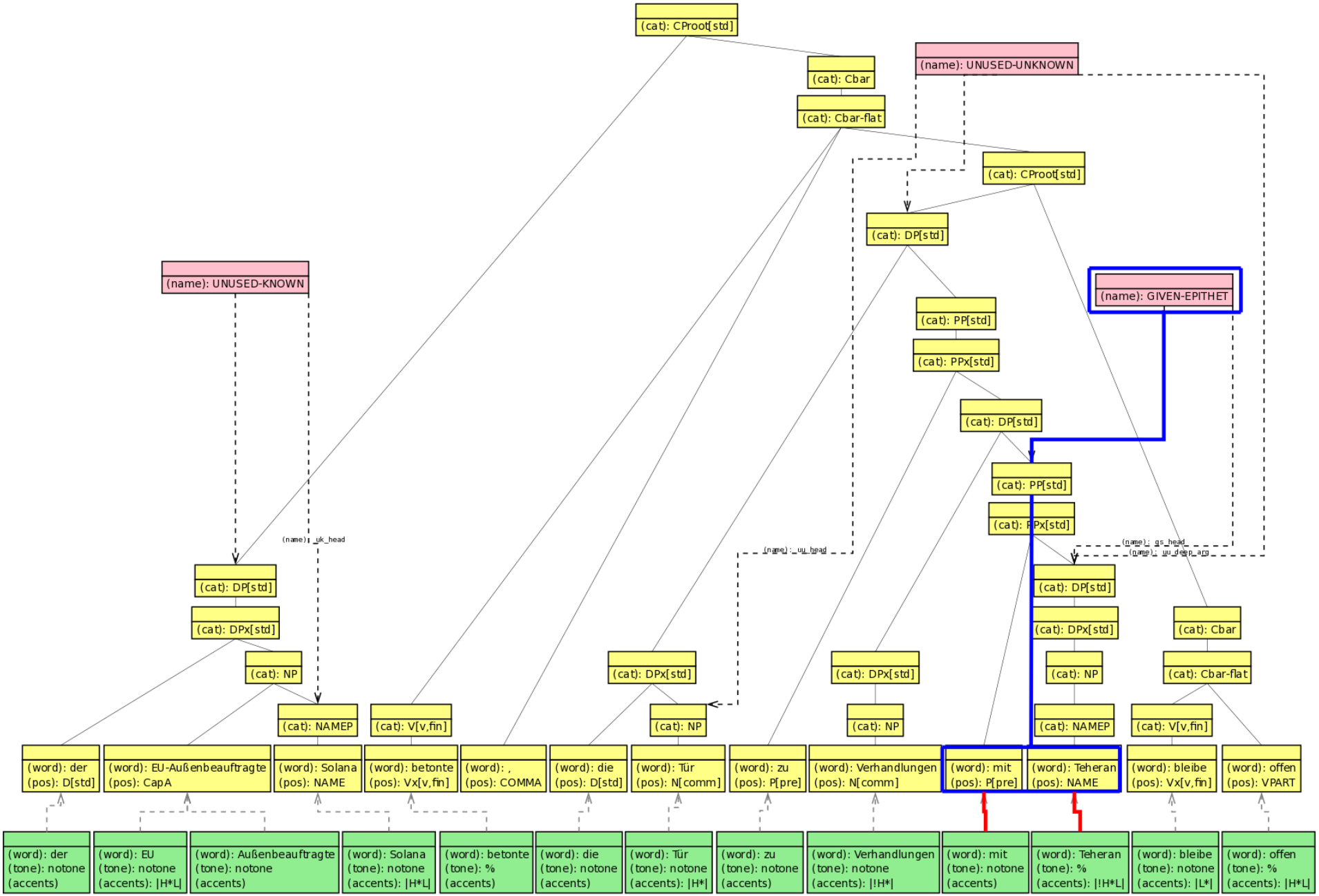

(cat): CProot[std]
(cat): Cbar
(cat): Cbar-flat
(name): UNUSED-UNKNOWN
(cat): CProot[std]
(cat): DP[std]
(name): UNUSED-KNOWN
(name): GIVEN-EPITHET
(cat): PP[std]
(cat): PPx[std]
(cat): DP[std]
(cat): PP[std]
(cat): PPx[std]
(name): uk_head
(name): uu_head
(name): gs_head
(name): uu_deep_arg
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): V[v,fin]
(cat): DPx[std]
(cat): NP
(cat): DPx[std]
(cat): NP
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): Cbar
(cat): Cbar-flat
(cat): V[v,fin]
(word): der
(pos): D[std]
(word): EU-Außenbeauftragte
(pos): CapA
(word): Solana
(pos): NAME
(word): betonte
(pos): Vx[v,fin]
(word): ,
(pos): COMMA
(word): die
(pos): D[std]
(word): Tür
(pos): N[comm]
(word): zu
(pos): P[pre]
(word): Verhandlungen
(pos): N[comm]
(word): mit
(pos): P[pre]
(word): Teheran
(pos): NAME
(word): bleibe
(pos): Vx[v,fin]
(word): offen
(pos): VPART
(word): der
(tone): notone
(accents)
(word): EU
(tone): notone
(accents): |H*L|
(word): Außenbeauftragte
(tone): notone
(accents)
(word): Solana
(tone): notone
(accents): |H*L|
(word): betonte
(tone): %
(accents)
(word): die
(tone): notone
(accents)
(word): Tür
(tone): notone
(accents): |H*|
(word): zu
(tone): notone
(accents)
(word): Verhandlungen
(tone): notone
(accents): |!H*|
(word): mit
(tone): notone
(accents)
(word): Teheran
(tone): %
(accents): |!H*L|
(word): bleibe
(tone): notone
(accents): |L*|
(word): offen
(tone): %
(accents): |H*L|

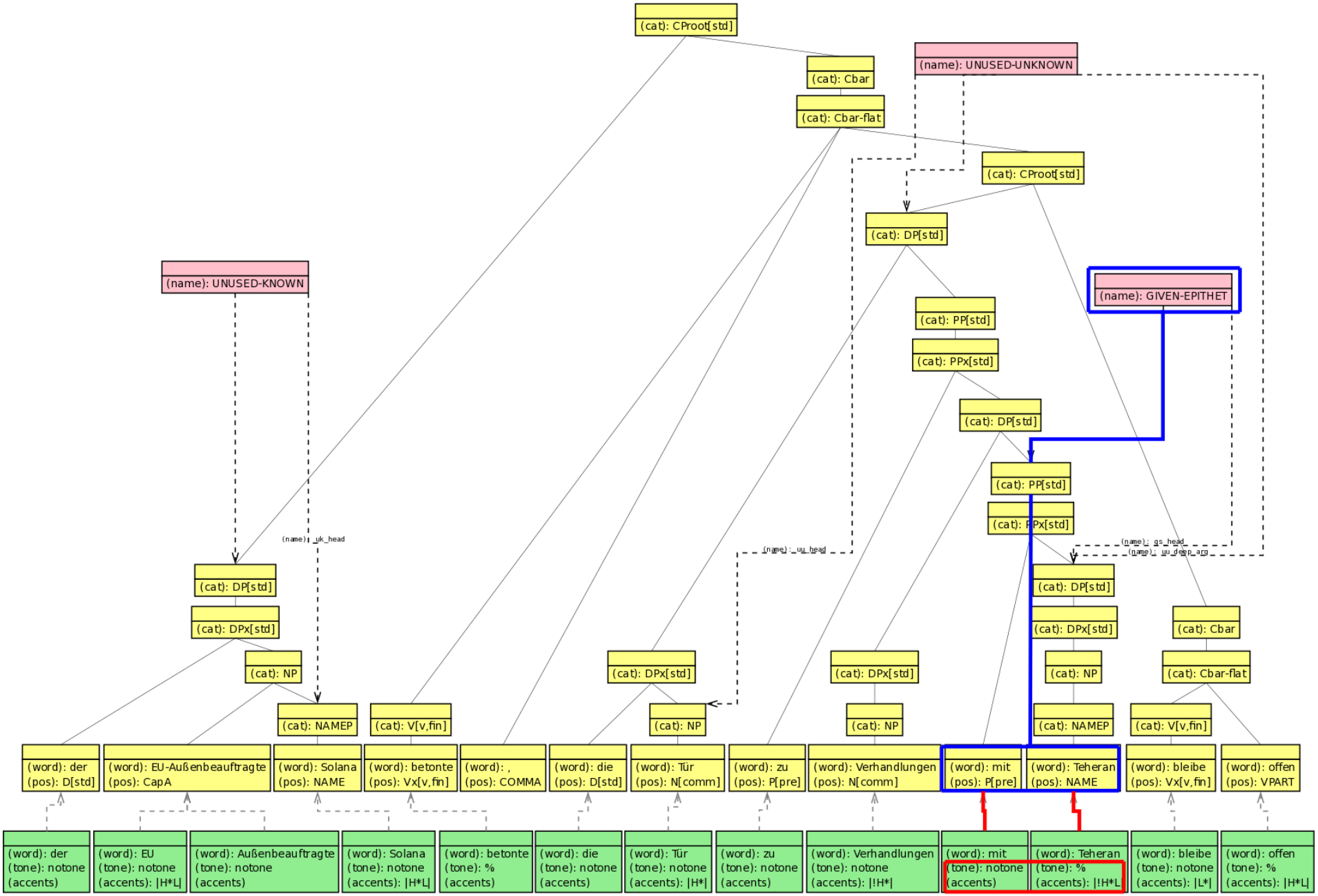

(cat): CProot[std]
(cat): Cbar
(cat): Cbar-flat
(name): UNUSED-UNKNOWN
(cat): CProot[std]
(cat): DP[std]
(name): UNUSED-KNOWN
(name): GIVEN-EPITHET
(cat): PP[std]
(cat): PPx[std]
(cat): DP[std]
(cat): PP[std]
(cat): PPx[std]
(name): uk_head
(name): uu_head
(name): gs_head
(name): uu_deep_arg
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): V[v,fin]
(cat): DPx[std]
(cat): NP
(cat): DPx[std]
(cat): NP
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): Cbar
(cat): Cbar-flat
(cat): V[v,fin]
(word): der
(pos): D[std]
(word): EU-Außenbeauftragte
(pos): CapA
(word): Solana
(pos): NAME
(word): betonte
(pos): Vx[v,fin]
(word): ,
(pos): COMMA
(word): die
(pos): D[std]
(word): Tür
(pos): N[comm]
(word): zu
(pos): P[pre]
(word): Verhandlungen
(pos): N[comm]
(word): mit
(pos): P[pre]
(word): Teheran
(pos): NAME
(word): bleibe
(pos): Vx[v,fin]
(word): offen
(pos): VPART
(word): der
(tone): notone
(accents)
(word): EU
(tone): notone
(accents): |H*L|
(word): Außenbeauftragte
(tone): notone
(accents)
(word): Solana
(tone): notone
(accents): |H*L|
(word): betonte
(tone): %
(accents)
(word): die
(tone): notone
(accents)
(word): Tür
(tone): notone
(accents): |H*|
(word): zu
(tone): notone
(accents)
(word): Verhandlungen
(tone): notone
(accents): |!H*|
(word): mit
(tone): notone
(accents)
(word): Teheran
(tone): %
(accents): |!H*L|
(word): bleibe
(tone): notone
(accents): |L*|
(word): offen
(tone): %
(accents): |H*L|


(cat): CProot[std]
(cat): Cbar
(cat): Cbar-flat
(name): UNUSED-UNKNOWN
(cat): CProot[std]
(cat): DP[std]
(name): UNUSED-KNOWN
(name): GIVEN-EPITHET
(cat): PP[std]
(cat): PPx[std]
(cat): DP[std]
(cat): PP[std]
(cat): PPx[std]
(name): uk_head
(name): uu_head
(name): gs_head
(name): uu_deep_arg
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): V[v,fin]
(cat): DPx[std]
(cat): NP
(cat): DPx[std]
(cat): NP
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): Cbar
(cat): Cbar-flat
(cat): V[v,fin]
(word): der (pos): D[std]
(word): EU-Außenbeauftragte (pos): CapA
(word): Solana (pos): NAME
(word): betonte (pos): Vx[v,fin]
(word): , (pos): COMMA
(word): die (pos): D[std]
(word): Tür (pos): N[comm]
(word): zu (pos): P[pre]
(word): Verhandlungen (pos): N[comm]
(word): mit (pos): P[pre]
(word): Teheran (pos): NAME
(word): bleibe (pos): Vx[v,fin]
(word): offen (pos): VPART
(word): der (tone): notone (accents)
(word): EU (tone): notone (accents): |H*L|
(word): Außenbeauftragte (tone): notone (accents)
(word): Solana (tone): notone (accents): |H*L|
(word): betonte (tone): % (accents)
(word): die (tone): notone (accents)
(word): Tür (tone): notone (accents): |H*|
(word): zu (tone): notone (accents)
(word): Verhandlungen (tone): notone (accents): |!H*|
(word): mit (tone): notone (accents)
(word): Teheran (tone): % (accents): |!H*L|
(word): bleibe (tone): notone (accents): |L*|
(word): offen (tone): % (accents): |H*L|

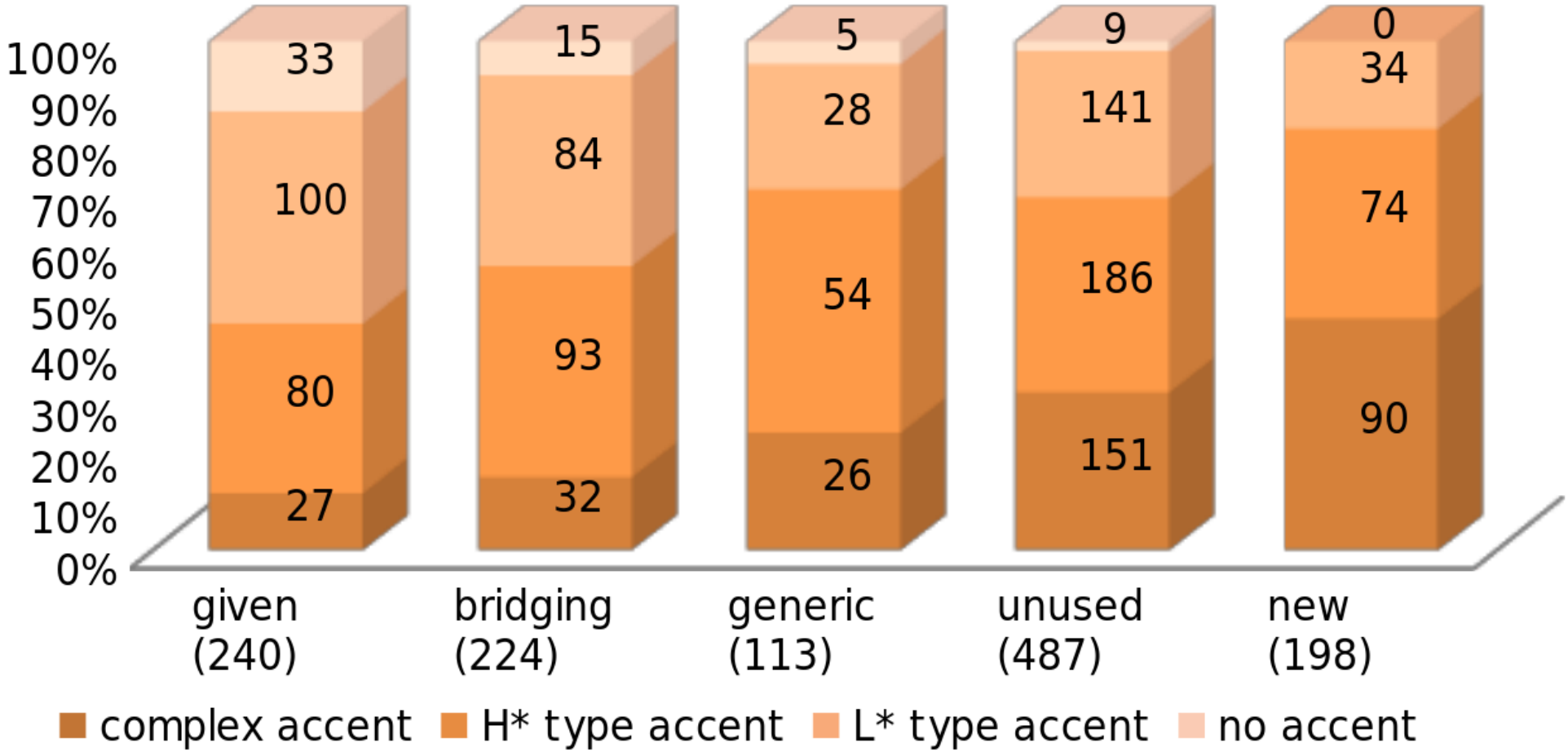

(cat): CProot[std]
(cat): Cbar
(cat): Cbar-flat
(name): UNUSED-UNKNOWN
(cat): CProot[std]
(cat): DP[std]
(name): UNUSED-KNOWN
(name): GIVEN-EPITHET
(cat): PP[std]
(cat): PPx[std]
(cat): DP[std]
(cat): PP[std]
(cat): PPx[std]
(name): uk_head
(name): uu_head
(name): gs_head
(name): uu_deep_arg
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): V[v,fin]
(cat): DPx[std]
(cat): NP
(cat): DPx[std]
(cat): NP
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): Cbar
(cat): Cbar-flat
(cat): V[v,fin]
(word): der
(pos): D[std]
(word): EU-Außenbeauftragte
(pos): CapA
(word): Solana
(pos): NAME
(word): betonte
(pos): Vx[v,fin]
(word): ,
(pos): COMMA
(word): die
(pos): D[std]
(word): Tür
(pos): N[comm]
(word): zu
(pos): P[pre]
(word): Verhandlungen
(pos): N[comm]
(word): mit
(pos): P[pre]
(word): Teheran
(pos): NAME
(word): bleibe
(pos): Vx[v,fin]
(word): offen
(pos): VPART
(word): der
(tone): notone
(accents)
(word): EU
(tone): notone
(accents): |H*L|
(word): Außenbeauftragte
(tone): notone
(accents)
(word): Solana
(tone): notone
(accents): |H*L|
(word): betonte
(tone): %
(accents)
(word): die
(tone): notone
(accents)
(word): Tür
(tone): notone
(accents): |H*|
(word): zu
(tone): notone
(accents)
(word): Verhandlungen
(tone): notone
(accents): |!H*|
(word): mit
(tone): notone
(accents)
(word): Teheran
(tone): %
(accents): |!H*L|
(word): bleibe
(tone): notone
(accents): |L*|
(word): offen
(tone): %
(accents): |H*L|


(cat): CProot[std]
(cat): Cbar
(cat): Cbar-flat
(name): UNUSED-UNKNOWN
(cat): CProot[std]
(cat): DP[std]
(name): UNUSED-KNOWN
(name): GIVEN-EPITHET
(cat): PP[std]
(cat): PPx[std]
(cat): DP[std]
(cat): PP[std]
(cat): PPx[std]
(name): uk_head
(name): uu_head
(name): gs_head
(name): uu_deep_arg
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): V[v,fin]
(cat): DPx[std]
(cat): NP
(cat): DPx[std]
(cat): NP
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): Cbar
(cat): Cbar-flat
(cat): V[v,fin]
(word): der
(pos): D[std]
(word): EU-Außenbeauftragte
(pos): CapA
(word): Solana
(pos): NAME
(word): betonte
(pos): Vx[v,fin]
(word): ,
(pos): COMMA
(word): die
(pos): D[std]
(word): Tür
(pos): N[comm]
(word): zu
(pos): P[pre]
(word): Verhandlungen
(pos): N[comm]
(word): mit
(pos): P[pre]
(word): Teheran
(pos): NAME
(word): bleibe
(pos): Vx[v,fin]
(word): offen
(pos): VPART
(word): der
(tone): notone
(accents)
(word): EU
(tone): notone
(accents): |H*L|
(word): Außenbeauftragte
(tone): notone
(accents)
(word): Solana
(tone): notone
(accents): |H*L|
(word): betonte
(tone): %
(accents)
(word): die
(tone): notone
(accents)
(word): Tür
(tone): notone
(accents): |H*|
(word): zu
(tone): notone
(accents)
(word): Verhandlungen
(tone): notone
(accents): |!H*|
(word): mit
(tone): notone
(accents)
(word): Teheran
(tone): %
(accents): |!H*L|
(word): bleibe
(tone): notone
(accents): |L*|
(word): offen
(tone): %
(accents): |H*L|


(cat): CProot[std]
(cat): Cbar
(cat): Cbar-flat
(name): UNUSED-UNKNOWN
(cat): CProot[std]
(cat): DP[std]
(name): UNUSED-KNOWN
(name): GIVEN-EPITHET
(cat): PP[std]
(cat): PPx[std]
(cat): DP[std]
(cat): PP[std]
(cat): PPx[std]
(name): uk_head
(name): uu_head
(name): gs_head
(name): uu_deep_arg
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): V[v,fin]
(cat): DPx[std]
(cat): NP
(cat): DPx[std]
(cat): NP
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): Cbar
(cat): Cbar-flat
(cat): V[v,fin]
(word): der
(pos): D[std]
(word): EU-Außenbeauftragte
(pos): CapA
(word): Solana
(pos): NAME
(word): betonte
(pos): Vx[v,fin]
(word): ,
(pos): COMMA
(word): die
(pos): D[std]
(word): Tür
(pos): N[comm]
(word): zu
(pos): P[pre]
(word): Verhandlungen
(pos): N[comm]
(word): mit
(pos): P[pre]
(word): Teheran
(pos): NAME
(word): bleibe
(pos): Vx[v,fin]
(word): offen
(pos): VPART
(word): der
(tone): notone
(accents)
(word): EU
(tone): notone
(accents): |H*L|
(word): Außenbeauftragte
(tone): notone
(accents)
(word): Solana
(tone): notone
(accents): |H*L|
(word): betonte
(tone): %
(accents)
(word): die
(tone): notone
(accents)
(word): Tür
(tone): notone
(accents): |H*|
(word): zu
(tone): notone
(accents)
(word): Verhandlungen
(tone): notone
(accents): |!H*|
(word): mit
(tone): notone
(accents)
(word): Teheran
(tone): %
(accents): |!H*L|
(word): bleibe
(tone): notone
(accents): |L*|
(word): offen
(tone): %
(accents): |H*L|


(cat): CProot[std]
(cat): Cbar
(cat): Cbar-flat
(name): UNUSED-UNKNOWN
(cat): CProot[std]
(cat): DP[std]
(name): UNUSED-KNOWN
(name): GIVEN-EPITHET
(cat): PP[std]
(cat): PPx[std]
(cat): DP[std]
(cat): PP[std]
(cat): PPx[std]
(name): uk_head
(name): uu_head
(name): gs_head
(name): uu_deep_arg
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): V[v,fin]
(cat): DPx[std]
(cat): NP
(cat): DPx[std]
(cat): NP
(cat): DP[std]
(cat): DPx[std]
(cat): NP
(cat): NAMEP
(cat): Cbar
(cat): Cbar-flat
(cat): V[v,fin]
(word): der (pos): D[std]
(word): EU-Außenbeauftragte (pos): CapA
(word): Solana (pos): NAME
(word): betonte (pos): Vx[v,fin]
(word): , (pos): COMMA
(word): die (pos): D[std]
(word): Tür (pos): N[comm]
(word): zu (pos): P[pre]
(word): Verhandlungen (pos): N[comm]
(word): mit (pos): P[pre]
(word): Teheran (pos): NAME
(word): bleibe (pos): Vx[v,fin]
(word): offen (pos): VPART
(word): der (tone): notone (accents)
(word): EU (tone): notone (accents): |H*L|
(word): Außenbeauftragte (tone): notone (accents)
(word): Solana (tone): notone (accents): |H*L|
(word): betonte (tone): % (accents)
(word): die (tone): notone (accents)
(word): Tür (tone): notone (accents): |H*|
(word): zu (tone): notone (accents)
(word): Verhandlungen (tone): notone (accents): |!H*|
(word): mit (tone): notone (accents)
(word): Teheran (tone): % (accents): |!H*L|
(word): bleibe (tone): notone (accents): |L*|
(word): offen (tone): % (accents): |H*L|

Results – pitch accents on two-word terms



Conclusions and outlook

The introduced structure...

(+) is able to handle different types of data,

(+) allows to keep track of changes in project workflows,

(+) allows for flexible linking of different annotation layers or annotated data sets

(−) requires detailed knowledge about the mapping of the annotations to the data structures,

(−) requires users to formulate (complex) SQL queries and deduce views

→ Export in GrAF format for publication via CLARIN-D platforms

→ Corrections to syntactic structure, linking of new versions

References

- Stefan Baumann and Arndt Riester. Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects. In G. Elordieta and P. Prieto, editors, Prosody and Meaning, Interface Explorations. De Gruyter Mouton, Berlin. To appear.
- Stefanie Dipper. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In Proceedings of Berliner XML Tage 2005, 39–50. Berlin. 2005.
- Kerstin Eckart, Kurt Eberle and Ulrich Heid. An Infrastructure for More Reliable Corpus Analysis. In Proceedings of the Workshop on Web Services and Processing Pipelines in HLT (LREC), pages 8–14, Valletta, Malta. 2010.
- Nancy Ide and Keith Suderman. GrAF: A Graph-based Format for Linguistic Annotations. Proceedings of the Linguistic Annotation Workshop, held in conjunction with ACL 2007, Prague, 1–8. 2007.
- Jörg Mayer. Transcription of German Intonation. The Stuttgart System. 1995.
- Arndt Riester, David Lorenz and Nina Seemann. A Recursive Annotation Scheme for Referential Information Status. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), pages 717–722, Valletta, Malta. 2010.

ANNIS/PAULA: developed at the Collaborative Research Centre 632, `http://www.sfb632.uni-potsdam.de/~d1/annis/`

Clarin-D: `http://de.clarin.eu/`

LAF/GrAF: ISO/FDIS 24612 Language resource management - Linguistic annotation framework (LAF).

PostgreSQL: `http://www.postgresql.org/`