# A Discourse Information Radio News Database for Linguistic Analysis

Kerstin Eckart, Arndt Riester, and Katrin Schweitzer

**Abstract** In this paper we present DIRNDL, an annotated corpus resource comprising syntactic annotations as well as information status labels and prosodic information. We introduce each annotation layer and then focus on the linking of the data in a standoff approach. The corpus is based on data from radio news broadcasts, i.e. two sets of primary data: spoken radio news files and a written text version which sometimes deviates from the actual spoken data. We utilize a generic relational database management system to bridge the gap between the deviating primary data as well as between the different properties of the annotation levels. We show how the resource can support data extraction concerning the interface between information status, syntax and prosody.

## 1 Introduction

We present the DIRNDL corpus (Discourse Information Radio News Database for Linguistic analysis), an annotated resource of news broadcasts from Deutschlandfunk, a German radio station, prepared for the investigation of the interfaces between prosody, information status and syntax.[1] The database contains audio files (approx. 5 hours of speech; 9 speakers: 5m, 4f), which were annotated for pitch accents and prosodic boundaries following GToBI(S) (Mayer, 1995). Furthermore, it comprises a treebank based on the written manuscripts of the news (3221 sentences), which were annotated for referential information status (given-new distinction), according to Riester et al (2010). The two types of data are aligned in a generic relational database management system described in Eckart et al (2010).

Kerstin Eckart · Arndt Riester · Katrin Schweitzer
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, Azenbergstr. 12, 70174 Stuttgart e-mail: \{eckartkn,arndt.riester,katrin-schweitzer\}@ims.uni-stuttgart.de

[1] News broadcasts from 25-27/03/2007; downloaded from http://www.dradio.de.

## 2 Two Annotation Pipelines

There exist two primary data sets: spoken data and a slightly deviating written version. The annotation layers are the results of two different processing pipelines: one from the written primary data to recursive information status labels, and the other from the spoken primary data to prosodic annotations.

### *2.1 Workflow Towards Information Status Annotations*

The written manuscripts of the news were parsed with the XLE system and the German LFG-grammar by Rohrer and Forst (2006). The resulting constituent trees[2] were converted into TIGER-XML using TIGERRegistry (Lezius et al, 2002). A sample is shown in Fig. 1.

Information status was annotated to syntactic nodes. We used the SALTO/SALSA tool (Burchardt et al, 2006) which allows for a free definition of annotation labels (in our case, information status labels), and which takes TIGER-XML as input, see Fig. 2. Information status (Prince, 1981, 1992) describes the degree of givenness of (referential) expressions. On a slightly different interpretation, it classifies terms as to whether they are anaphoric, inferable, deictic, or discourse-new. Notions closely related to information status are *salience*, *accessibility* and *cognitive status*.

Information status forms a subfield of information structure theory, since it is usually confined to referential expressions and furthermore leaves aside aspects of contrastive focus. For the annotation of the DIRNDL corpus, we made use of the scheme defined in Riester et al (2010), which is particularly suited to handle multiple embeddings, which are very frequent in news text, see Fig. 2. The scheme has been shown to reach an inter-annotator agreement of $\kappa = .66$ for the full scheme of 21 categories and $\kappa = .78$ for a core scheme of 6 main categories.

### *2.2 Workflow Towards Prosodic Annotations*

The spoken primary data set was automatically segmented into words, syllables and phonemes using forced alignment (Rapp, 1995). Pitch accents and prosodic boundaries were manually labelled according to GToBI(S) (Mayer, 1995). Word level annotations were mapped to the syllable-based prosodic labels using Festival (Taylor et al, 1998).

Fig. 3 shows the representation of time-aligned word boundaries, combined with phrase boundaries and pitch accents, all included as annotations in the corpus. While some words can be unaccented (e.g. the determiners and prepositions in Fig. 3),

[2] We always used the parses with the highest rank.

```
<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>
```

**Fig. 1** Sample phrase in TIGER-XML format

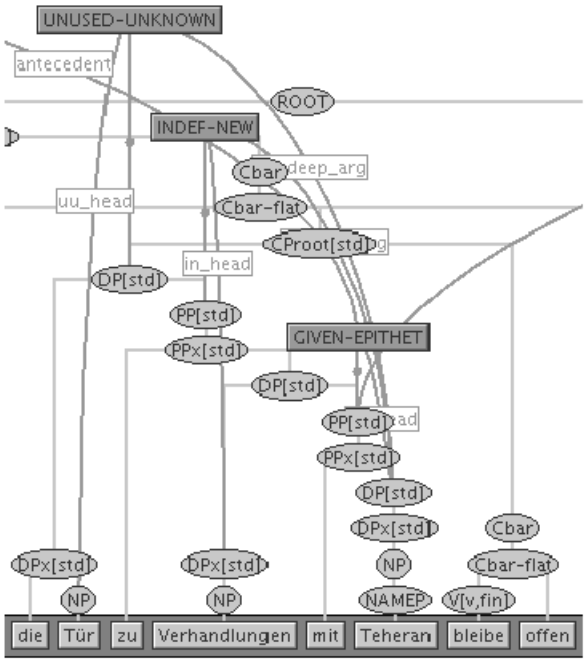


**Fig. 2** Annotation of the phrase from Fig. 1 in SALTO/SALSA

others, especially compounds, may carry more than one pitch accent. Such cases are represented as complex accents in the resource.

**Fig. 3** Sample phrase with prosodic annotations (time stamps denoting word boundaries, words, phrase boundaries, pitch accents)

| | | | |
|---|---|---|---|
| 54.480000 | die | NONE | NONE |
| 54.790000 | T"ur | NONE | \|H*\| |
| 55.060000 | zu | NONE | NONE |
| 55.750000 | Verhandlungen | NONE | \|!H*\| |
| 55.890000 | mit | NONE | NONE |
| 56.430000 | Teheran | % | \|!H*L\| |
| 57.180000 | bleibe | NONE | \|L*\| |
| 57.540000 | offen | % | \|H*L\| |

## *2.3 Differences in Annotation Structure*

There are two major differences between spoken and written language which have an influence on annotation decisions. First, speech has a temporal dimension. Every word token and every tonal event occurs at a specific time point or interval. Written language obviously lacks this temporal determination since it can be read at varying speed. A related issue, which, for lack of space, we cannot discuss in detail, is the fact that written language is often underspecified as regards its pragmatic impact. We want to mention, however, that the DIRNDL corpus is a good resource for studying meaning specification via prosody, since it contains many instances of repetitions of identical news features showing small prosodic deviations.

Second, as we pointed out in Sect. 2.1, to systematically annotate information status within complex news language, an (automatic) analysis of syntactic structure is indispensable in order to highlight hierarchical relations. As referential expressions are often embedded inside each other, so are information status labels. This cannot be adequately represented within a linearly organised phonetic analysis tool.

## *2.4 Deviations Within Primary Data*

When primary data is processed in different annotation pipelines, conflicting tokenizations may arise, which afterwards must be merged, cf. Chiarcos et al (2009). In our case, the two primary data sets, i.e. the written and the spoken one, already slightly deviate from each other due to slips of the tongue, see example (1), or other modifications. This requires additional handling.

(1) **Bundeskanzler** Köhler hat das **ich korrigiere** Bundespräsident Köhler hat das Gesetz zur Gesundheitsreform unterschrieben
'(Chancellor Köhler, correction) Federal President Köhler signed the bill on the health care reform'

As stated above, the processing of the data in different pipelines introduces even more deviations. Tokens in the prosodic pipeline refer to actually pronounced items. This leads to an inhomogeneous treatment of punctuation symbols. Hyphens, like in *EU-Außenbeauftragter* ('EU High Representative') are not pronounced and disappear in the transcriptions of the speech data, while the comma symbol in a numeric expression like *6,9* becomes a token of its own and is transcribed as the word *Komma*.

Choosing only one of the primary data sets means information loss in processing. On the one hand, slips of the tongue create problems for the parser. On the other hand, they have an influence on prosody. It is therefore not advisable to cut out parts of the speech data. To handle the differences between the primary data sets and the differences between the outputs of the processing pipelines we introduce links between the tokens created by each pipeline. That way, we are able to keep as much information as possible in the corpus and are even able to extract data for the study of specific phenomena such as the prosody of slips of the tongue.

## 3 A Generic Database Management System

Our database[3] is able to handle different data sets like primary data, metadata and linguistic annotations, cf. Eckart et al (2010). It meets the requirements for a resource like DIRNDL, as it is *extensible*, *theory-independent* and supports the versioning of annotations within a processing pipeline. Extensibility is important, as it allows to include more data sets into our resource at a later point. This is easily achieved since the generic data structures of the database allow the inclusion of new kinds of data without changes to the schema.

The database is conceptually divided into two different layers. At the *macroscopic* level each data set is represented as an object. Metadata about these objects are provided by sorting each object into a group (e.g. *corpus* for a set of primary data, or *analysis* for the result string produced by an analysis tool) and assigning it a type (e.g. *speech* for an object of group *corpus*). Versioning information is included in the form of a creation date. Other optional attribute-value-pairs can be used to add metadata, like author information etc.

Objects which contain further internal structure, such as a parse tree represented as a bracketed string, can be represented as graphs at the *microscopic* level. The data structures on the microscopic layer are mainly typed nodes and edges. The

[3] Implemented as PostgreSQL relational database system. `http://www.postgresql.org`

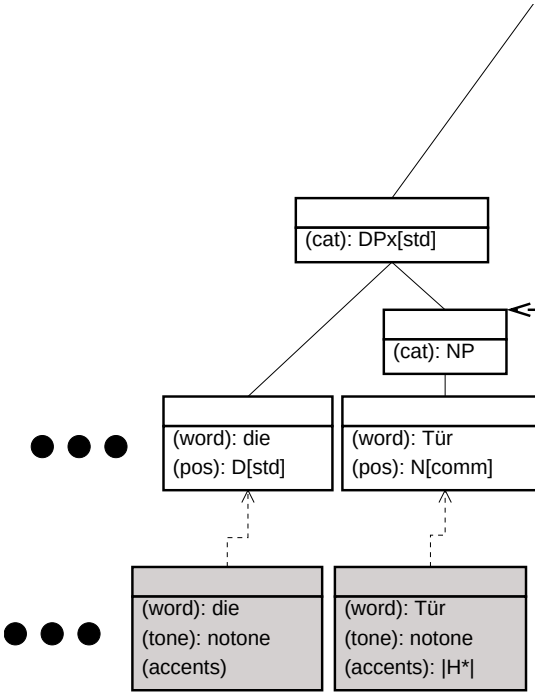


**Fig. 4** Linked annotation graphs in the database for the sentence: *Der EU-Außenbeauftragte Solana betonte, die Tür zu Verhandlungen mit Teheran bleibe offen.* ('The EU High Representative Solana stressed that the door for negotiations with Teheran remained open.')

schema is enhanced with structures based on the Graph Annotation Format (GrAF),[4] providing feature structures to annotate nodes and edges. For DIRNDL, we make use of GrAF-based data structures for all annotation layers.

## 4 Linking Annotations

With respect to each of the pipelines, the GrAF-based data-structures provide a standoff approach to the representation of each annotation layer in the database. The prosodic annotations are based on the spoken and time-aligned primary data set, the syntactic analyses from XLE are based on the written primary data set and the labels refer to the nodes in the constituent trees. Each layer is interpreted as a graph of a different type in the database:

[4] GrAF ist the XML serialization of the upcoming ISO-Standard LAF (Linguistic Annotation Framework, ISO/DIS 24612). LAF proposes a theory-independent exchange format based on a standoff approach.

- Each constituent tree (a sentence) is a graph; see the white nodes and the edges marked by continuous lines in Fig. 4.
- An information status graph contains all that refer to the same constituent tree; see the dark grey nodes in Fig. 4.
- A prosody graph comprises a complete broadcast rather than a single sentence; compare the light grey nodes in Fig. 4.

While the syntactic graphs include nodes and edges, prosody and information status are represented by unconnected graphs. They only consist of nodes. The prosody nodes become sequential when annotated with time-stamps while the information status graphs represent hierarchical information.

To prevent information loss, all information available in the results of the annotation pipelines is kept in the database in the form of annotations. This does not only comprise linguistic information like part-of-speech tags, but also the administrative information of the original SALSA-XML files (e.g. identifiers). The relations connecting the information status labels with their respective constituent trees are explicitly included in the SALSA output file. They are represented in the database as link edges between their respective information status and syntactic graphs; see the dashed edges in Fig. 4.

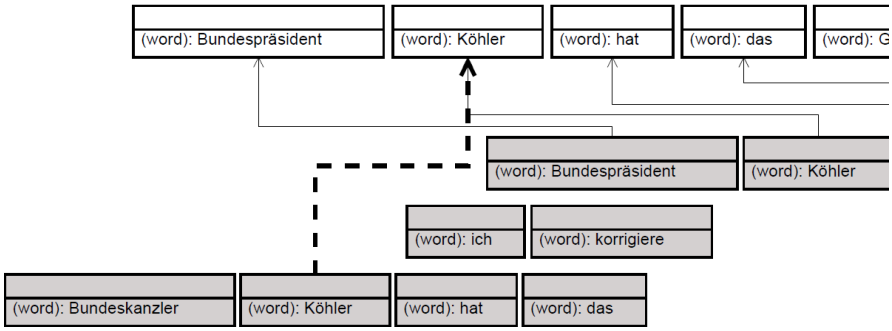


**Fig. 5** Links for example (1); tokens from written version (white), tokens from spoken version (grey), primary links, secondary link (dashed).

As a last step, we integrate the annotations of the two pipelines, by utilizing a semi-automatic mapping at token level, i.e. between the terminal nodes of the syntactic and the prosodic graphs. The algorithm takes a file with the terminal nodes from each data set as input and reads the first node from both files; if the tokens are identical or can be systematically mapped, like in the case of punctuation symbols (|*6,9*| vs. |*6*|*Komma*|*9*|; |*EU-Außenbeauftragter*| vs. |*EU*|*Außenbeauftragter*|), a link between the nodes can be inserted into the database. If the algorithm fails to map the tokens[5] the algorithm stops and prints out the tokens to the user. Then the user excludes problematic tokens from the input files and starts the mapping script

[5] The procedure is rather restrictive here to avoid mapping mismatches.

again. The user may now decide where to manually insert additional links. This is often the case with slips of the tongue, like in example (1). By also assigning types to the link edges, different mismatches can be identified and explicitly included in or excluded from queries, see Fig. 5.

## 5 Querying Information Status, Syntax and Prosody

As annotations from all layers are related via links, any combination of annotations can be used in a query. This means, however, that queries may become relatively complex, because all layers that must be included into or excluded from the query result need to be explicitly specified. In the trade-off between genericity and ease of query formulation, we have opted for the former.

In the following, we briefly describe a simple query, which is meant to demonstrate the interplay of the three linguistic levels of prosody, discourse (information status) and syntax. We want to investigate the prosodic realization of phrases consisting of exactly two words (in the written tokenization) which carry an information status label. This is formulated in the form of an SQL query, an excerpt of which is shown in Fig. 6. We run the query on a one-day subset of the data which at the time of publication of this paper has been integrated into the database.

For this query we generate the database table `is_syn_p`, which contains all information status labels, their corresponding text phrases and the respective accent patterns found when following the links from the tokens of the written to the tokens of the spoken dataset. We select the phrases which comprise two words (e.g. *mit Teheran*), see last line in Fig. 6, and obtain the results in Fig. 7, which show that the percentage of unaccented phrases on two-word expressions decreases along with the degree of salience: 14% of the coreference anaphora (GIVEN) are unaccented, 7% of the bridging anaphora, 4% of the generic terms, 2% of the discourse-new definites (UNUSED) and none of the specific indefinites (NEW).[6]

## 6 Availability

As the data structures of our resource are based on GrAF, which is already an exchange format, we intend to export the annotation layers in the GrAF XML format to make them available for research purposes. Figures 8 and 9 show parts of a GrAF-export for the sentence shown in Fig. 4: Fig. 8 is information on the UNUSED-KNOWN node from the information status graph, and Fig. 9 shows the representation

[6] The information status categories have been simplified in the following way: GIVEN subsumes GIVEN-EPITHET, -REPEATED, -SHORT; BRIDGING includes BRIDGING and BRIDGING-TEXT; UNUSED stands for UNUSED-KNOWN and UNUSED-UNKNOWN; NEW subsumes INDEF-NEW and INDEF-PARTITIVE; GENERIC combines INDEF-GENERIC and UNUSED-TYPE. For details, see Riester et al (2010) and Baumann and Riester (to appear).

```
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;
```

**Fig. 6** An excerpt of the SQL query discussed in Sect. 5

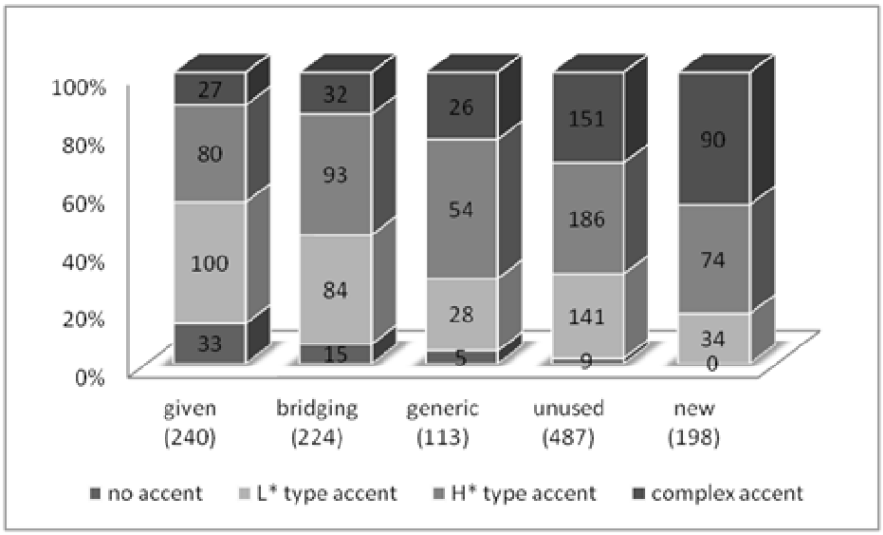


**Fig. 7** The pitch accents on two-word terms depending on information status. Results of the query in Fig. 6.

of the respective target node in the syntax tree (a *DP* node) and one of the terminal nodes in the phrase (*Tür*). The generic GrAF XML format is not only intended to be convertible into different tool input-formats but also into other graph-based generic formats, such as PAULA XML (Dipper, 2005). At the moment, different researchers also address the development of RDF and OWL linearizations of such graph-based generic formats: Cassidy (2010), for example, proposed an RDF linearization of GrAF, and Chiarcos (this vol.) developed an OWL/RDF linearization of PAULA XML. Through these recent developments, our approach is linked to the creation

```
<node xml:id="n215324_24941" />
<a ref="n215324_24941" label="a1_is_scheme">
 <fs xml:id="fs367562">
  <f value="UNUSED-UNKNOWN" name="name" />
 </fs>
</a>
<edge to="n151089_19406" from="n215324_24941"
      xml:id="e162443" />
<node xml:id="n151089_19406" />
<a ref="n151089_19406" label="xle_nonterminal">
 <fs xml:id="fs240027">
  <f value="DP[std]" name="cat" />
 </fs>
</a>
```

**Fig. 8** Samples of DIRNDL in GrAF format: UNUSED-KNOWN node from the information status graph, and its target node in the syntax tree (*DP*).

```
<node xml:id="n151049_19406" />
<a ref="n151049_19406" label="xle_terminal">
 <fs xml:id="fs239987">
  <f value="N[comm]" name="pos" />
  <f value="7" name="seq" />
  <f value="Tuer" name="word" />
 </fs>
</a>
<edge to="n151076_19406" from="n151089_19406"
      xml:id="e92409" />
<edge to="n151067_19406" from="n151089_19406"
      xml:id="e92410" />
```

**Fig. 9** Samples of DIRNDL in GrAF format: A terminal node (*Tür*) of the *DP* node referred to in Fig. 8.

of interoperable representations of multi-layer corpora by means of Semantic Web technologies, and to provide corpora as Linked Data.

**Acknowledgements** This work was funded by the German Research Foundation DFG, via the Collaborative Research Centre SFB 732 *Incremental Specification in Context*.

## References

Baumann S, Riester A (to appear) Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects. In: Elordieta G, Prieto P (eds) Prosody and Meaning, Interface Explorations, De Gruyter Mouton, Berlin

Burchardt A, Erk K, Frank A, Kowalski A, Padó S (2006) SALTO: A Versatile Multi-Level Annotation Tool. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), Genoa, Italy

Cassidy S (2010) An RDF realisation of LAF in the DADA annotation server. In: Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-5), Hong Kong

Chiarcos C (this vol.) Interoperability of corpora and annotations. P. 161-179

Chiarcos C, Ritz J, Stede M (2009) By all these lovely tokens... Merging Conflicting Tokenizations. In: Proceedings of the Third Linguistic Annotation Workshop, Association for Computational Linguistics, Suntec, Singapore, pp 35–43

Dipper S (2005) XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In: Proceedings of Berliner XML Tage 2005 (BXML 2005), Berlin, pp 39–50

Eckart K, Eberle K, Heid U (2010) An Infrastructure for More Reliable Corpus Analysis. In: Proceedings of the Workshop on Web Services and Processing Pipelines in HLT: Tool Evaluation, LR Production and Validation (LREC'10), Valletta, Malta, pp 8–14

Lezius W, Biesinger H, Gerstenberger C (2002) TIGERRegistry Manual. Tech. rep., IMS Stuttgart

Mayer J (1995) Transcription of German Intonation. The Stuttgart System, URL `http://www.ims.uni-stuttgart.de/phonetik/joerg/labman/STGTsystem.html`, ms.

Prince EF (1981) Toward a Taxonomy of Given-New Information. In: Cole P (ed) Radical Pragmatics, Academic Press, New York, pp 233–255

Prince EF (1992) The ZPG Letter: Subjects, Definiteness and Information Status. In: Mann W, Thompson S (eds) Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text, Benjamins, Amsterdam, pp 295–325

Rapp S (1995) Automatic Phonemic Transcription and Linguistic Annotation from Known Text with Hidden Markov Models – An Aligner for German. In: Proceedings of ELSNET Goes East and IMACS Workshop "Integration of Language and Speech in Academia and Industry" (Russia)

Riester A, Lorenz D, Seemann N (2010) A Recursive Annotation Scheme for Referential Information Status. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), Valletta, Malta, pp 717–722

Rohrer C, Forst M (2006) Improving Coverage and Parsing Quality of a Large-scale LFG for German. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), Genoa, Italy

Taylor P, Black AW, Caley R (1998) The Architecture Of The Festival Speech Synthesis System. In: Proceedings of the Third ESCA Workshop in Speech Synthesis, pp 147–151