German Radio Interviews: The GRAIN Release of the SFB732 Silver Standard Collection

Katrin Schweitzer, Kerstin Eckart, Markus Gärtner, Agnieszka Falenska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwien, Jonas Kuhn

University of Stuttgart, Institute for Natural Language Processing (IMS)
Pfaffenwaldring 5B, 70569 Stuttgart, Germany
{firstname.lastname}@ims.uni-stuttgart.de

Abstract

We present GRAIN (German RAdio INterviews) as part of the SFB732 Silver Standard Collection. GRAIN contains German radio interviews and is annotated on multiple linguistic layers. The data has been processed with state-of-the-art tools for text and speech and therefore represents a resource for text-based linguistic research as well as speech science. While there is a gold standard part with manual annotations, the (much larger) silver standard part (which is growing as the radio station releases more interviews) relies completely on automatic annotations. We explicitly release different versions of annotations for the same layers (e.g. morpho-syntax) with the aim to combine and compare multiple layers in order to derive confidence estimations for the annotations. Therefore, parts of the data where the output of several tools match can be considered clear-cut cases, while mismatches hint at areas of interest which are potentially challenging or where rare phenomena can be found.

Keywords: corpus, silver standard, text-speech-interface

1. Introduction

Over the years, many resources for natural language such as tools or corpora have been developed and are used in research and applications. Processing is usually focused on a specific type of primary data, which we call *canonical* data for the respective tool type or branch of research. However, with a large set of efficient tools available, the time has come to make the step beyond canonical data, at the same time connecting research branches such as those based on text and speech. The *SFB732 Silver Standard Collection* is intended to serve as a resource in this respect. It is a non-static collection, i.e. more and different primary data are added over time, as are different annotation layers.

This paper focuses on the first released part of the data set, containing richly annotated German radio interviews, made available for research and education in the GRAIN release of the collection. Different research groups contributed in its curation, among them were groups usually working primarily with written language data as well as groups who usually focus on speech, i.e. the tools applied are state-of-the art tools from both disciplines. Thereby, the resource can provide a starting point for joint exploration of speech and text. Moreover, bringing together tools from different research branches can result for some tools in encountering data that is non-canonical. For instance, parsers trained on newspaper text are to deal with incomplete sentences when being presented with spoken language data.

In the remainder of the paper, we first present the idea of a *Silver Standard*, relying on automatic annotations, then we describe the GRAIN release: the primary data is described in Section 3, details about the existing annotations can be found in Section 4. Section 5 outlines how the workflow of creating the resource is documented, Section 6 details its availability. Annotations currently under development are referred to in Section 7 and Section 8 concludes.

2. Silver Standard

The size and non-static nature of the data set make it inherently unfeasible to provide manual annotations for the entire resource. Instead only a small subset was chosen to be covered by gold standard annotations (see Section 3) and the remaining parts received automatically created annotations. For several automatic annotations we employed a "silver standard" approach (Rebholz-Schuhmann et al., 2010).

The idea behind this is that it is possible to provide a level of annotation quality, that is better than unchecked output of automated processing, even though it might not reach gold standard. To this end, automatic output has been enriched with additional confidence estimations (Gärtner and Eckart, 2017) that serve as quality indicators to increase the usability of the data. In this way, users can a) gauge the quality of the data they are working with, b) select subsets of the data where the annotations come with high confidence estimation or c) find areas of interest, for instance, when the confidence of the system is low.

Most automatic annotation systems do not provide quality indicators in their output even if they are internally aware of the relative reliability of their annotations, as is the case for all systems using probabilistic methods, e.g. stochastic parsers. Therefore, we relied on external methods for estimating confidence values, some of which have been presented in Eckart and Gärtner (2016). Note that all confidence estimations are provided as additional (meta-)-annotation layers and therefore can be used directly with regular linguistic features for visualization or search in tools such as ICARUS (Gärtner et al., 2013).

3. Primary Data

The primary data consists of German radio interviews, with a duration of just under 10 minutes each. For each in-

cluded interview an audio file (*.mp3*) and an edited transcript (mostly *.pdf*, sometimes *.doc*) were available from the radio station. The transcript has been intensely edited by the radio station to produce a version of the interview for reading and thereby omits features of orality, e.g. by excluding slips of the tongue and repetitions and by rephrasing utterances to adhere to written syntax (see Eckart and Gärtner (2016) for an example). Based on the *.pdf* and *.doc* files, which we consider raw data, primary data for the collection was extracted in the form of UTF-8 encoded plain text files.[1]

The audio files are heterogeneous in their characteristics: there are stereo as well as mono files, with either 44.1kHz sampling rate or 48kHz, and with varying audio bitrates (64-135kbps). The release contains the original mp3 files. For processing the files with our tools, we converted them to 16kHz mono wav-files. These more consistent files are made available, as well. Note that the basis though is mp3, i.e. the wav-files do not provide better quality. Each interview involves two speakers, a host and a guest. The guest appears in a professional role, and the questions of the host usually refer to a current political or social discussion at the time of recording.

Gold standard part and training interviews. A set of 20 interviews has been selected to serve as primary data for a gold standard part of the collection. Three additional interviews have been marked as training interviews for annotators being introduced to guidelines for manual annotation. Since the annotation efforts are conducted by several projects, globally defining these interviews within the collection minimizes the set of interviews for which specific restrictions apply, e.g. in evaluation settings.[2] The interviews for the gold standard part were balanced as well as possible with respect to sex and variety of the host[3] and sex and role of the guest[4].

Size of the dataset. Since the collection is non-static more interviews are added over time. The current status is 144 interviews, with about 221,000 word tokens and a duration of about 23 hours.

4. Available Annotation Layers

The GRAIN release provides annotation layers resulting from state-of-the-art text and speech processing and is therefore suited for examinations from either spoken or written language research, as well as studies at the interface of the two. The text-based annotation layers are linked with LAF anchors, the annotations referring to the audio signals are anchored via timestamps. The mapping between the two is done on the basis of word tokens.

[1] Apache PDFBox 1.8.7.

[2] It is of course possible to include further training interviews if needed or to annotate a training interview with gold quality.

[3] Two (of at that time three encountered) female hosts with five guests each, five (of at that time ten encountered) male hosts with two guests each.

[4] For the balancing we distinguished between guest interviewed w.r.t. to a political or a non-political role, i.e. the data is balanced for their role in the interview, not their usual background or former professions.

In what follows we discuss the annotations that are part of this GRAIN release.

4.1. Automatic Annotations

Automatic annotations are created for the entire dataset, i.e. all interviews for which the radio station provided both the audio file and the edited transcript. This part of the corpus, which contains only automatic annotations is what we refer to as the silver-standard part of the release, since it offers the possibility do combine various annotation levels to gauge the quality of the annotations by e.g confidence estimations.

Preprocessing. As a first step, anchors according to LAF (ISO 24612:2012) were introduced for the primary data text files. The base units of these files are UTF-8 characters, each identified by two numerical anchors, describing the one-character span in the data. The anchors allow for several layers of (stand-off) annotations to be linked to the primary data document. Further annotations for the textual data include speaker turn spans and document structure. Based on the described information available, several input formats for subsequent processing steps could be created.

Tokenizing and sentence segmentation. The data was tokenized with the TreeTagger (Schmid, 1994). Sentence segmentation was done on top, based on punctuation tokens.

Acoustic segmentation and alignment. The data was force-aligned for phone, syllable and word boundaries (Rapp, 1995).

Parametrized intonation events. PaIntE parameters were calculated for each syllable in the data (Möhler, 2001; Möhler and Conkie, 1998; Möhler, 1998). PaIntE stands for "Parametrized Intonation Events" and presents a way to describe the shape of local maxima in the pitch contour (that is, highly probable candidates for pitch accents or boundary tones) by means of six parameters. Parametrization is carried out using a function over time which approximates the fundamental frequency contour. The function comprises six free parameters that are fitted in such a way that the actual fundamental frequency curve is matched best. All six parameters are linguistically interpretable: parameter `d` corresponds to the height of the peak in Hertz, parameter `b` encodes its temporal anchoring within a three syllable window where the syllables are normalized for time, such that the current syllable ranges between 0 and 1. Parameters `c1` and `c2` stand for size of the increase before and the decline after the peak, again in Hertz, and parameters `a1` and `a2` encode the gradient of the rise and fall, respectively. Figure 1 (adapted from Möhler (2001)) displays the parameters in the 3-syllable window.

PaIntE-based prediction of intonation event types. Intonation events, in terms of GToBI(S) labels (Mayer, 1995) for pitch accents and boundary tones, were annotated automatically with the procedure described in Schweitzer (2010). This method takes into account PaIntE parametrizations and normalized phone durations, phonological features and higher linguistic information.

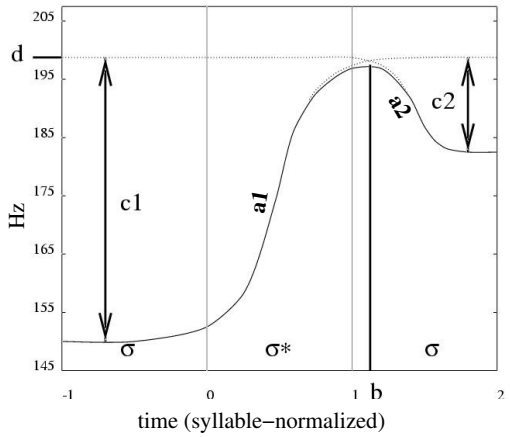


Figure 1: The PaIntE function. Approximation takes place over three syllables. $\sigma*$ marks the syllable for which approximation is currently being carried out. In the depicted case, the parameter setting corresponds to a peak in the following syllable (encoded in parameter b) with a pronounced rise before the peak (c1) and a small fall after (c2). Absolute peak height is encoded in the value of parameter d, and the gradients of the rise and fall can be derived from parameters a1 and a2.

CNN-based prediction of pitch accent placement. The annotations in this layer consist of binary placement information for pitch accents at the word level. Even though no type of pitch accent was assigned, this layer makes it possible to apply the silver standard idea to prosodic annotations: placement of the automatically predicted pitch accent labels derived from PaIntE features (see above) can be compared and combined with this layer for confidence estimations. Speech signals and time aligned word labels were input to a convolutional neural network-based binary classifier (CNN) that predicts for each word whether it carries a pitch accent or not (Stehwien and Vu, 2017). The input to the CNN was a frame-based representation of the speech signal using low-level acoustic descriptors extracted using OpenSMILE (Eyben et al., 2013). The model was trained on the German radio news corpus DIRNDL (Eckart et al., 2012).

Agreement on pitch accent placement. Since GRAIN contains no manually annotated intonation events, the prediction accuracy could only be estimated on a small amount of data labeled by a human expert. In the following, we report the agreement with respect to word-level pitch accent placement between the two annotation layers (PaIntE-based and CNN-based) and compared to human annotation. Table 1 shows the performance of the two tools measured against the reference annotations. In terms of accuracy, both tools provide annotations of similar quality. The CNN-based tool yields a higher recall, while the PaIntE-based annotations have a higher precision. Table 2 compares the precision on either label classes *accent* and *none* obtained using both methods and when counting only the labels that the two tools agree on. These results show that in cases where both annotations agree wrt presence or absence of pitch accents, the precision increases considerably for both classes (up 10% for the accent class) compared to when used alone. While this evaluation can only provide an estimation of performance (especially when using only one human labeler), the reported numbers do show that both tools, using entirely different methods of prosodic modeling, complement each other and that using both can be used to estimate the confidence of automatic annotation. It also demonstrates our idea of a silver standard: combining two layers of automatic annotation on the same annotation level achieves a better quality than just one level alone.

	PaIntE-based	CNN-based
accents	494	629
accuracy	80.9%	80.0%
precision	67.2	62.2
recall	65.2	76.8
F1-score	66.2	68.7

Table 1: Agreement with human labeling of word-level pitch accent placement: The PaIntE-based and CNN-based automatic annotations are compared on one example file containing 1778 words and 509 pitch accents.

class	PaIntE-based	CNN-based	both
accents	67.2	62.2	76.0
none	86.2	89.7	94.2

Table 2: Precision for both label classes *accent* and *none* obtained using either pitch accent labeling method and when taking only the labels into account on which both tools agree.

Additional phonetic features. The data was preprocessed using the Festival (Black, 1997) version of the University of Stuttgart (IMS Festival, 2010) to retrieve some of the features needed in the automatic annotation process for prosodic events. We release some of the syllable-based features, for convenience. We provide the duration of each syllable, its position in the word, and the number of phonemes in onset and rhyme, as well as the Van Santen/Hirschberg Classification (van Santen and Hirschberg, 1994)
of onset and rhyme.

Morpho-syntax. On the morpho-syntactic level we employed a series of very different pipeline implementations (BitPar (Schmid, 2006; Schmid, 2004), IMS-SZEGED-CIS (Björkelund et al., 2013), Mate (Bohnet and Nivre, 2012; Bohnet, 2010), IMSTrans (Björkelund and Nivre, 2015; Björkelund et al., 2016) and Stanford CoreNLP components such as the Stanford Parser (Chen and Manning, 2014)) to generate automatic parses and underlying morpho-syntactic annotations for the entire data set (see Table 3). Since we do not have (morpho-)syntactic gold standard annotations for this release, we post-processed the automatic system output to improve its usability, but without actually changing or correcting it. That is, following the idea of a silver-standard, we introduced additional confidence estimations as meta annotations for individual predictions based on the agreement between different systems (see Figure 2a for an example of trees predicted by three different parsing systems for the same sentence and Figure 2b for corresponding global confidence estimations).

While this extra step does not directly increase the annotation quality per se, it provides valuable information about the relative reliability of individual annotations. Researchers can then use those indicators to find data points which might be of interest or should be ignored for certain research questions.

System	Constituency	Dependency
BitPar	+	
IMS-SZEGED-CIS	+	+
Mate		+
IMSTrans		+
Stanford Parser	+	+

Table 3: List of automatic (pipeline) systems for parsing used to generate concurrent annotations for the corpus.

4.2. Manual Annotations

Manual annotations were conducted on the interviews of the gold standard part of the collection. That is, the manual annotations are additional annotations (not corrections) besides the automatic annotations. We refer to the part of the corpus for which manual annotations are available as the gold-standard part of this release. This part constitutes a subset of the silver-standard part, but has been labeled independently from any automatic annotations.

Unnormalization. To provide a textual version of the interviews suited for several processing pipelines, the edited versions were modified (cf. Eckart and Gärtner (2016) for a motivation and additional details of this additional layer): Based on the audio signal, some features of orality were re-introduced. However, fillers and partially uttered words were not included. This resulted in transcripts that are slightly closer to an orthographic transcription of the utterances as compared to the edited versions from the radio station. We call this process *unnormalization*.[5] Guidelines have been defined and each interview was modified independently by two annotators and adjudication was then done by a third person.
In Eckart and Gärtner (2016) we quantified the difference between the edited version provided by the radio station and the unnormalized versions in terms of the quality of automatic parsing. For the current release of GRAIN, we additionally computed a raw measure of difference between the edited versions and the result of our unnormalization. Using Levenshtein Distance on entire interviews and treating each token as a symbol we calculated edit distances that ranged between 21 and 148 with an average of about 54.
The manual annotations described in the following sections have been conducted on the unnormalized version of the interviews. For details on which automatically generated annotations also use this layer we refer to the documentation that is part of the release.

Part-of-speech tagging. The interviews were annotated with part-of-speech labels based on the STTS guidelines (Schiller et al., 1999) including the modifications from the TIGER corpus (Albert et al., 2003). Some additional guidelines were set up for the interview corpus (Seeker, 2016) but due to the specificities of the interviews no further categories were needed (cf. Westpfahl et al. (2017) for a broader set for spoken data). Three annotators were involved in the process and each interview was annotated by two of them independently, applying the Synpathy tool[6]. The annotators achieved pair-wise agreement with a Cohen's κ of 0.97, ranging between 0.96 and 0.98. In an adjudication step all three annotators then decided on the annotation, and remaining hard cases were discussed in the project context and documented separately. After all interviews had been manually annotated and discussed, an implementation of the DECCA-Tools (Dickinson and Meurers, 2003) in ICARUS (Thiele et al., 2014) was applied to the interviews, automatically finding potential cases of inconsistent annotation.

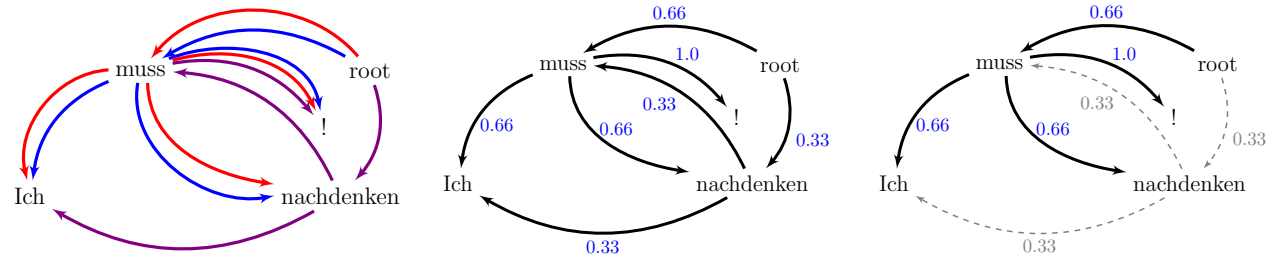
Referential information status. From the gold part 20 interviews and the three training interviews were annotated with referential information status (Baumann and Riester, 2012), following the guidelines in Riester and Baumann (2017). This means that all referring expressions in the interviews (and a number of verb phrases and sentences functioning as antecedents for abstract anaphors) were categorized as to whether they are given/coreferential, bridging anaphors, deictic, discourse-new, idiomatic etc. The interviews furthermore contain coreference chains and bridging links. Each of the interviews was annotated independently by two annotators, applying the Slate tool (Kaplan et al., 2012). Adjudication was either done by a third person, or in a discussion round of the project group.
The inter-annotator-agreement has been computed for markables with the same span, where we have achieved substantial agreement, with a Cohen's κ of 0.75. Five different annotators were involved in the annotation (all students of computational linguistics) and the pair-wise agreement for different annotator pairs (Cohen's κ) ranges between 0.64 and 0.82. For more details on the inter-annotator agreement, please refer to Pagel (2018) and Draudt (2018).

Questions under discussion (QUD). From the gold part, ten interviews and the three training interviews were analyzed according to the *QUD-tree* method (Reyle and Riester, 2016; Riester et al., to appear), which involves a new (sub-sentential) text segmentation into information-structurally relevant discourse units, the reconstruction of implicit *questions under discussion* (QUDs) for each unit, based on a number of pragmatic principles, and the construction of question-based discourse trees (QUD trees) with TreeAnno (De Kuthy et al., 2018). The annotations were created by two annotators. Adjudication was subsequently done within the project group. More detail and evaluation of the annotation of QUDs can be found in De Kuthy et al. (2018).

Information structure. As described in Riester et al. (to appear), the QUD-tree method is a joint approach for the

[5] This term is inspired by a step called normalization often applied to map non-canonical representations to some sort of standard or processable forms.

[6] http://www.mpi.nl/tools/synpathy.html



(a) Predictions from three different parsers. (b) Confidence estimations. (c) Merged tree.

Figure 2: An example sentence "Ich muss nachdenken!" (eng. "I have to think!") and different layers of automatically predicted dependency annotations.

analysis of both discourse structure and information structure. The implicit QUDs define which parts of the discourse units receive either of the labels focus, contrastive topic, background and non-at-issue. We used the annotation tool Slate (Kaplan et al., 2012) for this annotation task. For more details on the annotation of discourse structure and information structure we also refer to De Kuthy et al. (2018).

5. Documentation

Besides the primary data and our various annotation layers we also created detailed documentation for the entire workflow of resource creation. The specific version of each annotation tool, the versions and nature of data used to configure or train it and also the settings used for the actual analysis are all crucial information needed to properly evaluate the output and its suitability (in this case of the final corpus resource) in the context of a certain research question. We therefore used a simple metadata scheme similar to the one proposed by Gärtner et al. (2018) for recording *process metadata*.

Metadata for manual annotations includes amongst other information annotator ids[7], details of manual curation as well as applied annotation guidelines. In the case of automatic steps the recorded metadata is very similar and additionally contains version information for involved resources and/or tools, where available. Figures 3 and 4 show (condensed) instances of this metadata for manual and automatic processing steps, respectively. The entire process metadata is available as part of the corpus release together with the overall documentation.

6. Availability

Due to the high number of different tools involved in the creation of GRAIN and in order to accommodate researchers from different communities, various representation formats are part of this release. They contain, for instance, popular tabular formats such as those used in the CoNLL Shared Tasks of 2009 (Hajič et al., 2009) and 2012 (Pradhan et al., 2012) and extended versions (Björkelund et al., 2014), XML-based formats such as TIGER XML (König et al., 2003), or Praat TextGrids (Boersma, 2001). For an exhaustive list of formats used and annotation layers contained in them we refer to the official documentation

[7] In anonymized form.

```
[result: ["swr2-interview-der-woche-20150620.txt-mod"],
 input: ["swr2-interview-der-woche-20150620.txt","swr2-
     interview-der-woche-20150620.6444m.mp3"],
 workflowSteps:[
   [description: "transcription modification",
    mode: "manual",
    operators:[
      [name: "OP01",
       components:[
         [name: "transcription-change-guidelines",
          version: "1.1.0",
          type: "guidelines"
         ]]
      ]
      [name: "OP02",
       components:[
         [name: "transcription-change-guidelines",
          version: "1.1.0",
          type: "guidelines"
         ]]
      ]]
   ]
   [description: "conflict resolution",
    mode: "manual",
    operators:[
      [name: "OP04",
       components:[
         [name: "transcription-change-guidelines",
          version: "1.1.0",
          type: "guidelines"
         ]]
      ]]
   ]]
]
```

Figure 3: Example process metadata for two independent manual annotation steps by different annotators (`operators`) for unnormalization and a subsequent adjudication step by a third person.

that is part of the resource. Individual annotation files in the corpus also follow a simple naming scheme that contains the part of the primary data the annotations are associated with and reflects the processing step that created the data. The release, as well as a detailed documentation is published in the framework of CLARIN[8] and available via a persistent identifier[9] in order to ensure sustainability.

7. Annotation Layers under Development

The SFB732 Silver Standard Collection is non-static, i.e. primary data and annotation layers will continue to be added. The annotations currently under development or planned for future releases are listed below.

[8] `https://www.clarin.eu/`

[9] `http://hdl.handle.net/11022/1007-0000-0007-C632-1`

```
{"result": ["swr2-interview-der-woche-20140517.3.2.1.0.
    tag-tt"],
 "input": ["swr2-interview-der-woche-20140517.1.0.0.0.
     tok-tt-mod"],
 "workflowSteps":[
   {"description":"TreeTagger: lexicon lookup, part-of-
       speech and lemma annotations, error correction",
    "mode": "automatic",
    "operators":[
      {"name": "tree-tagger-3.2.1-german-notok-nosgmlrec
          ",
       "version": "3.2.1 (TreeTagger)",
       "parameters": "",
       "components":[
         {"name": "german-lexicon-utf8.txt",
          "version": "3.2.1 (TreeTagger)",
          "type": "lexicon"
         },
         {"name": "german-utf8.par",
          "version": "3.2",
          "type": "parameter file"
         }]
      }]
   }]
}
```


Figure 4: Example process metadata for an automatic processing step consisting of part-of-speech tagging and lemmatization with a subsequent automatic error correction.

7.1. Automatic Annotations

In addition to the annotations listed in Section 4.1 we plan to include the following annotations in subsequent releases of the silver standard part of the corpus.

Merged dependency parses. As part of the parsing outputs described in Section 4.1 several parallel dependency trees are available for every sentence. While this provides a rich foundation for comparison, it can also be challenging for users to work with multiple concurrent trees. We will therefore provide additional merged versions of dependency trees. That is, we will include a majority decision of the dependency parsers under tree constraints.
This is done by employing blending, also known as reparsing (Sagae and Lavie, 2006). We combine all the silver standard trees for a sentence into one graph and assign scores to arcs depending on the confidence estimations (see Figure 2b for an example of a combined graph). We then use the Chu-Liu-Edmonds algorithm (Chu and Liu, 1965; Edmonds, 1967) to find the maximum spanning tree in the combined graph (see Figure 2c for the resulting maximum spanning tree). For every resulting arc we select the most frequent label across all the labels previously assigned to it. This additional layer of automatic annotations has two purposes. It improves the usability of the syntactic layer for users who may prefer to work with only a single dependency tree instead of multiple predicted ones. Secondly, it increases the reliability of the syntactic annotations, since Björkelund et al. (2017) showed that blending can achieve higher performance than single parsers.

CNN-based prediction of boundary tone placement. In addition to the pitch accent placement labels predicted using a CNN-based classifier (described in Section 4.1), a similar model (extended to include duration and pause information) will be trained to label each word as bearing a phrase boundary tone or not. This model will require more annotated data from additional English sources e.g. from BURNC (Ostendorf et al., 1995).

7.2. Manual Annotations

The manual annotation layers under development will be added to the gold standard subset of the corpus, which was also used for the annotations described in Section 4.2.

Unedited orthographic transcripts. Currently, orthographic transcriptions of another granularity are being created, additionally to the edited version provided by the radio station (see Section 3) and the "unnormalized" version which was used as a basis for the text-based manual annotations (see Section 4.2). This version is as close to the audio files as possible, i.e. it contains fillers and other non-lexical information, partially uttered words, mispronunciations, non-standard pronunciations etc., and gives information about overlap between the speakers. Guidelines have been defined which are based on the guidelines used for the GECO corpus (Schweitzer and Lewandowski, 2013) and some aspects of the definition of the "verbal tier" in the HIAT guidelines (Rehbein et al., 2004).

8. Conclusion

We presented the GRAIN release of the SFB732 Silver Standard Collection. The data comprises audio files of German radio interviews and their transcripts provided by the broadcasting station. We provide (manual) gold standard annotations for a subset of 20 radio interviews. These annotations include a transcript which is closer to the audio files than the edited transcript, POS tags, referential information status annotations, and questions under discussion. Additionally, a much larger data set (currently 160 interviews) has been annotated with silver standard annotations, i.e. automatic annotations which can be combined and compared, both across as well as between layers, in order to make it possible to infer a confidence estimation for the annotations. These annotations include information about speakers and their roles, time-aligned word, phone and syllable labels, parametrized intonation events, GToBI(S) intonation labels, CNN-based annotation of pitch accents, additional syllable features and morpho-syntactic annotations. For the syntax annotations, confidence estimations are already provided in this release. The silver standard part of the data is growing: as the radio station releases more interviews, they are being collected and automatically processed. , new annotation layers are currently being created. These will comprise automatic annotations in the form of merged dependency parses, CNN-based boundary tone placement as well as manually created information structure labels and a version of the transcript with all features of orality.

9. Acknowledgements

We would like to thank the numerous dedicated annotators for their contributions. This work was funded by the German research foundation (DFG) via SFB 732, projects INF, A4, A6, A8 and D8.

10. Bibliographical References

Albert, S., Anderssen, J., Bader, R., Becker, S., Bracht, T., Brants, S., Brants, T., Demberg, V., Dipper, S., Eisenberg, P., Hansen, S., Hirschmann, H., Janitzek, J., Kirstein, C., Langner, R., Michelbacher, L., Plaehn, O., Preis, C., Pußel, M., Rower, M., Schrader, B., Schwartz, A., Smith, G., and Uszkoreit, H., (2003). *TIGER Annotationsschema*. Universität des Saarlandes, Universität Stuttgart, Universität Potsdam.

Baumann, S. and Riester, A. (2012). Referential and lexical givenness: Semantic, prosodic and cognitive aspects. In Gorka Elordieta et al., editors, *Prosody and Meaning*, pages 119–161. De Gruyter Mouton, Berlin.

Björkelund, A. and Nivre, J. (2015). Non-Deterministic Oracles for Unrestricted Non-Projective Transition-Based Dependency Parsing. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 76–86, Bilbao, Spain, July. Association for Computational Linguistics.

Björkelund, A., Cetinoglu, O., Farkas, R., Mueller, T., and Seeker, W. (2013). (re)ranking meets morphosyntax: State-of-the-art results from the SPMRL 2013 shared task. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 135–145, Seattle, Washington, USA, October. Association for Computational Linguistics.

Björkelund, A., Eckart, K., Riester, A., Schauffler, N., and Schweitzer, K. (2014). The extended DIRNDL corpus as a resource for coreference and bridging resolution. In Nicoletta Calzolari et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3222–3228, Reykjavik.

Björkelund, A., Faleńska, A., Seeker, W., and Kuhn, J. (2016). How to train dependency parsers with inexact search for joint sentence boundary detection and parsing of entire documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1924–1934, Berlin, Germany, August. Association for Computational Linguistics.

Björkelund, A., Falenska, A., Yu, X., and Kuhn, J. (2017). Ims at the conll 2017 ud shared task: Crfs and perceptrons meet neural networks. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 40–51. Association for Computational Linguistics.

Black, A. W. (1997). The Festival speech synthesis system. www.cstr.ed.ac.uk/projects/.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Bohnet, B. and Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea, July. Association for Computational Linguistics.

Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August. Coling 2010 Organizing Committee.

Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In Alessandro Moschitti, et al., editors, *EMNLP*, pages 740–750. ACL.

Chu, Y. and Liu, T. (1965). On the shortest aborescence of a directed graph. *Science Sinica*, 14:1396–1400.

De Kuthy, K., Reiter, N., and Riester, A. (2018). QUD-based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, Miyazaki, JP.

Dickinson, M. and Meurers, W. D. (2003). Detecting errors in part-of-speech annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 107–114, Budapest, Hungary. Association for Computational Linguistics.

Draudt, A.-C. (2018). "Inter-Annotator Agreement von Informationsstatus-Annotationen im GRAIN-Korpus" (BSc thesis).

Eckart, K. and Gärtner, M. (2016). Creating Silver Standard Annotations for a Corpus of Non-Standard Data. In Stefanie Dipper, et al., editors, *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, volume 16 of *BLA: Bochumer Linguistische Arbeitsberichte*, pages 90–96, Bochum, Germany.

Eckart, K., Riester, A., and Schweitzer, K. (2012). A discourse information radio news database for linguistic analysis. In Christian Chiarcos, et al., editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 65–75. Springer, Heidelberg.

Edmonds, J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards*, 71(B):233–240.

Eyben, F., Weninger, F., Groß, F., and Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838.

Gärtner, M. and Eckart, K. (2017). In support of self-assessment – exploiting available information from tools. Poster accepted at the DGfS CL Poster Session 2017, Saarbrücken.

Gärtner, M., Thiele, G., Seeker, W., Björkelund, A., and Kuhn, J. (2013). ICARUS – an extensible graphical search tool for dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Sofia, Bulgaria, August. Association for Computational Linguistics.

Gärtner, M., Hahn, U., and Hermann, S. (2018). Supporting sustainable process documentation. In Georg Rehm et al., editors, *Language Technologies for the Challenges of the Digital Age*, pages 284–291, Cham. Springer International Publishing.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D.,

Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., and Zhang, Y. (2009). The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 1–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

IMS Festival. (2010). IMS German Festival home page.

Kaplan, D., Iida, R., Nishina, K., and Tokunaga, T. (2012). Slate - A Tool for Creating and Maintaining Annotated Corpora. *Journal for Language Technology and Computational Linguistics*, 26(2):89 – 101.

König, E., Lezius, W., and Voormann, H., (2003). *TIGERSearch 2.1 User's Manual. Chapter V - The TIGER-XML treebank encoding format*. IMS, Universität Stuttgart.

Mayer, J. (1995). Transcribing German intonation – the Stuttgart system. Technical report, Universität Stuttgart.

Möhler, G. and Conkie, A. (1998). Parametric modeling of intonation using vector quantization. In *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, pages 311–316.

Möhler, G. (1998). Describing intonation with a parametric model. In *Proceedings of the International Conference on Spoken Language Processing*, volume 7, pages 2851–2854.

Möhler, G. (2001). Improvements of the PaIntE model for F_0 parametrization. Technical report, Institute of Natural Language Processing, University of Stuttgart. Draft version.

Ostendorf, M., Price, P. J., and Shattuck-Hufnagel, S. (1995). The Boston University Radio News Corpus. Technical Report ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA.

Pagel, J. (2018). Rule-based and Learning-based Approaches for Automatic Bridging Detection and Resolution in German (MSc thesis).

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *EMNLP-CoNLL: Shared Task*, pages 1–40, Jeju Island, Korea, July.

Rapp, S. (1995). Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov models—An aligner for German. In *Proc. of ELSNET Goes East and IMACS Workshop "Integration of Language and Speech in Academia and Industry" (Russia)*.

Rebholz-Schuhmann, D., Jimeno-Yepes, A. J., van Mulligen, E. M., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Tomanek, K., Beisswanger, E., and Hahn, U. (2010). The calbc silver standard corpus for biomedical named entities - a study in harmonizing the contributions from four independent named entity taggers. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Rehbein, J., Schmidt, T., Meyer, B., Watzke, F., and Herkenrath, A., (2004). *Handbuch für das computergestützte Transkribieren nach HIAT*.

Reyle, U. and Riester, A. (2016). Joint information structure and discourse structure analysis in an Underspecified DRT framework. In Julie Hunter, et al., editors, *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem)*, pages 15–24, New Brunswick, NJ, USA.

Riester, A. and Baumann, S. (2017). *The RefLex Scheme – Annotation Guidelines*, volume 14 of *SinSpeC. Working Papers of the SFB 732*. University of Stuttgart.

Riester, A., Brunetti, L., and De Kuthy, K. (to appear). Annotation guidelines for Questions under Discussion and information structure. In Evangelia Adamou, et al., editors, *Information Structure in Lesser-Described Languages: Studies in Syntax and Prosody*. Benjamins, Amsterdam.

Sagae, K. and Lavie, A. (2006). Parser combination by reparsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 129–132, New York City, USA, June. Association for Computational Linguistics.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Schmid, H. (2004). Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

Schmid, H. (2006). Trace prediction and recovery with unlexicalized pcfgs and slash features. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 177–184, Sydney, Australia, July. Association for Computational Linguistics.

Schweitzer, A. and Lewandowski, N. (2013). Convergence of articulation rate in spontaneous speech. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013, Lyon)*, pages 525–529.

Schweitzer, A. (2010). *Production and Perception of Prosodic Events – Evidence from Corpus-based Experiments*. Doctoral dissertation, Universität Stuttgart.

Seeker, W. (2016). Guidelines for the Annotation of Syntactic Structure in the IMS Interview Corpus.

Stehwien, S. and Vu, N. T. (2017). Prosodic event detection using convolutional neural networks with context information. In *Proceedings of Interspeech*, pages 2326–2330.

Thiele, G., Seeker, W., Gärtner, M., Björkelund, A., and

Kuhn, J. (2014). A graphical interface for automatic error mining in corpora. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 57–60, Gothenburg, Sweden, April. Association for Computational Linguistics.

van Santen, J. and Hirschberg, J. (1994). Segmental effects on timing and height of pitch contours. In *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP 94)*, pages 719–722, Yokohama, Japan, 09.

Westpfahl, S., Schmidt, T., Jonietz, J., and Borlinghaus, A. (2017). Stts 2.0. guidelines für die annotation von pos -tags für transkripte gesprochener sprache in anlehnung an das stuttgart tübingen tagset (stts).

11. Language Resource References

Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue.

Eckart, Kerstin and Riester, Arndt and Schweitzer, Katrin. (n.d.). *DIRNDL – Discource Information Radio News Database for Linguistic analysis*.

Ostendorf, Mari and Price, Patti J. and Shattuck-Hufnagel, Stefanie. (1995). *The Boston University Radio News Corpus*. Electrical, Computer and Systems Engineering Department, Boston University.

Schweitzer, Antje and Lewandowski, Natalie. (2013). *German Conversations: The IMS GECO database*. IMS, University of Stuttgart.