Creating Silver Standard Annotations for a Corpus of Non-Standard Data

Kerstin Eckart **Markus Gärtner**
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Pfaffenwaldring 5b, D-70569 Stuttgart
{kerstin.eckart,markus.gaertner}@ims.uni-stuttgart.de

Abstract

We present our approach for annotating a large collection of non-standard multimodal data. Its automatically created silver standard annotations lack the quality of their manual counterparts but will be enriched with confidence estimations which allow an assessment of an annotation's expected correctness. For this we first aim at providing many different annotation layers with multiple concurrent annotations. The approach is exemplified on a collection of German radio interviews and their transcripts. Finally we argue for inclusion and consideration of a tool's own confidence values in annotations and research.

1 Introduction

Non-standard data is data which is not typical for a specific application or line of research. And since in text and speech processing, many tools nowadays work well on their typical data, time has come to take the next logical step towards other domains, modalities, languages, registers and time periods. This includes of course the handling of additional phenomena. Switching the domain might change the vocabulary and switching the modality might blur the basic structure of processing, e.g. when a parser which bases its analysis on the unit of a sentence is applied to spontaneous speech. Amongst others, shared tasks have fostered the development of approaches to domain adaptation (Petrov and McDonald, 2012), and the development of approaches that can be applied to several languages (Seddah et al., 2014).

For many tools, a set of high-quality annotated data is needed to train them on, or be adapted to. For German, the NoSta-D corpus (Dipper et al., 2013) provides a collection of non-standard data, including historical data, chat data, learner data, literary prose and spoken data from a map task. Since all parts have been manually annotated, the corpus can be used in training and evaluation. However, to study specific or less frequent phenomena, huge corpora might be necessary (Zarrieß et al., 2013).

Our goal is to provide a large collection of non-standard data for various research fields which includes two non-standard areas, spoken and web data. The data will be enhanced with several annotation layers, including interaction of tools from text and speech processing. Due to the size no full manual annotation is possible, therefore we opt for a silver standard approach, as exemplified in (Rebholz-Schuhmann et al., 2010). The silver standard provides annotation quality between gold standard and uncontrolled automatic annotation. For this, we combine information from multiple tools and annotation layers, include manual and automatic annotations, and argue for a visible confidence estimation along with annotations. We present the silver standard idea in Section 3, and focus on a current set of speech data, for which we introduce an "unnormalized" layer that constitutes non-standard data for both speech and text processing pipelines.

2 Data

The data set we focus on here is a collection of German radio interviews. The primary data available from the radio station consists of recordings of the interviews (.mp3) and edited transcripts (.pdf)[1].

The data set is non-static, i.e. more interviews are being added. At the time of writing the set comprises ca. 100 interviews of about 10 minutes length each, collected from broadcasts between May 2014 and July 2016.

The setting of the interviews is such that a host from the radio station interviews a guest on topics from the (at that time) current political and social discussion. The guest appears in a professional role

[1] For a few transcripts a .doc file was made available instead of a .pdf file.

(political representative, commissioner, founder of an association, managing director, etc.).

The definition of non-standard data varies with the task or line of research in which the data is applied. The NoSta-D corpus (Dipper et al., 2013) contains several different subcorpora of data that is considered non-canonical; and while Hirschmann et al. (2007) state that non-canonical cases can only be defined with respect to a canon – in their case a linguistic framework or an annotation scheme, Petrov and McDonald (2012) go further in the direction of processability by a tool. Transferring the latter to speech corpora includes e.g. data that is non-canonical due to recording settings. Additionally, what is non-standard data for one setting might be completely canonical in another.

Since our goal is to enhance data with various layers of annotations, we consider this data non-standard in various respects.

Regarding spoken data, planned or read speech recorded under laboratory conditions is clearly more canonical for processing and annotating than spontaneous conversations recorded in a noisy environment. Our data set is somewhere in between: semi-planned speech[2], recorded in the studio of a professional radio station. Despite the latter, the available audio recordings contain both speakers in the same file and while there is only little overlap, we regard the data as non-standard with respect to processing. An additional dimension for non-standard speech is the eloquence of the speaker. While the hosts are professional speakers from the radio station the guests vary along this dimension.

Regarding written data, newspaper text is adequately processable by most tools. Thus, non-standard data for these tools includes e.g. web data, historical data, and also written representation of features of orality. The transcripts which the radio station provides are however an edited version of the interview. The transcriber introduces sentence borders, corrects the syntax and even adds words where necessary to form a sentence. Thereby the transcripts are rather canonical data for text processing and neither include fillers, false starts or repairs nor do they necessarily keep the original syntax. Since it is our goal to adapt our text processing tools (in small steps) to more non-standard data, we reintroduce some of the features of orality to the transcripts, cf. Section 4, i.e. we create a closer transcription of what was actually said.

[2]Topics of the interview are probably known in advance.

3 Silver Standard Approach

The data described in Section 2 is part of an ongoing initiative to create a so called *silver standard collection* in the SFB732 [3]. It is meant to contain a large number of annotated resources that vary with regards to modality, language, domain and (non-)canonicality. Since manual annotations are not feasible for such a large data set[4], annotations need to be created automatically. For this the term "silver standard" describes a level of annotation quality between a manually created gold standard and the unchecked output of automatic processing. Sections 3.1 to 3.3 outline the annotation project and describe methods usable to ensure an adequate level of annotation quality or to provide quality indicators.

3.1 Variety of Annotations

Besides previously introduced radio interviews the silver standard collection will contain French radio conversations and a selection of already available German and English web corpus data. It covers different modalities (speech, written transcripts, textual web data), languages (German, French, English) and domains (interviews, conversations, blog/forum posts), making it an ideal source of non-standard data for many research fields.

While the goal is to provide a large number of automatically annotated resources that contain various types of non-standard phenomena, we still need a small set of manually annotated gold data for training or evaluation. For the German radio interviews we selected a subset of 20 interviews and their transcripts, totaling ~ 3 hours of audio and ~ 36.000 written tokens. They are being annotated for part-of-speech (TIGER/STTS guidelines by Brants et al. (2004) and Schiller et al. (1999) with additions by Seeker (2016)), information status (RefLex scheme by Baumann and Riester (2012)) and discourse.

The different data sets in the silver standard collection will then receive stand-off annotations created automatically by several tools (cf. Section 3.2) for multiple layers. Figure 1 shows a simplified version of the annotation workflow. It indicates where

[3]SFB: Sonderforschungsbereich (Collaborative Research Center) `http://www.uni-stuttgart.de/linguistik/sfb732/`

[4]For example the time cost for prosodic annotations of speech data according to the Tones and Break Indices (ToBI) system alone is around 100-200 times the real time (Syrdal et al., 2001) for experienced annotators.

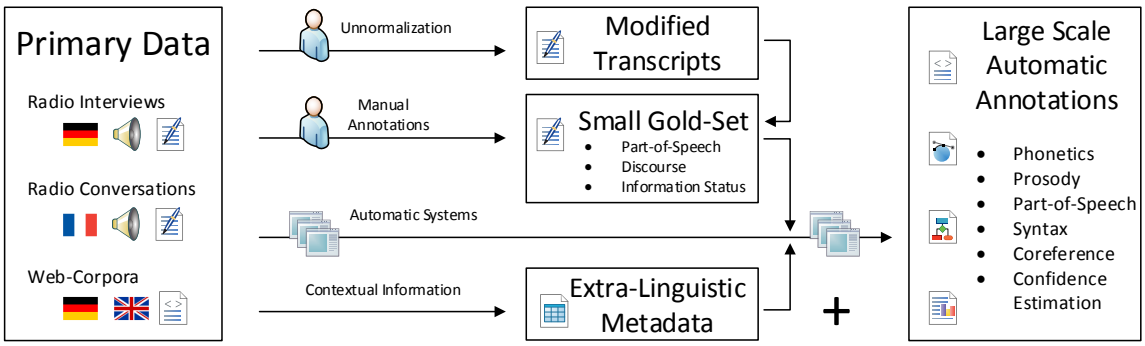

Figure 1: Composition of the silver standard collection and annotation workflow for the subset of German radio interviews (cf. Section 2).

direct human work is involved in the annotation process and which types of automatic annotations we plan to make available. Besides this vertical variety of annotation types there will also be horizontally concurrent annotations for (ideally) each type. That is, we intend to use multiple tools to create annotations for the same level. Those (potentially different) outputs can both help to get a better understanding of the data at hand or offer the basis for confidence estimations (cf. Section 3.3) or indicators for processability of data points.

We will also include extra-linguistic information[5] as additional annotations, if available. This allows to incrementally take more context into account when analyzing data. This is especially true for speech data, where Lewandowski (2013) showed the relevance of personality-related information for phonetic convergence.

Besides extra-linguistic information derived directly from the primary data, we also aim to attach the confidence or scoring values for automatically created annotations, retrieved from the respective tools. We argue that the difficulty for automatic processing presented by non-standard data makes it particularly valuable to not only look at annotations in isolation when analyzing, but also at the relative confidence with which the respective tools made those predictions. By making this information available as additional (meta-)annotation layers in our corpus it can directly be used in exploration tools such as ICARUS (Gärtner et al., 2013) for investigation together with regular linguistic features.

[5]No additional identity related data is included.

3.2 Automatic Processing

This section gives a (non-exhaustive) overview of the systems used for automatic annotations in the silver standard collection.

For processing of text resources we mainly employ pipeline systems covering multiple annotation layers, e.g.: 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), FSPar (Schiehlen, 2003), TreeTagger (Schmid, 1994). Table 1 shows which annotation layers are covered by those systems.

In addition the IMS HotCoref DE system by Roesiger and Kuhn (2016) is used for German text to create automatic coreference annotations.

System	Syntax	Lemma	PoS	Morph.
BitPar	C		+	+
ISC	C+D		+	+
Mate	D	+	+	+
IMSTrans	D			
FSPar	D	+	+	+
TT		+	+	

Table 1: List of systems planned to be used for text processing and the annotation layers they cover (C: constituency, D: dependency, ISC: IMS-SZEGED-CIS, TT: TreeTagger).

Our pipeline for speech resources is very similar to the one applied by Schweitzer and Lewandowski (2013) for the GECO corpus. It uses IMS Festival Morphology [6] and IMS Aligner (Rapp, 1995) to

[6]http://hdl.handle.net/11022/1007-0000-0000-8E71-1

produce various annotations on the segment, syllable and word level. We further include an approximation of the F_0 contour using PaIntE (Möhler, 1998; Möhler, 2001) and on top of this categorical prosody labels (e.g. following GToBI(S) by Mayer (1995)) predicted automatically (Schweitzer, 2010; Schweitzer and Möbius, 2009).

3.3 Evaluation and Quality

To obtain meaningful confidence estimations for automatic annotations we employ different strategies. For local (i.e. within one and the same annotation layer) inconsistencies detection is facilitated using the approach developed by DECCA (Boyd et al., 2008). An implementation of their idea for part-of-speech and dependency syntax annotations with an interactive visual front-end exists in one of the plugins (Thiele et al., 2014) for ICARUS.

Taking information from multiple annotation layers into account, we can exploit various redundancies. In-level (or horizontal) redundancy constitutes for example the output of different tools for the same annotation type. It can be used to produce confidence statements based on the agreement of those tools as shown by Haselbach et al. (2012) for parser outputs. A pilot study for the web data part (George, 2016) used a token-based comparison of the output from three parsers with respect to the aspects *head*, *label*, and the combination of both (cf. also the "disagree" method from Smith and Dickinson (2014)). Confidence was derived from the number of parsers that agreed for a specific token and mapped to a respective color scheme.

Cross-level (or vertical) redundancy on the other hand exists when multiple annotation layers describe aspects that are related. If support or contradiction exists between information from different layers, we can use this to assign tentative confidence or simply mark those data points. Dickinson (2015) refers to this as making use of annotation layer inconsistencies, and gives examples for methods taking part-of-speech, syntactic and semantic information into account. With our spoken data, additional annotation layers can be taken into account, e.g. with respect to syntactic and prosodic phrase recognition.

Conventional evaluation of the tools used for automatic annotations will be performed using small gold subsets, e.g. the one mentioned in 3.1. This provides us with performance information that we can attach to entire annotation layers as metadata.

Note that all these confidence or performance values (including a tool's own confidence estimation) are not meant to be used for some a priori cleaning of the data. Instead they are treated as an annotation layer and act as possible indicators for data points which might be of interest or should be ignored for certain research questions. One can then produce excerpts of the entire data set based on the required level of confidence.

4 "Unnormalization": Including Features of Orality for Text Processing

As discussed in Section 2, an aspect of the available interview transcriptions is the omission of features of orality. While the edited transcript is suited for text processing, it is unfit for the speech processing pipeline, when trying to align text and audio data. For the interviews which are part of the gold standard, we reintroduced some of the omitted features in a way that the result is neither canonical data nor an unsolvable puzzle for one of the processing pipelines. Since a step that produces canonical forms from non-standard data is often referred to as *normalization*, we call this step *unnormalization*.

An important fact is that we consider both types of available primary data (audio and edited transcript) as equal in status. The text files are not seen as 'wrong' transcriptions or annotations, which can just be changed, but as an interesting source in its own right, e.g. for research on typing errors or aspects of edition. Thus, the original primary data is kept and the modified transcripts are created as an additional layer based on the primary data. Furthermore, the decisions made in the original transcription process are taken into account in the process of unnormalization. That is, in cases where several transcriptions are possible and the original transcription is among them, it is kept.

4.1 Process

The unnormalization is similar to processes of normalization and annotation. Guidelines have been defined and each interview is modified by two annotators independently. Adjudication is done by a third person. The guidelines comprise cases of spelling errors; missing, additional or different words; word order; repairs; repeats; and unrepaired slips of the tongue. Thereby the main principles are: (i) correct and completely heard words should be part of the modified transcript, while (ii) the transcript is changed as little as possible, such that

the decisions of the transcriber are still reflected. The results include all fully spoken words (including repetitions) in the original word order from the audio file. This is helpful for the aligner but introduces non-standard features for the text processing. On the other hand, the modified transcript does not include any fillers or words that have been uttered only partially, which would pose a vocabulary problem for the text processing, but this way the result provides still no optimal representation for the aligner. Example (1) shows a case where a word and a filler from an utterance in the audio file were not included in the transcript (2)[7]. In our process of unnormalization, the filler was also ignored, but the completely heard word *in* was added (3).

(1) obwohl die [...] in vielfach **äh** günstiger sind als
although they in several times ehm cheaper than
'although [...] they have become (in) several times cheaper than'

(2) obwohl die [...] vielfach günstiger sind als

(3) obwohl die [...] in vielfach günstiger sind als

4.2 Quantification

To give a rough quantification, that unnormalization is a first step to non-standard data for text processing, we apply a method from Faaß and Eckart (2013). They use the robust rule-based dependency parser FSPar (Schiehlen, 2003), which includes all input tokens into a dependency graph, but attaches parts which could not be properly embedded to the artificial root node. Based on the number of these attachments and the number of tokens in the sentence, Faaß and Eckart (2013) compute an *error rate* and exclude sentences from a web corpus which are considered less processable.

For our small study we parsed 10 transcripts and their manually modified counterparts with FSPar. Since FSPar comes with its own pre-processing pipeline we leave sentence border detection to this pipeline and only mark each speaker turn as its own text.[8] Table 2 shows the results of the comparison between the original and modified transcripts. The error rate increases slightly for the modified transcript. Thus, the parser encountered more tokens it could not attach properly to the dependency graph in the transcripts after the unnormalization step, i.e. the data became a bit more non-canonical for the parser. Still, we are far away from the sentences being hardly parsable at all, which is due to the official interview situation where at least one of the participants is a professional speaker from the radio station.[9]

	orig. transcript	mod. transcript
error rate	0.157	0.163

Table 2: Processability values based on FSPar.

5 Conclusions and Future Work

We presented our approach to create a large silver standard collection of non-standard data. In particular we discussed one ongoing annotation project for German radio interviews and their written transcripts. With the size of resources involved making an exhaustive manual annotation impossible we instead use existing tools to create (concurrent) annotations on various linguistic levels. To gain indicators for annotation quality we estimate confidence values for individual annotations or entire layers based on consistency checks or redundancy along horizontal and vertical annotation axes. This places the silver standard somewhere between true gold standards and raw automatic annotations in terms of quality. Data from the SFB732 silver standard collection will be made available for research purposes, along with CMDI[10] metadata and a persistent identifier for each release.

For the future we also plan to further raise awareness regarding the integration of a tool's own confidence estimation in its output. This is to motivate developers of both processing tools and data formats to consider those meta-annotations in their work, as well as to encourage their usage in research and development. We are aware of the current lack of standardization or comparability for this type of annotation, and therefore will investigate sensible ways of normalization to make confidence annotations a valuable part of NLP data.

Acknowledgments

This work was funded by the German Research Foundation (DFG) via the SFB 732, project INF.

[7] The subject is renewable energy and the larger context gives a strong indication for this reading.

[8] This decision is debatable, since a sentence might be continued by another speaker, and overlap might occur at speaker turns.

[9] Faaß and Eckart (2013) deleted sentences with an error rate above 0.7, however in a corpus with tables, etc. from raw web data.

[10] Component Metadata Infrastructure from CLARIN, https://www.clarin.eu/

References

Stefan Baumann and Arndt Riester. 2012. Referential and Lexical Givenness: semantic, prosodic and cognitive aspects. In Gorka Elordieta and Pilar Prieto, editors, *Prosody and Meaning*, number 25 in Interface Explorations. Mouton de Gruyter, Berlin.

Anders Björkelund and Joakim Nivre. 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.

Anders Björkelund, Ozlem Cetinoglu, Richárd Farkas, Thomas Mueller, and Wolfgang Seeker. 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.

Anders Björkelund, Agnieszka Faleńska, Wolfgang Seeker, and Jonas Kuhn. 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.

Bernd Bohnet and Joakim Nivre. 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.

Bernd Bohnet. 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.

Adriane Boyd, Markus Dickinson, and W.Detmar Meurers. 2008. On detecting errors in dependency treebanks. *Research on Language and Computation*, 6(2):113–137.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther Knig, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Markus Dickinson. 2015. Detection of annotation errors in corpora. *Language and Linguistics Compass*, 9(3):119–138. LNCO-0526.R1.

Stefanie Dipper, Anke Lüdeling, and Marc Reznicek. 2013. NoSta-D: A corpus of German non-standard varieties. In Marcos Zampieri and Sascha Diwersy, editors, *Non-standard Data Sources in Corpus-based Research*, pages 69–76. Shaker.

Gertrud Faaß and Kerstin Eckart. 2013. Sdewac a corpus of parsable sentences from the web. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer Berlin Heidelberg.

Markus Gärtner, Gregor Thiele, Wolfgang Seeker, Anders Björkelund, and Jonas Kuhn. 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.

Tanja George. 2016. Confidence estimation for automatic parsing of large web data sets. Masterarbeit, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Boris Haselbach, Kerstin Eckart, Wolfgang Seeker, Kurt Eberle, and Ulrich Heid. 2012. Approximating theoretical linguistics classification in real data: the case of German "nach" particle verbs. In *Proceedings of COLING 2012*, pages 1113–1128, Mumbai. The COLING 2012 Organizing Committee.

Hagen Hirschmann, Seanna Doolittle, and Anke Lüdeling. 2007. Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*, Birmingham.

Natalie Lewandowski. 2013. Phonetic convergence and individual differences in non-native dialogs. Abstract presented at the New Sounds Conference in Montréal.

Jörg Mayer. 1995. Transcription of German Intonation. The Stuttgart System. ms.

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

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

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, 59.

Stefan Rapp. 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).

Dietrich Rebholz-Schuhmann, Antonio Jos Jimeno-Yepes, Erik M. van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger, and Udo Hahn. 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), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ina Roesiger and Jonas Kuhn. 2016. Ims hotcoref de: A data-driven co-reference resolver for german. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Michael Schiehlen. 2003. A cascaded finite-state parser for German. In *Proceedings of EACL 2003*, pages 163–166, Budapest.

Anne Schiller, Simone Teufel, Christine Stckert, and Christine Thielen. 1999. Guidelines fr das Tagging deutscher Textcorpora mit STTS.

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

Helmut Schmid. 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.

Helmut Schmid. 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.

Antje Schweitzer and Natalie Lewandowski. 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.

Antje Schweitzer and Bernd Möbius. 2009. Experiments on automatic prosodic labeling. In *Proceedings of Interspeech 2009*, pages 2515–2518.

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

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109, Dublin, Ireland, August. Dublin City University.

Wolfgang Seeker. 2016. Guidelines for the Annotation of Syntactic Structure in the IMS Interview Corpus.

Amber Smith and Markus Dickinson. 2014. Evaluating parse error detection across varied conditions. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13*, pages 230–241, Tübingen, Germany.

Ann K. Syrdal, Julia Hirschberg, Julie McGory, and Mary Beckman. 2001. Automatic tobi prediction and alignment to speed manual labeling of prosody. *Speech Commun.*, 33(1-2):135–151, January.

Gregor Thiele, Wolfgang Seeker, Markus Gärtner, Anders Björkelund, and Jonas Kuhn. 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.

Sina Zarrieß, Florian Schäfer, and Sabine Schulte im Walde. 2013. Passives of reflexives: a corpus study. Abstract at LinguisticEvidence – Berlin Special.