University of Stuttgart
Germany

GRAIN

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

GRAIN Corpus

- collection of German radio interviews
- brings together state-of-the-art tools from *text* and *speech* processing
- non-canonical for text-based tools: spontaneous speech with features of orality, but experienced public speakers

Primary Data

- German radio interviews, just under 10 minutes each (mp3)
- their edited transcripts from radio station (pdf or doc)
- 20 interviews chosen for gold-standard annotations
- remainder of the interviews represent silver-standard part
- non-static resource: collection continuously grows as radio station releases more interviews
- current corpus size: 140 interviews, about 221,00 word tokens and about 23 hours of audio

The Silver Standard Idea

Provide annotation quality better than unchecked automatic annotation (though it might not reach gold-standard), cf. [Rebholz-Schuhmann et al., 2010].

- combining annotations for the same layer
- adding confidence estimations as additional annotation layer

→ can be used in visualization and search

Useful for:

- gauging the quality of (subsets of) the data
- selecting subsets with high confidence estimations
- find areas of interest (with low confidence)

Documentation & Availability

- thorough workflow documentation
- various annotation formats
- published in the CLARIN framework under persistent identifier `http://hdl.handle.net/11022/1007-0000-0007-C632-1`

Overview of Annotations (Silver vs. Gold Part)

Annotation	Tool/Guidelines
Character Anchors	intern
Tokenization and PoS	TreeTagger [Schmid, 1994]
Sentence Segmentation	intern
Phones	Aligner [Rapp, 1995]
Syllables	Aligner [Rapp, 1995]
Words	Aligner [Rapp, 1995]
Painte-based prediction of GToBI pitch accents & boundary types	Prosody Recognition [Schweitzer, 2010]
Pitch accent placement	CNN-based accent detector [Stehwien and Vu, 2017]
PoS Tagging	Stanford Tagger [Toutanova et al., 2003]
Dependency parses	IMS-SZEGED-CIS [Björkelund et al., 2013]
Dependency parses	Mate [Bohnet and Nivre, 2012]
Dependency parses	IMSTrans [Björkelund and Nivre, 2015]
Dependency parses	Stanford Parser [Chen and Manning, 2014]
Dependency parses – merged	intern [Sagae and Lavie, 2006]
Constituency parses	BitPar [Schmid, 2006]
Constituency parses	IMS-SZEGED-CIS [Björkelund et al., 2013]
Constituency parses	Stanford Parser [Klein and Manning, 2003]
Unnormalization	Manual [Eckart and Gärtner, 2016]
PoS Tagging	Manual [Schiller et al., 1999, Seeker, 2016]
Referential information status	Manual [Riester and Baumann, 2017]
Information structure	Manual [Riester et al., 2018]
Questions under Discussion	Manual [Riester et al., 2018]
QUD trees (Discourse structure)	Manual [Riester et al., 2018]

Annotator Agreement

Task	Agreement
PoS Tagging	Cohen's κ of 0.97
Referential information status	Cohen's κ of 0.75

Annotations Under Development

- CNN-based prediction of intonation boundary placement
- unedited orthographic transcripts (preserving all features of orality)

Morphosyntactic Layers

Tool	Lemma	PoS	Morph	D-Syn	C-Syn
ImsTrans				x	
Mate	x	x	x	x	
BitPar		x			x
ISC	x	x	x	x	x
Stanford		x		x	x
TreeTagger	x	x			

References

[Björkelund et al., 2013] 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. Association for Computational Linguistics.

[Björkelund and Nivre, 2015] 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. Association for Computational Linguistics.

[Bohnet and Nivre, 2012] 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. Association for Computational Linguistics.

[Chen and Manning, 2014] Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In Moschitti, A., Pang, B., and Daelemans, W., editors, *EMNLP*, pages 740–750. ACL.

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

[Klein and Manning, 2003] Klein, D. and Manning, C. D. (2003). Fast exact inference with a factored model for natural language parsing. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 3–10. MIT Press.

[Rapp, 1995] 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 et al., 2010] 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 Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[Riester and Baumann, 2017] 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 et al., 2018] Riester, A., Brunetti, L., and De Kuthy, K. (2018). Annotation guidelines for Questions under Discussion and information structure. In Adamou, E., Haude, K., and Vanhove, M., editors, *Information Structure in Lesser-Described Languages: Studies in Syntax and Prosody*. Benjamins, Amsterdam. Manuscript under submission.

[Sagae and Lavie, 2006] 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. Association for Computational Linguistics.

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

[Schmid, 1994] 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, 2006] 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. Association for Computational Linguistics.

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

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

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

[Toutanova et al., 2003] Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

I M S

www.ims.uni-stuttgart.de

This work was funded by the German research foundation (DFG) via SFB 732, projects INF, A4, A6, A8 and D8.