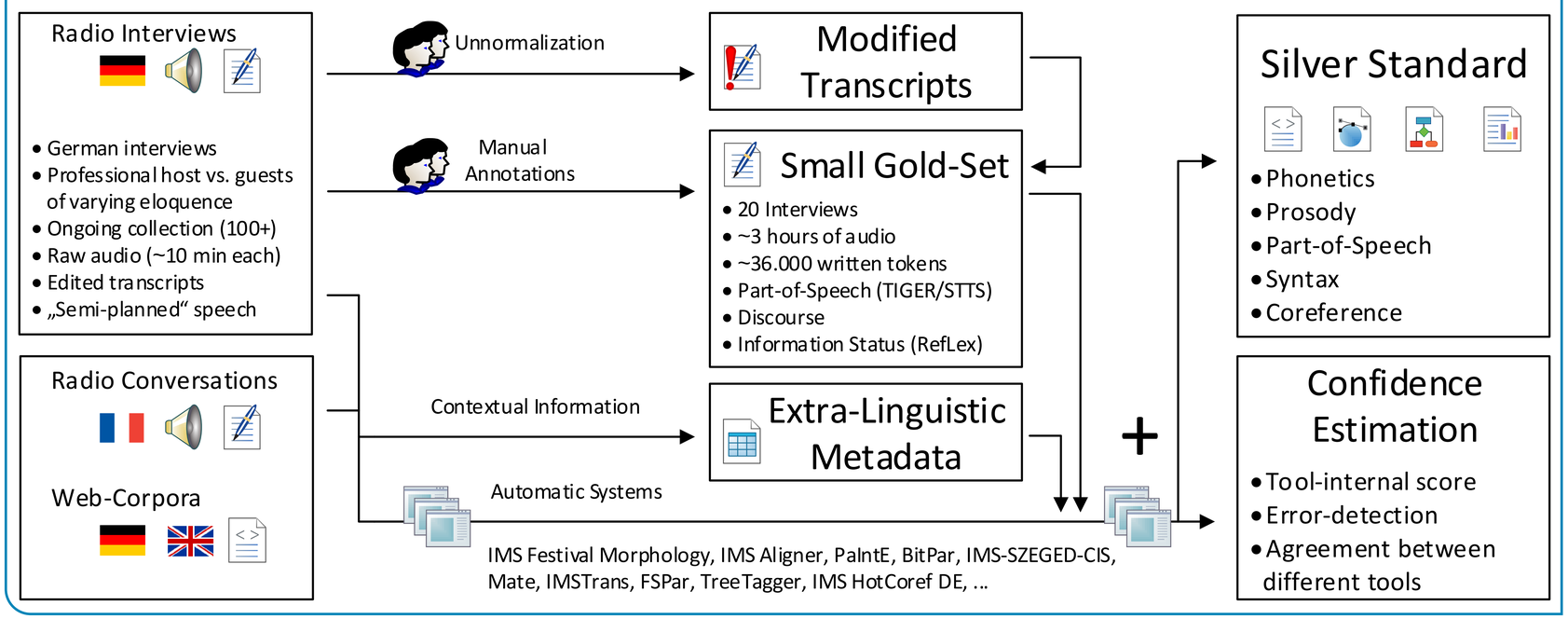
Creating Silver Standard Annotations for a Corpus of Non-Standard Data

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

Annotation Workflow



Data

- Non-static collection of German radio interviews (currently 100 interviews of about 10 minutes each, starting from May 2014)
- Multi-modal data: audio recordings and written transcripts
- Both types of available primary data are equal in status: transcript reflects decisions of transcriber
- Small manually annotated gold-standard set
  - additional manual unnormalization step for transcripts to further move towards non-canonicality
  - 20 interviews totalling about 3 hours and 36.000 written tokens
  - balanced for gender of host and guest and the guest's respective political role
  - annotated for part-of-speech with TIGER/STTS guidelines by Brants et al. (2004) and Schiller et al. (1999) with additions by Seeker (2016), information status according to RefLex scheme by Baumann and Riester (2012), and discourse.
- Canonicality of the data defined by its processability (Petrov and McDonald, 2012)
- Interviews represent a setup between planned/read speech (laboratory conditions: canonical data) and spontaneous conversation (non-canonical data)

Confidence Estimation

- Provided as additional (meta-)annotation layers so they can easily be used in exploration tools like ICARUS (Gärtner et al., 2013)
- Exploiting annotation redundancies
  - horizontally (multiple annotations of same type)
  - vertically (related annotations of different types)
- Multiple sources for confidence estimation
  - exposure of tool-internal scores
  - counting relative number of tool-specific fallback decisions, e.g. root attachment of leftover tokens (Faaß and Eckart, 2013)
  - error or inconsistency detection (DECCA, Boyd et al., 2008)
  - agreement scores between multiple tools (e.g. George, 2016)
  - evaluation results on appropriate gold-standard subset
- Ideally at least one confidence estimation per "real" annotation
- No standardization defined for confidence values yet
  - no general interpretation – only tool-internal comparability
  - no comparability for confidence values from different sources
- Confidence information can then be used to create excerpts of the data suitable for a given research question or application

Unnormalization

- Introducing features of orality into the edited transcript
- Process: detailed annotation guidelines, two annotators per transcript, conflict resolution by a thrid person
- Main principles:
  - correct and completely heard words should be part of the modified transcript
  - the transcript is changed as little as possible
- Result: non-canonical data for speech and text processing alike

Examples: audio, transcript, unnormalization

- Insertion of omitted words, but no inclusion of filled pauses
  - obwohl die [...] in vielfach **äh** günstiger sind als
  - obwohl die [...] vielfach günstiger sind als
  - obwohl die [...] in vielfach günstiger sind als
- Insertion of repeats, but no inclusion of false starts
  - teilweise aber nicht für **je**.. für alle Geräte
  - Teilweise, aber nicht für alle Geräte.
  - Teilweise, aber nicht für, für alle Geräte.
- Reversion of syntactic correction
  - Von Hamburg bis Rheinland-Pfalz **gibt es** regelrechte Konjunkturprogramme für den Bau von Holzhäusern, die man später auch für studentische Wohnungen und für anderen Bedarf nutzen kann.
  - Von Hamburg bis Rheinland-Pfalz werden regelrechte Konjunkturprogramme für den Bau von Holzhäusern, die man später auch für studentische Wohnungen, für anderen Bedarf nutzen kann.
- Omission of inserted words:
  - Und ich glaube, oder ich weiß auch
  - Und ich weiß auch

References

Baumann, S. and Riester, A. (2012). Referential and Lexical Givenness: semantic, prosodic and cognitive aspects. In Elordieta, G. and Prieto, P., editors, *Prosody and Meaning*, number 25 in Interface Explorations. Mouton de Gruyter, Berlin.

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

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Faaß, G. and Eckart, K. (2013). Sdewac – a corpus of parsable sentences from the web. In Gurevych, I., Biemann, C., and Zesch, T., editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer Berlin Heidelberg.

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. Association for Computational Linguistics.

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

Petrov, S. and McDonald, R. (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.

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

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

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