Published in: Haugh, Michael/Ruhi, Sukryie/Schmidt, Thomas/Wömer, Kai (eds.): Proceedings of the LREC-12 Workshop on Best Practices for Speech Corpora in Linguistic Research. Istanbul, Turkey, May 2012. - European Language Resources Association (ELRA), 2012. pp. 47-50.

Developing Solutions for Long-Term Archiving of Spoken Language Data at the Institut für Deutsche Sprache

Peter M. Fischer, Andreas Witt
Institut für Deutsche Sprache
R5 6-13, 68161 Mannheim, Germany
peter.fischer@ids-mannheim.de, witt@ids-mannheim.de

Abstract

This document presents ongoing work related to spoken language data within a project that aims to establish a common and unified infrastructure for the sustainable provision of linguistic primary research data at the Institut für Deutsche Sprache (IDS). In furtherance of its mission to "document the German language as it is currently used", the project expects to enable the research community to access a broad empirical base of working material via a single platform. While the goal is to eventually cover all linguistically relevant digital resources of the IDS, including lexicographic information systems such as the IDS German Vocabulary Portal, OWID, written language corpora such as the IDS German Reference Corpus, DeReKo, and spoken language corpora such as the IDS German Speech Corpus for Research and Teaching, FOLK, the work presented here predominantly focuses on the latter type of data, i.e. speech corpora. Within this context, the present document pictures the project's contributions to the development of standards and best practice guidelines concerning data storage, process documentation and legal issues for the sustainable preservation and long-term accessibility of primary linguistic research data.

Keywords: Best-Practice, Long-Term Archiving, Spoken Language Data

1. Introduction

This document presents ongoing work related to spoken language data within the project called *Zentrum für germanistische Forschungsprimärdaten* (Center for Primary Research Data in German Linguistics), funded by Germany's largest funding organization, the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) and based at Germany's principal research facility in modern German linguistics, the *Institut für Deutsche Sprache* (IDS, German Language Research Institute), in Mannheim, Germany.

In furtherance of IDS's mission to "document the German language as it is currently used", this project aims to establish a common and unified infrastructure for the sustainable provision of germane primary research data, thus to enable the research community to access a broad empirical base of working material via a single platform. While the goal is to eventually cover all linguistically relevant digital resources, including lexicographic information systems such as the IDS German Vocabulary Portal, OWID, written language corpora such as the IDS German Reference Corpus, DeReKo, and spoken language corpora such as the IDS German Speech Corpus for Research and Teaching, FOLK, the work presented here predominantly focuses on the latter type of data, i.e. speech corpora. Within this context, the present document pictures the project's contributions to the development of standards and best practice guidelines concerning data storage, process documentation and legal issues for the sustainable preservation and long-term accessibility of primary linguistic research data.

2. Spoken Language Data at the IDS

IDS maintains a variety of spoken language data resources, painstakingly collected, processed, archived and published by the Department of Pragmatics over the last few decades. Concomitantly, IDS has developed and continuously refined strategies and techniques both to optimize, facilitate the work on the data involved and to ensure appropriate and sustaining access to these resources, resulting in high proficiency in storage matters, documentation tasks and legal issues. Today, separate projects carry out different tasks of this elaborated workflow. Generally speaking, the project *Archiv für Gesprochenes Deutsch* (AGD, German Spoken Language Archive) is the first point of contact when it comes to general processing and principal archiving of corpora and all kinds of files related to them. It is also responsible for rendering them serviceable enabling the *Datenbank für Gesprochenes Deutsch* (DGD, German Spoken Language Database) to provide access to selected parts of the archive through an interactive web-based interface (cf. Fiehler and Wagener, 2005).

2.1 The archive project AGD

The AGD maintains a constantly growing portfolio of German speech data, currently comprising 44 corpora containing approx. 6,000,000 tokens and 6,000 hours of audio recordings. They include collections of numerous German dialects, colloquial and standard language and different types of conversations, which were acquired and processed by various internal and external research projects. To this end, the ADG continually accepts data from institutions, external donor projects and individual researchers who conduct language surveys or otherwise contribute spoken language data.

2.2 The database project DGD

A part of the data managed by the AGD is accessible through an interactive web-based interface maintained by the DGD. The provided content consists of manifold data collections such as recordings, documentations or aligned transcriptions. However, the corpus data accessible through DGD are limited and restricted due to their original user agreements, mostly due to the privacy rights of individuals in the recordings.

In the course of recent modernization efforts a new DGD release as German National Speech Corpus is scheduled for 2012 (cf. Deppermann and Hartung, 2011) and will be accompanied by a significant increase of accessible data, among others as part of the FOLK corpus which is included in the set of corpora available via the DGD interface.

2.3 The FOLK corpus

The *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (FOLK, German Speech Corpus for Research and Teaching; cf. Deppermann and Hartung, 2011) strives to document the German language as it is spoken today by giving insight into the reality of social communication in present-day Germany and other German-speaking regions, on the basis of a wide-ranged yet balanced collection of speech data from different areas of social life such as work, leisure, education and media. As a reference corpus for the spoken language, it aims to render the recurrent process of data acquisition in many linguistic and discourse analysis research projects unnecessary, and to provide illustrative examples of today's spoken German for academic education and language teaching performances. Hence, FOLK seeks to discover new opportunities in the fields of research and teaching.

The available resources comprise not merely the recordings but also their respective textual transcriptions as well as fine-grained alignments between transcriptions and their recording. These additional resources are created with the transcription editor FOLKER (cf. Schmidt and Schütte, 2010) following the GAT 2 transcription conventions (cf. Selting et al., 2009) that specify a modified orthography providing a blend of both pronunciation resemblance and proper (retrievable) spelling, along with rules for typical conversational phenomena such as pauses, breathing, coughing, laughing, etc. and the handling of uncertain or incomprehensible passages. Extensive metadata is also available for all resources, including the conversational circumstances and socio-demographic data on the speakers.

3. Long-Term Archiving and Best Practice

The main goal of this project is the establishment of a common and unified infrastructure for the sustainable provision of primary research data at IDS. In order to achieve this, many obstacles on different levels have to be overcome. This section of the document addresses these levels individually discussing their challenges and sketches the project's current progress in developing a solution which, since the project has just launched in December 2011, is often merely an outline albeit always with a clear line of approach. It should be noted that, although the solutions given here are narrowed down to spoken language resources, in particular to those introduced in the preceding section, however, the issues themselves in general hold equally for other kinds of language resources and presumably for other sustainability-oriented infrastructures as well.

The ground strategy for installing a long-term archiving environment is fourfold:

- Workflow models need to be developed that define concrete practical measures on how to prepare and preprocess resources archived in a "living" storage solution like the AGD, in order to appropriately channel them into the archives of a long-term storage solution like the one to be developed by this project. This particularly involves much-debated issues on the definition of adequate comprehensive and stable formats (cf. Witt, 1998; Rehm et al., 2010; Schmidt, 2011).

- Various strategies regarding the archive's accessibility must be developed, among others to guarantee that legal usage regulations are met which is a pervasive concern to speech corpora (as recordings usually raise privacy issues) and to tackle the findability problem which primary research data collections commonly used to face. The latter is closely associated with a standardized, consistent and reliable referencing system for resources.

- After technically setting up a permanent, reliable and secure storage infrastructure, the repository may then outgrow its simple data-serving functionality by particularly supporting input and output strategies as developed in the two preceding points, respectively.

- In order to sustain the environment's longevity, one must ensure that the technical systems are continuously adapted to ongoing developments and that the devised workflows become firmly established. As for the latter, intensive contact with users applying these guidelines in their own environments and feeding their experiences back to the community immediately supports the long-term effort to streamline best-practice procedures.

The following sections will expand upon the aforementioned points covering the development of best-practice guidelines.

3.1 Standards for Primary Data

As long-term archives comprise more than just their underlying repository, by committing primary research data to such an environment, users expect additional

features such as searching the data, referencing the findings, defining and accessing subsets of them or means for discovering unknown data, alongside their albeit permanent, reliable and secure but plain storage. Yet, the ability of an archive to fully support such features heavily depends on the interpretability of the data committed.

This project therefore encourages the utilization of standard data types and formats in the process of acquisition or preparation of the primary data by developing and defining a list of low-level but obligatory requirements to be met when committing data to the archive. Figuratively speaking, this can be understood as an admission ticket certifying the archivability of a resource.

3.2 Standards for Metadata

In principle, the requirement for standards for primary data holds equally for metadata, except that the range of reasonable metadata, especially in rich-metadata environments as with spoken language recordings, is entirely too broad, in order to have a comprehensive yet expedient cover of standards for it. As a consequence, a sufficiently dynamic metadata schema with a small fixed core set of obligatory items is the preferred approach. The Component Metadata Infrastructure CMDI offers possibilities to implement such a dynamic schema (cf. Broeder et al., 2011).

The aim of the project here is to define such a minimal core set of descriptors that have pan-corpus relevance along with their sets of corresponding possible values. This is important in order to render the metadata harvestable and thus the resource searchable. This is supported by the fact that the CMDI specifications fully comply with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

3.3 Identifying Resources

In 2011, the International Organization for Standardization (ISO) adopted an international standard for persistent reference to electronic language resources (ISO 24619) prepared by their Technical Committee TC37 (Terminology and other language and content resources), Subcommittee SC4 (Language resource management). It introduces a method to uniquely and persistently identify language resources by issuing persistent identifiers (PI) that can be assigned (once) to single resources rendering them referenceable (cf. Broeder et al., 2007).

This functionality is indispensable in the context of long-term archiving as the preservation of primary research data and their consequent availability necessitate their permanent referenceability and retrievability. Therefore, this project will apply the standard to all data committed to the archive. By doing so, it will have to implement a resolver as part of the repository to properly handle the translation to the location where a resource is actually stored.

3.4 Citing Resources

Another aim in the field of standard procedures for bibliographically cataloguing resources is the introduction of metadata covering the linking between primary data and the publications citing them. It is important for researchers to get acknowledged for compiling primary data, because publications are often the only impact factors. Moreover, it can be useful to find out what research has been conducted on a specific resource. On the other hand, in publications, the information about an empirical base is often given in a very informal way, occasionally encoded within the text. Making this information more explicit is therefore highly desirable.

This project recently entered into cooperation with the project called *Integration von Forschungsdaten und Literatur in den Sozialwissenschaften* (InFoLiS, Integrating Research Data and Literature in the Social Sciences), a fellow project that is based at Germany's largest infrastructure institution for the Social Sciences, GESIS, and aims to automatically detect and cross-link publications.

3.5 Technical awareness

The software system that conceptually underlies such a repository must be considered very carefully, as its sufficiently modular architecture is crucial to achieve technical sustainability. In particular, the encapsulation of storage-related matters from internal business logics and, in turn, from front-end applications prepares for incessant adjustments that inevitably come with technical progress over time.

This project has evaluated some available open-source repository systems that (sometimes in combination and to their respective degree) meet the aforementioned requirements. Fully stand-alone developments include OPUS4[1] (licensed under the GNU GPL) and DSpace[2] (shared under a BSD license), whereas systems based on Fedora Commons[3] (licensed under a Creative Commons License) include eSciDoc[4] (distributed under the CDDL) and the interaction of the content management platform Drupal[5] with the digital asset management system Islandora[6] (both licensed under the GNU GPL). At this stage, this project favours the latter solution.

3.6 Community awareness

DFG has acknowledged that long-term preservation of primary research data is necessary and important, and is funding a multidisciplinary list of various projects that are engaged in this direction. However, as solutions are being developed, two opposing implementation principles

[1] http://www.kobv.de/opus4
[2] http://www.dspace.org/
[3] http://www.fedora-commons.org/
[4] http://www.escidoc.org/
[5] http://www.drupal.org/
[6] http://www.islandora.ca/

emerge: *in situ* repositories primarily processing subject-specific data with highly specialized applications in their respective fields on the one hand, and fairly interdisciplinary, usually centralized repositories with a focus on catholicity and general interoperability on the other hand.

This project attempts to effect a compromise by implementing a technically centralized storage solution with leaving the control over the data (like structuring, documentation, accessing policies, etc.) as far as possible to the data providers. Also, the repository specializes on linguistic research data to its most general extent, as one goal of the project is to eventually cover all linguistically relevant IDS-internal digital resources, including lexicographic information systems, written language and spoken language corpora alike. To this end, the project has also entered into cooperation with some external partners like the *Hamburger Zentrum für Sprachkorpora* (HZSK, Hamburg Center for Speech Corpora), based at the University of Hamburg, Germany, and the *Institut für Deutsche Sprache und Linguistik* (German Language and Linguistics Research Institute), based at the Humboldt University of Berlin, Germany.

4. Perspective

While proximate development clearly focuses on means to allow for linguistic resources other than speech-related data to be included in the preservation process, in the long term the project is contemplating opening up the platform for third-party participation by providing some kind of upload mechanism in order to enable external partners to have their non-IDS data preserved.

5. References

Broeder, Daan; Declerck, Thierry; Kemps-Snijders, Marc; Keibel, Holger; Kupietz, Marc; Lemnitzer, Lothar; Witt, Andreas; Wittenburg, Peter (2007). Citation of Electronic Resources: Proposal for a new work item in ISO TC37/SC4. ISO TC37/SC4-Document N366. http://www.tc37sc4.org/new_doc/ISO_TC37_SC4_N366_NP_CitER_Annex.pdf.

Broeder, Daan; Schonefeld, Oliver; Trippel, Thorsten; Van Uytvanck, Dieter; Witt, Andreas (2011). A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI). In: Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies, vol. 7. doi: 10.4242/BalisageVol7.Broeder01.

Deppermann, Arnulf; Hartung, Martin (2011). Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des 'Forschungs- und Lehrkorpus Gesprochenes Deutsch' (FOLK) am Institut für Deutsche Sprache (Mannheim). In: Felder, Ekkehard; Müller, Marcus; Vogel, Friedemann (Eds.). Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. Berlin, New York: de Gruyter, pp. 414-450.

Fiehler, Reinhard; Wagener, Peter (2005). Die Datenbank Gesprochenes Deutsch (DGD) – Sammlung, Archivierung und Untersuchung gesprochener Sprache als Aufgaben der Sprachwissenschaft. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion. 6/2005, pp. 136-147. http://www.gespraechsforschung-ozs.de/heft2005/px-fiehler.pdf.

Rehm, Georg; Schonefeld, Oliver; Trippel, Thorsten; Witt, Andreas (2010). Sustainability of Linguistic Resources Revisited. In: Proceedings of the International Symposium on XML for the Long Haul. Issues in the Long-term Preservation of XML. Balisage Series on Markup Technologies, vol. 6. doi:10.4242/Balisage Vol6.Witt01.

Schmidt, Thomas; Schütte, Wilfried (2010). FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction. In: Calzolari, Nicoletta et al. (Eds.). Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta: European Language Resources Association (ELRA), pp. 2091-2096. http://www.lrec-conf.org/proceedings/lrec2010/pdf/18_Paper.pdf.

Schmidt, Thomas (2011). A TEI-based approach to standardising spoken language transcription. In: Journal of the Text Encoding Initiative, 1. http://jtei.revues.org/142

Selting, Margret; Auer, Peter; Barth-Weingarten, Dagmar; Bergmann, Jörg; Bergmann, Pia; Birkner, Karin; Couper-Kuhlen, Elizabeth; Deppermann, Arnulf; Gilles, Peter; Günthner, Susanne; Hartung, Martin; Kern, Friederike; Mertzlufft, Christine; Meyer, Christian; Morek, Miriam; Oberzaucher, Frank; Peters, Jörg; Quasthoff, Uta; Schütte, Wilfried; Stukenbrock, Anja; Uhmann, Susanne (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung (10), pp. 353-402. http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf.

Witt, Andreas (1998). TEI-based XML-Applications: Transcriptions. In: ALLCACH98, Joint Conference of the ALLC and ACH, Debrecen, pp. 170-174.