Published in: Opas-Hänninen, Lisa Lena/Jokelainen, Mikko/ Juuso, Ilkka/Seppänen, Tapio (eds.): Digital Humanities 2008.
Book of Abstracts - Oulu: University of Oulu, 2008. pp. 27-29.

Paper 3: Sustainability of Annotated Resources in Linguistics

Georg Rehm, Andreas Witt, Erhard Hinrichs, Marga Reis

Introduction

In practically all scientific fields the task of ensuring the sustainability of resources, data collections, personal research journals, and databases is an increasingly important topic – linguistics is no exception (Dipper et al., 2006, Trilsbeek and Wittenburg, 2006). We report on ongoing work in a project that is concerned with providing methods, tools, best-practice guidelines, and solutions for *sustainable* linguistic resources. Our overall goal is to make sure that a large and very heterogeneous set of ca. 65 linguistic resources will be accessible, readable, and processible by interested parties such as, for example, other researchers than the ones who originally created said resources, in five, ten, or even 20 years time. In other words, the agency that funded both our project as well as the projects who created the linguistic resources – the German Research Foundation – would like to avoid a situation in which they have to fund yet another project to (re)create a corpus for whose creation they already provided funding in the past, but the "existing" version is no longer available or readable due to a proprietary file format, because it has been locked away in an academic's hidden vault, or the person who developed the annotation format can no longer be asked questions concerning specific details of the custom-built annotation format (Schmidt et al., 2006).

Linguistic Resources: Aspects of Sustainability

There are several text types that linguists work and interact with on a frequent basis, but the most common, by far, are linguistic corpora (Zinsmeister et al., 2007). In addition to rather simple word and sentence collections, empirical sets of grammaticality judgements, and lexical databases, the linguistic resources our sustainability project is primarily confronted with are linguistic corpora that contain either texts or transcribed speech in several languages; they are annotated using several incompatible annotation schemes. We developed XML-based tools to normalise the existing resources into a common approach of representing linguistic data (Wörner et al., 2006, Witt et al., 2007b) and use interconnected OWL ontologies to represent knowledge about the individual annotation schemes used in the original resources (Rehm et al., 2007a).

Currently, the most central aspects of sustainability for linguistic resources are:

- markup languages
- metadata encoding
- legal aspects (Zimmermann and Lehmberg, 2007, Lehmberg et al., 2007a,b, Rehm et al., 2007b,c, Lehmberg et al., 2008),
- querying and search (Rehm et al., 2007a, 2008a, Söhn et al., 2008), and
- best-practice guidelines (see, for example, the general guidelines mentioned by Bird and Simons, 2003).

None of these points are specific to the field of linguistics, the solutions, however, are. This is exemplified by means of two of these aspects.

The use of markup languages for the annotation of linguistic data has been discussed frequently. This topic is also subject to standardisation efforts. A separate ISO Group, ISO TC37 SC4, deals with the standardisation of linguistic annotations.

Our project developed an annotation architecture for linguistic corpora. Today, a linguistic corpus is normally represented by a single XML file. The underlying data structures most often found are either trees or unrestricted graphs. In our approach we transform an original XML file to several XML files, so that each file contains the same textual content. The markup of these files is different. Each file contains annotations which belong to a single annotation layer. A data structure usable to model the result documents is a multi-rooted tree. (Wörner et al., 2006, Witt et al., 2007a,b, Lehmberg and Wörner, 2007).

The specificities of linguistic data also led to activities in the field of metadata encoding and its standardisation. Within our project we developed an approach to handle the complex nature of linguistic metadata (Rehm et al., 2008b) which is based on the metadata encoding scheme described by the TEI. (Burnard and Bauman, 2007). This method of metadata representation splits the metadata into the 5 different levels the primary information belongs to. These levels are: (1) setting, i.e. the situation in which the speech or dialogue took place; (2) raw data, e.g., a book, a piece of paper, an audio or video recording of a conversation etc.; (3) primary data, e.g., transcribed speech, digital texts etc.; (4) annotations, i.e., (linguistic) markup that add information to primary data; and (5) corpus, i.e. a collection of primary data and its annotations.

All of these aspects demonstrate that it is necessary to use field specific as well as generalised methodologies to approach the issue "Sustainability of Linguistic Resources".

References

Bird, S. and Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, 79:557–582.

Burnard, L. and Bauman, S., editors (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium.

Dipper, S., Hinrichs, E., Schmidt, T., Wagner, A., and Witt, A. (2006). Sustainability of Linguistic Resources. In Hinrichs, E., Ide, N., Palmer, M., and Pustejovsky, J., editors, *Proceedings of the LREC 2006 Satellite Workshop Merging and Layering Linguistic Information*, pages 48–54, Genoa, Italy.

Lehmberg, T., Chiarcos, C., Hinrichs, E., Rehm, G., and Witt, A. (2007a). Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System. In Schmidt, S., Siemens, R., Kumar, A., and Unsworth, J., editors, *Digital Humanities 2007*, pages 164–166, Urbana-Champaign, IL, USA. ACH, ALLC, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Lehmberg, T., Chiarcos, C., Rehm, G., and Witt, A. (2007b). Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten. In Rehm, G., Witt, A., and Lemnitzer, L., editors, *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, pages 93–102. Gunter Narr, Tübingen.

Lehmberg, T., Rehm, G., Witt, A., and Zimmermann, F. (2008). Preserving Linguistic Resources: Licensing – Privacy Issues – Mashups. *Library Trends*. In print.

Lehmberg, T. and Wörner, K. (2007). Annotation Standards. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics, Handbücher zur Sprach-und Kommunikationswissenschaft (HSK)*. de Gruyter, Berlin, New York. In press.

Rehm, G., Eckart, R., and Chiarcos, C. (2007a). An OWL-and XQuery-Based Mechanism for the Retrieval of Linguistic Patterns from XML-Corpora. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., and Nikolov, N., editors, *International Conference Recent Advances in Natural Language Processing (RANLP 2007)*, pages 510–514, Borovets, Bulgaria.

Rehm, G., Eckart, R., Chiarcos., C., Dellert, J. (2008a). Ontology-Based XQuery'ing of XML-Encoded Language Resources on Multiple Annotation Layer. In *Proceedings of LREC 2008*, Marrakech, Morocco.

Rehm, G., Schonefeld, O., Witt, A., Lehmberg, T., Chiarcos, C., Bechara, H., Eishold, F., Evang, K., Leshtanska, M., Savkov, A., and Stark, M. (2008b). The Metadata-Database of a Next Generation Sustainability Web-Platform for Language Resources. In *Proceedings of LREC 2008*, Marrakech, Morocco.

Rehm, G., Witt, A., Zinsmeister, H., and Dellert, J. (2007b). Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections. In Schmidt, S., Siemens, R., Kumar, A., and Unsworth, J., editors, *Digital Humanities 2007*, pages 166–170, Urbana-Champaign, IL, USA. ACH, ALLC, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Rehm, G., Witt, A., Zinsmeister, H., and Dellert, J. (2007c). Masking Treebanks for the Free Distribution of Linguistic Resources and Other Applications. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*, number 1 in Northern European Association for Language Technology Proceedings Series, pages 127–138, Bergen, Norway.

Schmidt, T., Chiarcos, C., Lehmberg, T., Rehm, G., Witt, A., and Hinrichs, E. (2006). Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources. In *Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art*, East Lansing, Michigan.

Söhn, J.-P., Zinsmeister, H., and Rehm, G. (2008). Requirements of a User-Friendly, General-Purpose Corpus Query Interface. In *Proceedings of LREC 2008 workshop on Sustainability of Language Resources and Tools for Natural Language Processing*, Marrakech, Morocco.

Trilsbeek, P. and Wittenburg, P. (2006). Archiving Challenges. In Gippert, J., Himmelmann, N. P., and Mosel, U., editors, *Essentials of Language Documentation,* pages 311–335. Mouton de Gruyter, Berlin, New York.

Witt, A., Rehm, G., Lehmberg, T., and Hinrichs, E. (2007a). Mapping Multi-Rooted Trees from a Sustainable Exchange Format to TEI Feature Structures. In *TEI@20: 20 Years of Supporting the Digital Humanities*. The 20th Anniversary Text Encoding Initiative Consortium Members' Meeting, University of Maryland, College Park.

Witt, A., Schonefeld, O., Rehm, G., Khoo, J., and Evang, K. (2007b). On the Lossless Transformation of Single-File, Multi-Layer Annotations into Multi-Rooted Trees. In Usdin, B. T., editor, *Proceedings of Extreme Markup Languages 2007*, Montréal, Canada.

Wörner, K., Witt, A., Rehm, G., and Dipper, S. (2006). Modelling Linguistic Data Structures. In Usdin, B. T., editor, *Proceedings of Extreme Markup Languages 2006*, Montréal, Canada.

Zimmermann, F. and Lehmberg, T. (2007). Language Corpora – Copyright – Data Protection: The Legal Point of View. In Schmidt, S., Siemens, R., Kumar, A., and Unsworth, J., editors, *Digital Humanities 2007*, pages 162–164, Urbana-Champaign, IL, USA. ACH, ALLC, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Zinsmeister, H., Kübler, S., Hinrichs, E., and Witt, A. (2008). Linguistically Annotated Corpora: Quality Assurance, Reusability and Sustainability. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics, HSK*. de Gruyter, Berlin etc. In print.