LREC10-W4

Language Resource and Language Technology Standards – state of the art, emerging needs, and future developments

18th of May, 2010

Linguistic tool development between community practices and technology standards

Thomas Schmidt
SFB 538 'Multilingualism'
Max Brauer-Allee 60
D-22765 Hamburg
E-mail: thomas.schmidt@uni-hamburg.de

Abstract

This contribution addresses the workshop topic of "standardising policies within eHumanities infrastructures". It relates 10 years of experience with language resource standards, gained in the development of EXMARaLDA, a system for the construction and exploitation of spoken language corpora. Section 2 gives an overview of the EXMARaLDA system focussing on its relationship with existing and evolving standards for language resources. Section 3 presents the HIAT system as an example of an established community practice. Section 4 then addresses several issues that where encountered when trying to bring together HIAT, EXMARaLDA and the wider standard world.

1. Introduction

This contribution addresses the workshop topic of "standardising policies within eHumanities infra-structures". It relates 10 years of experience with language resource standards, gained in the development of EXMARaLDA, a system for the construction and exploitation of spoken language corpora.

EXMARaLDA is targeted mainly at an audience of non-technologically oriented linguists who study, for instance, pragmatic aspects of natural interaction, language acquisition in children and adults, dialectal variation, or special forms of multi-lingual interaction like interpreting. While awareness in these different communities about the importance of standards for data exchange and sustainability is growing, there is still a large gap between their own established practices of data processing and high-level standardisation efforts in currently evolving e-infrastructures such as CLARIN. We as tool developers have therefore come to accept a role as a mediator between established community practices on the one hand, and "true" technological standards on the other hand, and it is from this perspective that I will look at language resource standards in this contribution.

The paper is structured as follows: section 2 gives an overview of the EXMARaLDA system focussing on its relationship with existing and evolving standards for language resources. Section 3 presents the HIAT system as an example of an established community practice. Section 4 then addresses several issues that where encountered when trying to bring together HIAT, EXMARaLDA and the wider standard world.

2. Standards in EXMARaLDA

EXMARaLDA, under development since 2000 at the Research Centre on Multilingualism at the University of Hamburg, is a system of data models, formats and tools for the construction and exploitation of spoken language corpora. Its main areas of application are conversation and discourse analysis, research on learner language and dialectology (see Schmidt/Wörner 2009).

2.1 EXMARaLDA data model

EXMARaLDA's data model[1] is an application of the Annotation Graph Formalism (AG, Bird/Liberman 2001). It is represented in two XML-based data formats of different structural complexity:

1. An EXMARaLDA Basic-Transcription is an annotation graph with a single, fully ordered timeline and a partition of annotation labels into a set of tiers (aka the "Single timeline multiple tiers" data model: STMT). It is suitable to represent the temporal structure of transcribed events, as well as their assignment to speakers and to different levels of description (e.g. verbal vs. non-verbal).
2. An EXMARaLDA Segmented-Transcription is an annotation graph with a potentially bifurcating timeline in which the temporal order of some nodes may remain unspecified. It is derived automatically from a Basic-Transcription and adds to it an explicit representation of the linguistic structure of annotations, i.e. it segments temporally motivated annotation labels into units like utterances, words, pauses etc.

A more detailed description of EXMARaLDA's data model can be found in Schmidt 2005.

2.2 Interoperability with ELAN, ANVIL, etc.

Annotation tools like ELAN, ANVIL, Praat etc. work with data models which are very similar to that of an EXMARaLDA Basic-Transcription. Schmidt et al. (2009) discusses the different variants of the STMT data model used by these tools and formulates a suggestion for a an XML exchange format based on the Atlas Interchange Format (Laprun et al. 2002) which ensures that the common denominator information of their data models can be

[1] EXMARaLDA also caters for metadata descriptions, but I will restrict myself in this paper to data models and formats for representing spoken language transcriptions.

exchanged. In practice, EXMARaLDA users can profit from this interoperability by employing different tools for different tasks in their annotation workflows.

2.3 Compatibility with TEI

The principal challenge in establishing compatibility between time-based data models like AG or its different STMT derivatives and more hierarchy-oriented approaches like the TEI's is to find suitable structural units inside a directed acyclic graph (DAG) which can be ordered sequentially and underneath which other structural units of that graph nest in an ordered fashion, thus giving rise to an ordered hierarchy of content objects (OHCO) This problem can probably not be solved generically (i.e. for every conceivable type of data representable in a DAG), but, as argued in Schmidt 2005, mechanisms can be found which are at least applicable across a wider range of data types. EXMARaLDA uses one such mechanism – the combination of temporally contiguous annotation labels assigned to the same speaker – to derive a list-like representation of an annotation document from a Segmented-Transcription. This list can then be represented in an XML document following the TEI guidelines for transcriptions of speech. In terms of interoperability and data exchange, this is especially important because it creates a link between the most common way of representing time-series data (i.e. DAG) and the "natural" way of representing written language (i.e. OHCO).

The same mechanism is also used to establish interoperability between EXMARaLDA and transcription tools built on a more hierarchy-oriented conception of data – most importantly the CLAN editor of the CHILDES system.

2.4 Compatibility with LAF and GENAU

In the practice of spoken corpus construction, the Linguistic Annotation Framework (LAF) has so far not played any important role, if for no other reason than the fact that there is no transcription or annotation tool that uses or directly supports the LAF data model. Work on PAULA and the ANNIS database (Zeldes et al. 2009), however, shows at least that EXMARaLDA data can be integrated into LAF-based frameworks and thus be made accessible for analysis together with other data whose annotation follows the same principle.

Similarly, GENAU and the SPLICR platform (Rehm et al. 2008) have shown – as a proof of concept at least – that EXMARaLDA data can be transformed into data models based on the idea of multiple annotation of identical primary data (Witt 2002).

3. HIAT

HIAT is an acronym of Halbinterpretative Arbeitstranskriptionen ("semi-interpretative working transcriptions"). It is a transcription convention originally developed in the 1970s for the transcription of classroom interaction. The first versions of the system (Ehlich/Rehbein 1976) were designed for transcription with pencil or typewriter and paper. HIAT's main characteristic is the use of so-called Partitur (musical score) notation, i.e. a two-dimensional transcript layout in which speaker overlap and other simultaneous actions can be represented in a natural and intuitive manner.

Not least because editing such musical scores is technically challenging, HIAT was computerized relatively early in the 1990s in the form of two computer programs – HIAT-DOS for DOS (and later Windows) computers, and syncWriter for Macintoshes. Large corpora of classroom discourse, doctor-patient communication and similar interaction types were constructed with the help of these tools. However, standardization and data exchange being a minor concern at the time, these data turned out to be less sustainable than their non-digital predecessors: The data format produced by HIAT-DOS is purely presentation-oriented and thus does not allow any structural transformations based on the actual semantics of the data. Even more problematically, syncWriter uses a largely undocumented binary format, readable and writable by no other application than syncWriter itself. The realisation that data produced by two functionally almost identical tools on two different operating systems could not be exchanged and, moreover, the prospect that large existing bodies of such data might become completely unusable on future technology, raised awareness in the HIAT community for the need for standards and was one of the major motivations for initiating the development of EXMARaLDA.

4. EXMARaLDA and HIAT

As discussed in the previous sections, EXMARaLDA as a system based on and actively supportive of different existing and developing standards for language resources, has increased the potential of transcription data to be exchanged between different applications and to be integrated into more generic frameworks for linguistic data processing. From the point of view of the HIAT community, the major challenge was to adapt the existing data processing practices in such a way that they could be realized inside the EXMARaLDA system. And, conversely, EXMARaLDA's development had to be sensitive to the needs of that community. The following sections therefore discuss how various types of standards and other – more or less conventionalized – practices continue to interact and compete with each other in this assimilation of HIAT and EXMARaLDA.

4.1 Legacy data

One non-negotiable condition for the acceptance of EXMARaLDA by the HIAT community was that it must be able to accommodate the existing bodies of data created with HIAT-DOS and syncWriter. This condition translates into three more specific requirements:

1) The data model and formats must contain the model(s) underlying the legacy data, i.e. every structural relation represented in the legacy data must also be representable in EXMARaLDA. Since musical score transcripts are based on a similar logic as annotation graphs, this requirement was relatively straightforward to fulfil.
2) Wherever the data model or formats stipulate con-

structs that go beyond the legacy data structure, they must still tolerate data that does not (yet) contain (or worse: that deals inconsistently with) such constructs. As an example, take the assignment of stretches of transcription to absolute times in the recording. While it is certainly desirable for EXMARaLDA's data model to contain a construct for this information, neither syncWriter nor HIAT-DOS provide a place for it. In order to be able to efficiently transform legacy data into and inside EXMARaLDA, the system must therefore also be able to process transcriptions *without* temporal alignment[2], and it must also provide the means of adding this information *ex post*. Yet, when new data is produced with the system, it should allow the user to record this kind of information at the same time the actual annotation is entered. Legacy data and new data thus pose competing requirements to the tools.

3) There must be efficient methods for systematically transforming legacy data into the new data model and formats. As the legacy data are known to be deficient in terms of structure and consistency, the expectation is not a fully automatic conversion procedure, but rather a workflow in which manual and automatic processing steps are combined in a maximally efficient manner. For the HIAT legacy data, this workflow consisted in a method for reading out data from the older tools, followed by a couple of semi-automatic methods for correcting structural inconsistencies, followed by several manual steps in which additional information lacking in the original data (like the above-mentioned media alignment) was added.

Of course, on top of these requirements to *enable* legacy data conversion, a further prerequisite was to find the resources to actually *carry* it *out* – a non-trivial requirement given that legacy data conversion (even if supported by adequate tools) is very demanding in terms of man-hours. After several years of work, a number of HIAT legacy corpora have now been fully transformed to EXMARaLDA[3], and further data are in the waiting line. Experience with the data converted so far will hopefully help to speed up future transformations (see Schmidt/Bennöhr 2008 for a more detailed discussion of this aspect).

4.2 Community practices

The HIAT transcription convention is a documented community practice. It gives instructions on what phenomena to describe in an interaction, and on how to describe them. The latter type of instruction is, in principle, a formal one – it picks out certain symbols from the alphabet, assigns them certain semantics inside the transcription, and formulates rules about which combinations of such symbols are permissible and which are not. For instance, one such rule states that descriptions of pauses should have the form "((1,2s))", i.e. a decimal number followed by an 's' between a pair of double round brackets. In EXMARaLDA, the transformation of Basic-Transcriptions into Segmented-Transcriptions relies on these formal regularities as the basis for a finite state parsing of annotation strings (see Schmidt 2005).

However, in times of pencil and paper transcription and also during the early computerized days of HIAT, no mechanism was available (nor was one needed) to actually *check* the "formal correctness" of a given HIAT transcription. Consequently, the formal rules were followed only loosely in practice and different dialects of HIAT developed over the years to accommodate annotation needs not covered by the "official" conventions. When the first legacy corpora had been converted and the formal regularities of HIAT were to be exploited in automatic processing of the data, it therefore soon became apparent that the conventions were in need of a revision. In Rehbein et al. (2004), the formal transcription rules were thus formulated in a more rigid manner (e.g. by providing Unicode codepoints for all symbols), and additional regulations were introduced to ensure a firm basis for automatic processing of the data. Not surprisingly (HIAT being a community practice with a tradition) this change of practice met with some opposition. In the long run, however, the additional processing methods enabled through EXMARaLDA seem to work in favour of an acceptance of the changes. In any case, the modification of the conventions naturally also had an impact on the legacy data conversion described above – the converted data now had to be checked for correctness against the new version of HIAT.

Another change of community practice became necessary in the area of *workflows*. As long as corpora were not made available to a larger audience, and no methods existed to automatically query a larger corpus of transcriptions, analyses were usually carried out by a small number of researchers on a small number of transcripts. If errors or inaccuracies in these transcripts were found, they could be corrected immediately without having to take into account how the change would affect the overall corpus or other people analysing the same data. Also, corpora could grow and be completed according to the analysis needs of a single project.

As Bird/Simons (2002) have pointed out, however, the *immutability* of a resource is an important aspect of its usability once it has been made available to a wider audience. Moreover, techniques like standoff-annotation also usually require certain parts of the data to remain unchanged in order for pointers to remain valid. Last but not least, publishing a resource also means agreeing on a certain date at which no further modifications on its current version are allowed. The new technology and new uses for the old data thus required HIAT users to think about issues like version and quality control, and to develop practicable workflows not only for creating, but also for publishing resources.

4.3 Other tools

When the development of EXMARaLDA started, only

[2] Note that, for instance, Praat or ANVIL cannot deal with such data – they expect the nodes in their DAGs to correspond to some location in a recording.

[3] These corpora are available through http://corpora.exmaralda.org

Praat and CHAT were available as robust editors for creating transcriptions, and these were, at the time, judged inadequate by the HIAT community for their purposes. This situation has changed fundamentally: tools like ELAN, ANVIL (also Praat in its newer versions) now all run stable and each of them offers interesting features that the others don't. As a further change in community practice, the more innovation friendly members of the HIAT community thus began looking for ways of using different tools side-by-side, exploiting their individual strengths, e.g. doing orthographic transcription in EXMARaLDA, gesture analysis in ANVIL or ELAN and phonetic analysis in Praat. The import and export methods described in 2.2 provide the basis for this. However, given that each of the tools employs a data model that is optimized for its own functionality, data exchange between two of such tools is usually not lossless in both directions. As a further aspect of data creation workflows, processing chains involving different tools and the optimal way of combining them had therefore to be considered.

4.4 Standards

Apart from the fact that they are built on general document standards like XML and Unicode and that they implement specific versions of more general frameworks like AG, neither EXMARaLDA nor the data models and formats of other tools mentioned in the previous sections are "standards" in the strict (ISO) sense of the word. The CLARIN Standardisation Action Plan thus does not list them under the heading of "standards", but under "community practices". It seems to me important to note, however, that they are different from a community practice like HIAT (in its pre-EXMARaLDA version at least) insofar as they have an explicit formal specification and technical realisations that actually exploit this formal basis.[4]

From the frameworks listed under "standards" in this document, at least TEI and LAF are potentially relevant for the users of HIAT and EXMARaLDA. As discussed in 2.3 and 2.4, there seem to be no principal obstacles to converting EXMARaLDA to one of these standards. From the point of view of the HIAT user community, however, these standards currently do not play any important role. Their main reason for this is that they do not yet offer any additional value in terms of data processing or interoperability that would be relevant to the researchers' work. When details of conversion methods have to be worked out for these standards, it might get difficult to motivate the community to further changes of their practices as long as this additional value is not clearly visible to them.

5. Conclusions

This paper has sketched some issues encountered on the way from an informal community practice to more general standards for language resources. It has shown that existing bodies of legacy data, existing codifications of community practices and existing workflows, as well as parallel development of different tools all co-determine the standardisation process.

The most important lesson learned in the assimilation of EXMARaLDA and HIAT is that, tedious as the method of carefully and iteratively adapting established practices exemplified here may be, it has turned out to be a reasonably successful standardising policy.

If evolving eHumanities infra-structures want to serve a diverse audience, it may be a key requirement that more such community practices with a potential for standardisation are identified. The development of "generic" standards should then ideally be realised as a stepwise approximation between the concrete practices of specific communities and the high-level abstractions underlying current standardisation efforts in language technology.

[4] The CLARIN document lists CHAT (CHILDES) as a community practice comparable to HIAT insofar as "it is not formally specified as a schema, but a set of widely used tools work on the resources [...]."

6. Acknowledgements

Work on EXMARaLDA is financed by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

7. References

Bird, S. & Liberman, M. (2001). *A formal framework for linguistic annotation.* In: Speech Communication 3, 323-60.

Bird, S. & Simons, G. (2002). *Seven Dimensions of Portability for Language Documentation and Description.* In: Language 79, 557-582.

Laprun, C.; Fiscus, J.; Garofolo, J. & Pajot, S. (2002). *Recent Improvements to the ATLAS Architecture.* Proceedings of HLT 2002, Second International Conference on Human Language Technology, San Francisco, 2002.

Schmidt, T. (2005). *Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech.* In: Arbeiten zur Mehrsprachigkeit, Folge B 62.

Rehbein, J.; Schmidt, T.; Meyer, B.; Watzke, F. & Herkenrath, A. (2004). *Handbuch für das computergestützte Transkribieren nach HIAT.* In: Arbeiten zur Mehrsprachigkeit, Folge B 56.

Rehm, G.; Schonefeld, O.; Witt, A.; Chiarcos, C. & Lehmberg, T. (2008). *SPLICR: A Sustainability Platform for Linguistic Corpora and Resources.* In: Konferenz zur Verarbeitung natürlicher Sprache, September 30–October 02, Berlin, Germany.

Schmidt, T. & Wörner, K. (2009). *EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research.* In: Pragmatics 19.

Schmidt, T.; Duncan, S.; Ehmer, O.; Hoyt, J.; Kipp, M.; Magnusson, M.; Rose, T. & Sloetjes, H. (2009). *An Exchange Format for Multimodal Annotations.* In: Michael Kipp, Jean-Claude Martin, P. P. & Heylen, D. (ed.): Multimodal Corpora, Lecture Notes in Computer Science 207-221. Springer.

Witt, A. (2002). Multiple Informationsstrukturierung mit Auszeichnungssprachen. XML-basierte Methoden und deren Nutzen für die Sprachtechnologie. PhD Thesis, Universität Bielefeld.