Originally published in: Lino, Maria Teresa/Xavier, Maria Francisca/Ferreira, Fátima/Costa, Rute/Silva, Raquel (Eds.):
Proceedings of the fourth international conference on language resources and evaluation (LREC'04). 26 May - 28 May 2004, Lisbon, Portugal.
– Paris: European Language Resources Association (ELRA). Pp. 1213-1216.

Consistent storage of metadata in inference lexica: the MetaLex approach

Thorsten Trippel, Felix Sasaki, Dafydd Gibbon

Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld, Germany
{ttrippel,gibbon}@spectrum.uni-bielefeld.de , felix.sasaki@uni-bielefeld.de

Abstract

With *MetaLex* we introduce a framework for metadata management where information can be inferred from different areas of metadata coding, such as metadata for catalogue descriptions, linguistic levels, or tiers. This is done for consistency and efficiency in metadata recording and applies the same inference techniques that are used for lexical inference. For this purpose we motivate the need for metadata descriptions on all document levels, describe the different structures of metadata, use existing metadata recommendations on different levels of annotations, and show a usecase of metadata inference.

1. Objectives

We introduce a framework which uses similar procedures for metadata management as for the management of structured lexical information using inference techniques over descriptions and structures. Metadata categories such as the general *Dublin Core* (DC, DCMI Usage Board, 2003) set, the *ISLE Metadata Initiative* (IMDI, ISLE Metadata Initiative, 2001) linguistic set, and the *Open Language Archive Community* (OLAC, Simons and Bird, 2002) repository set, are conventionally classified hierarchically and are marked up in XML in ways which are similar to lexicon microstructures.

Metadata are used both to identify a corpus and to describe the corpus itself, and enable access to corpora for many different purposes. Relevant relations between corpora can be described in terms of

similarity , which can be used for the classification, grouping and sorting of (sub-)corpora,

coherence , for example if two resources refer to each other, a common external resource or more general in using the same knowledge or refering to the same contexts.

For the linguistic description of corpora the similarity relations are most relevant, as coherence relations are corpus internal references that are delt with in other areas of linguistic analysis.

2. Metadata for different structures and applications

The application area is distinguished from the data structures. Metadata categories depend on the intended use of a resource or corpus. As the recording of metadata is expensive and too large a number of categories results in confusion and hinders usability, the categories are restricted to areas where a possible use can be assumed.

2.1. Metadata applications

Metadata applications can be divided into two main application areas: first for human use, i.e. allowing human selection of (sub-)corpora, and second for machine use.

We distinguish between the abstract underlying *document structure*, its content or *information structure* and its media rendering or *presentation structure* (Gibbon and Trippel, 2000). The requirments for these areas are different, though the information structure will be similar. For human use the goal is to enable the user to classify the content of a corpus instantneously, therefore the presentation structure is of great importance. For machine applications the document structure is of greater importance, including a standardized document grammar.

2.2. Metadata levels

The description of metadata is crucial for linguistic data storage, and relates to different levels of annotations and corpora.

Traditional metadata standards such as DCMI Usage Board, 2003 and Simons and Bird, 2002 target the level of metadata catalogues, i.e. repositories of general descriptions for corpora. This may be sufficient for cases where one resource has only one set of metadata for all annotation components, for example annotation on a single annotation layer following one standard and for only one recording session. However, linguistically richly annotated corpora for spoken language tend to be of a different nature, where each annotation tier may require a different description, hence different metadata. In addition to this, metadata categories for the description of the individual tiers can differ from the descriptions on the catalogue level.

In one case, top level catalogue metadata may be adequate, while another resource may require top level metadata as well as metadata for each component of the resource, down to individual segments. Human uses of a corpus require at least a classification and description of the entire corpus. Additional metadata on lower levels are more likely to be needed for system use, for instance where they contain format information.

3. Requirements for the inference of metadata

Arbitrary subsets of low level segments, tiers, recording sessions, subcorpora may share metadata. This will clearly involve massive redundancy if the metadata is spelled out in all contexts. It is also often difficult to ensure reliable input of metadata, especially in difficult fieldwork situations and for this reason, too, it is desirable to avoid multiple insertion of the same metadata: uncontrolled redundancy

provokes inconsistency, increases maintenance problems, is time consuming, and expensive.

In order to handle these issues, we advocate the use of a redundancy-reducing and consistency-preserving mechanism such as a type or default inheritance hierarchy for the inference of data from tested premises. Using this inheritance mechanism, lower level metadata are by default inherited from a catalogue level or another level of annotation, and form a well-defined hierarchy of metadata. All of these structures are found in different metadata representation structures.

3.1. Representation structures for metadata

At least three different methods are commonly used for representing metadata; all are related to knowledge representation formalisms:

Attribute Value Structures (AVS): Every metadata category is an attribute which is assigned a value, the content of the metadata.

Functions: An AVS can equivalently expressed as a function. This mechanism is used in knowledge representation for example by KIF, 1998

Trees or dependency hierarchies: hierarchical relations between different metadata categories are modelled in trees, for example the IMDI recommendation (ISLE Metadata Initiative, 2001) or *Text Encoding Initiative* (TEI) metadata header (Sperberg-McQueen and Burnard, 2001) define their data categories in this kind of structure. Nested AVS are equivalent to trees with labelled branches and leaf nodes.

3.2. Inheritance in multilevel corpus metadata

When targeting an individual tier, information that is available on superordinate levels, such as the session level in spoken language corpora, does not need to be repeated on the subordinate structures if it does not deviate. This can for example be true for intellectual property rights, availability, etc. which tend to be the same for all parts of one corpus. For example, in richly annotated spoken language corpora the information can be used for all tiers.

From subordinate structure reverse inference of other information may be required. For example the list of contributors, which is relevant for catalogue level annotations, can be inferred from contributor information for each individual tier or annotation layer.

3.3. Inheritance in structured metadata representations

In some metadata systems, especially when two systems are used in combination, there is or can be a choice of granularity, and some categories may share the values. We use DC and OLAC as an example. First, for corpora in fieldwork situations the DC value for contributor, creator, publisher may refer to the same person and therefore, this can be relevant for a common representation, especially if not only a form of identifier (such as a name) but more detailed information is provided.

Second in the OLAC metadata set the DC category *format* is subdivided into *cpu*, *encoding*, *markup*, *os*, and *sourcecode*; nevertheless the information is present and a query for the metadata category should result in the available information.

4. Multi-tier metadata

In practice, metadata are not necessarily homogeneous. Metadata standards for language data such as DC or TEI (Sperberg-McQueen and Burnard, 2001) describe a resource according to a rather stable metadata set, largely neglecting this issue. Optionality of data categories, and multiple use of the same data categories, sometimes with additional qualifiers, is insufficient to handle the problem of inhomogeneous bits of information for parts of a resource and the specification of some parts for only small portions of the resource. The OLAC portal set solves the problem by using a small core category set.

A first attempt to define metadata for different annotation levels was made by the IMDI working group, distinguishing between catalogue and session metadata. Catalogue metadata refers to the same abstraction level as DC and OLAC, describing the resource as a whole for archives, data repositories and other types of catalogue. The IMDI standard is intended for corpora based on signal data such as audio recordings. These types of recordings may not have been recorded in one single recording session but at different occasions, times and places by different persons. Hence, metadata are also available for each session. It is also possible in the model to treat every annotation tier as well as every textual annotation based on an identical source as separate annotations, obscuring the identity of the shared primary signal source.

To allow for the representation of metadata at every available position in the annotation format, the introduction of metadata needs to be possible on all levels of annotation. In other words, wherever a change of metadata occurs, there has to be a way of marking it. To avoid unwanted redundancies a formalism is needed in which it becomes possible to inherit information between different evels of the annotation.

In computational lexicography, considerable use has been made of highly structured inheritance lexica (Flickinger, Daelemans, Gazdar, Gibbon). As metadata structures are quite similar to lexicon microstructures the use of inheritance-based redundancy-reducing and consistency-preserving mechanisms seems to be a likely candidate for solving the problems addressed here. The structural similarities are rather conspicous: metadata categories, and resemble lexical data categories: the resource to be described corresponds to the lexical lemma.

Following this approach, which we term the *MetaLex Model*, metadata categories are classified and organised hierarchically as follows:

Level-bound metadata: technically connected to a specific level of the corpus, e.g. intellectual property rights (IPR, cf. DC publisher); technical recording information for multimodal data (like recording devices for each session), annotation conventions (like

CoGesT (Trippel et al., 2004) for gesture, ToBI (Silverman et al., 1992) for prosody, GOLD (Farrar and Langendoen, 2003) for morphosyntax), or specific information on specific segments (like character encoding or comments).

Aggregated metadata: inherited piecewise from subordinate structures, e.g. the list of all annotators at catalogue level inherited from individual sessions, tiers or segments.

Claaified metadata: inherited from superordinate structures, e.g. in a fieldwork situation, the person responsible for recording may also be the annotator for all tiers, and copyright holder, etc. In order to cope with inhomogeneities, a default-override strategy is used: if information on the annotator is specified on the catalogue level, the information will also be usable with subordinate units via default inheritance, but local differences override the information.

Metadata information is organised into an implicational hierarchy, for which an XML model is defined. The inheritance mechanism provides inference machinery for using the hierarchy. Depending on the property specifications of each object in the hierarchy, alternative opportunistic mappings to DC, IMDI, and other specifications are selectable. Presently the metadata lexicon is implemented using the DATR formalism (Evans and Gazdar, 1996).

5. Inference of metadata

5.1. Case study: German spoken fairy tale

The German fairy tale *Das Eselein* by Johann and Jacob Grimm was recorded with a semi-professional story teller on audio and video and annotated on 10 levels, including prosody, gesture, word. The whole session has 9 metadata categories, one of them serving as the container for 57 IMDI data categories. Each annotation tier/layer currently has 7 to 15 additional metadata categories, including identification categories (for session identification, layer type, name), application categories (font, etc.), and data warehousing information (annotator description and annotation process description). At present the individual segments have a maximum of 1 technical metadata element.

5.2. Preprocessing

For taking full advantage of the *MetaLex* approach a number of preprocessing steps are needed in order to use the Metadata that is currently embedded in the corpus. These are:

1. Extraction of the metadata into a metadata repository. As the metadata needs to be transformed into a different data format for extensive use of non-XML technology based techniques, the metadata needs to be separated from the corpus itself. In practice this is done with a simple XQuery (Boag et al., 2003) expression.

2. Transformation of metadata into AVS in DATR syntax, using the identifier of the superordinate structure as headword. The reason for using the identifier is to enable the reallocation of metadata with the annotation tier, segment, and subcorpus. As the metadata is supposed to describe the characteristics of the superordinate structure, this structure needs to be referred in the headword; as the whole structure is not intended to be there, the identifier serves this purpose. This step involves transformation of a whole document and is performed with XSLT (XSLT, 1999).

3. Querying the metadata using a DATR inference engine. This allows global inheritance and default overriding, thus serving the purpose[1].

For the conversion into DATR format there are certain problems to be solved that are related to the original DATR syntax and the available format for the metadata structures:

- Special or reserved characters such as angled brackets have a special meaning in DATR.

- Character sets that are non-ASCII are an issue. To enable Unicode processing we internationalized an existing Java implementation of DATR, in order to permit other character sets such as metadata coded in Japanese, Chinese or other non ASCII coding systems.

- The transformation of hierarchical tree structures such as IMDI metadata into DATR format involves *shredding*, in which potential terminal symbols of a tree structure are used to define the arity of a database table as DATR does not use tree structures but directed graphs for interconnections between different hierarchies. This could lead to ambiguities because the substructures of the metadata do not necessarily require identifiers for unambiguous identification if they are determined in the tree context.

For the present proof of concept we have tested values using different glyph systems. The category names are based on ASCII, mostly words from English derived from XML coded source corpora. All metadata categories to be used in inference are unique on all levels. For the future we plan to use a more flexible inference engine in order to allow arbitrary character strings represented in Unicode (cf. McGuinness and van Harmelen, 2004 or related systems).

5.3. Case study: Analysis of coreference on a general and language-specific level

In Sasaki and Witt, 2004 we describe the annotation of a corpus with data from Japanese task-oriented dialogues. The phenomenon under investigation is coreference. Annotations are made on several annotation levels within textual data, making use of annotational categories which are specific to Japanese.

Another focus is the realization of coreference in various languages. For this purpose, a methodology is necessary to describe the relations between language-specific categories which are used to express functions related to

[1] For this purpose the KATR (http://www.cs.uky.edu/∼gstump/katrsite/) implementation was enhanced with Unicode I/O handling and a GUI.

coreference. To accomplish this the MetaLex approach is used.

Coreference is making use of *antecedent expressions* and *coreferential expressions*. These are realized language dependently by various parts of speech. For example in German, pronouns are prototypical coreferential expressions, where in Japanese a similar function is fulfiled by so called *numeral classifiers*.

Specifying *pronouns* and *numeral classifiers* as coreferential expressions, queries on linguistic functions can be generated, e.g. the manner of expressing coreference, without specifying the annotational categories which realize them, i.e. *pronouns* or *numeral classifier*. This is a starting point to specify queries for data from various languages and within various data formats. The query methodology is described in more detail within (Sasaki et al., 2004).

6. Evaluation

The MetaLex model has been evaluated by providing heterogeneous but consistent metadata levels for a number of different corpora, including

1. a Japanese corpus of textual data annotated using primary identical data on more than 17 linguistic levels,
2. a German corpus based on video data annotated on different levels including phonemic, prosodic, grammatic and gesture levels,
3. an Anyi (Ivory Coast) corpus based on audio data annotated on 10 different tier, including syllables, tones, and gloss.

6.1. Sample queries

After bringing the metadata into DATR format, we were able to infer, for example,

- the date of the release of an annotation tier. The release date is available on the session level, and the individual layer does not have this information,
- the author of an annotation tier, which sometimes is the default annotator inferred from the session level and sometimes another person.

7. References

Boag, Scott, Don Chamberlin, Mary F. Fernandez, Daniela Florescu, Jonathan Robie, and Jérôme Siméon, 2003. XQuery 1.0: An XML query language. http://www.w3.org/TR/xquery/. W3C Working Draft 02 May 2003.

DCMI Usage Board, 2003. DCMI metadata terms. URL: http://dublincore.org/documents/2003/ 03/04/dcmi-terms/.

Evans, Roger and Gerald Gazdar, 1996. Datr : A language for lexical knowledge representation. *Computational Linguistics*.

Farrar, Scott and D. Terence Langendoen, 2003. Markup and the gold ontology. In *Workshop on Digitizing and Annotating Text and Field Recordings*. LSA Insitute, Michigan State University. Published at http://saussure.linguistlist.org/ cfdocs/emeld/workshop/2003/langoen-paper.pdf.

Gibbon, Dafydd and Thorsten Trippel, 2000. A multiview hyperlexcion resource for speech and language system development. In *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens.

ISLE Metadata Initiative, 2001. Metadata elements for session descriptions, draft proposal version 2.5. URL: http://www.mpi.nl/world/ISLE/documents/draft/ ISLE_MetaData_2.5.pdf.

KIF, 1998. Knowledge interchange format. http://logic.stanford.edu/kif/.

McGuinness, Deborah L. and Frank van Harmelen, 2004. OWL web ontology language. URL: http://www.w3.org/TR/owl-features/. W3C Recommendation 10 February 2004.

Sasaki, F. and A. Witt, 2004. Co-reference in japanese task-oriented dialogues: A contribution to the development of language-specific and general annotation schemes and resources. In *Proceedings of LREC 2004*. Lisbon.

Sasaki, F., A. Witt, D. Gibbon, and T. Trippel, 2004. Concept-based queries: Combining and reusing linguistic corpus formats and query languages. In *Proceedings of LREC 2004*. Lisbon.

Silverman, K., M. E. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, 1992. Tobi: a standard for labelling english prosody. In *ICSLP-92*, volume 2.

Simons, Gary and Steven Bird, 2002. OLAC metadata. URL: http://www.language-archives.org/OLAC/metadata.html.

Sperberg-McQueen, C. M. and Lou Burnard, 2001. TEI P4 guidelines for electronic text encoding and interchange. URL: http://www.tei-c.org/P4X/index.html.

Trippel, Thorsten, Alexandra Thies, Karin Looks, Ulrike Gut, Jan-Torsten Milde, Benjamin Hell, and Dafydd Gibbon, 2004. Cogest: a formal transcription system for conversational gesture. In *Proceedings of LREC 2004, this volume*. Lisbon.

XSLT, 1999. XSL Transformations (XSLT) version 1.0. URL: http://www.w3.org/TR/xslt.

Acknowledgment

The work presented in this paper was funded mainly by the German Research Council (DFG) grant to the project *Theory and Design of Multimodal Lexica*, Research Group *Text Technological Information Modelling.* Too many colleagues and students have helped with critical feedback following guest lectures and conference presentations to be named here; we are grateful to them all.