Knowledge Management for Small Languages

Huib Verweij, Menzo Windhouwer, and Peter Wittenburg
Max Planck Institute for Psycholinguistics - The Language Archive
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
E-mail(s): Huib.Verwey@mpi.nl, Menzo.Windhouwer@mpi.nl, Peter.Wittenburg@mpi.nl

Abstract. *In this paper an overview of the knowledge components needed for extensive documentation of small languages is given. The Language Archive is striving to offer all these tools to the linguistic community. The major tools in relation to the knowledge components are described. Followed by a discussion on what is currently lacking and possible strategies to move forward.*

Keywords. knowledge management, language resources, annotations, lexica, conceptual spaces, metadata, concept registries.

1. Introduction

Currently we have about 6500 languages that are still spoken in the world, however, each week one is dying, i.e. many of them are highly endangered since they are only spoken by a few elderly speakers [1]. Since knowledge about the world we are living in and our history is encoded in some form in language, we will lose with each language becoming extinct part of our cultural heritage and knowledge about the world, the environmental system, our roots etc. Thus documenting languages embedded in the cultural background it is spoken is already one form of managing knowledge posing many challenges for those who carry out the documentation task. Obviously it is a very important task to store this special knowledge in a suitable form so that future generations can refer back to this knowledge when understanding their roots and when getting an idea about the diversity of language systems once having populated our globe. A recent paper from Levinson and Evans [2] may give an impression about the variety of different language systems nature has created over thousands of years. Archiving the knowledge about our langages could also offer future generations to refer back to proven structures in a time where a blurring of language systems seems to happen everywhere.

There are a couple of relevant initiatives worldwide currently focusing on this challenging task such as DOBES [3], HRELP [4], AILLA [5], E-MELD [6], PARADISEC [7] and ELF [8] to just mention a few important ones. Most of these initiatives are targeting languages spread around the world and in particular in areas with still a large variety of languages spoken. This documentation occurs in a time where we are facing the digital revolution. This revolution allows us to not only imprint a relatively small amount of characters or symbols on walls, clay tablets etc., but allows us to create and store audio and video recordings and even to record time series information of various sort. Of course it is one of our fundamental societal tasks to take measures to preserve such authentic recordings and a variety of derived resources in digital form. Yet we are not sure how this can be done given the enormous technological innovation rate which is affecting hardware, software as well as encoding and structuring standards.

The term "knowledge management" obviously has also a much more technological connotation, since it addresses the questions about which types of knowledge we are speaking and about the best ways of representing this knowledge to foster easy exploitation. Cultural and linguistic knowledge can be covered in

- recordings preserving verbal and non-verbal communicative acts in various culturally relevant situations;
- annotations of various sorts (transcription, translations, morphosyntactic and syntactic analysis, semantics);
- lexica as focal points of covering language knowledge around words, expressions and morphemes;
- conceptual spaces as a matter to preserve conceptualizations and the relations between the relevant concepts of a culture;
- notes of various types describing additional aspects of languages;

- geographical information hooking up language information such as micro-variation to geographical locations;
- digital representations of physical objects;
- metadata information covering contextual information and all sorts of relations between instantiations of the above mentioned object types.

The DOBES program, which started in 2000 covering about 46 teams operating all over the world and documenting about 100 languages, is creating, storing and analyzing the above mentioned object types and making use of state-of-the-art technology in all relevant fields. In doing so it brings the two levels of knowledge management closely together. Regular discussions between the field linguists and the technologists at the Max Planck Institute ensure that the close link is being maintained.

In the following we will first describe the different object types from the linguistic view and then describe the technical solutions found. Finally we will draw some conclusions.

2. Knowledge Components

In this chapter we want to discuss the major knowledge components that are required to document a language.

2.1. Resources and Annotations

Primary audio and video recordings document the languages as they are spoken to communicate between language community members to organize their living. It is the task of the researchers - linguists, ethnologists, musicologists, ethnobiologists, etc. - to select a variety of situations that are typical for the chosen culture to document the usages of the corresponding language. Modern audio and video technology simplifies making recordings even in very complicated fieldwork situations and thus we see a continuous increase in recording hours. Due to the special recording situations the quality of the recording is very much varying, making the application of modern automatic speech and image analysis techniques very difficult. Thus the process of annotation is mainly a manual activity and very time consuming. Often there is even no “standard orthography” for the language in focus and little a priori information making the discovery of the “language system” a difficult puzzle. Often early assumptions about word morphology need to be withdrawn after more material has been studied and additional specific questions have been discussed with the speakers. Thus finding out the units of a language and its rules for combining them is an iterative process of building up first knowledge in the researchers' mind and then making this explicit in a layer of annotations typically covering a transcription, a translation to a major language, a morphosyntactic breakdown, syntactical and semantic descriptions and partly annotations for example describing gesture or prosodic phenomena. Recordings and layers of annotations form bundles of resources that share the same time axis, thus all annotations must be anchored in time periods. Linguistic annotations, however, can form partial hierarchies where partly discontinuous annotations are anchored in symbolic fragments. While the bundles are bound together by contextual information called metadata, the resulting annotation lattices are realized by resource internal pointers to time periods and symbols. Such annotation structures can become utterly complex, if multiple streams (modalities such as speech, intonation, gesture, gaze, etc.) are being annotated in parallel.

2.2. Lexicon Organization

While annotations describe the languages and their system along the utterances as they are found in communications, thus along a time axis, a lexicon extracts constructional units that bear a meaning be it as meaningful morphemes, as words or as more complex expressions such as idioms. Thus the lexicon is the place where the knowledge about the units of a language, their construction and their meaning is stored. The grammar of a language is in so far complementary as it describes how to combine the different units to come to understandable utterances. Storing this complex information about languages which can be so different requires a rather flexible mechanism where we cannot take the Western European languages as basis for designing structural templates. In Wichita, an agglutinative Native American language for example, complex units will form the headword where it is obvious that the begin phoneme used is a marker whether the complex string needs to be interpreted as a verb. For interpreting such a lexicon from a linguistic point of view it is thus of great importance to chop of the prefix, however, for a native speaker the resulting segment is not interpretable anymore. While full-form lexica are often the starting point, abstractions need to be carried out where possible to extract root forms dependent on the language in focus. Different forms are then being aggregated to paradigms.

Meaning is very closely related to units as they occur in the syntactic and semantic contexts, thus lexica without referring to the concrete examples would widely be useless. Thus in addition to their complex internal structure lexica include a variety of pointers to the annotation structures. All in all lexical structures can be very different dependent on the language in focus and they are part of a collection with a dense relation structure. An analysis of several concrete lexica for different languages in the DOBES program revealed the large variation in structural needs [9].

2.3. Conceptual Spaces

Conceptual spaces [10] represent the knowledge type that relates the units of a language that have important cultural relevance. Since each real word is part of a very dense and multi-layered network of relations, all our modern knowledge engineering methods are very poor and utterly reductionist ways of describing the semantic reality [11]. Thus native people are very clear about relevant concepts in their culture such as the "Coconut" and "Fish" for Marquesan and Tuomotuan - islands in the Austronesian archipelago. It is also easy for them to denote related concepts such as "medicin, cooking receipts, housing, clothing, etc." all involving different parts of the coconut tree. However, formalizing the relations, as we are used to when discussing in abstract knowledge engineering terms [12], is an almost impossible undertaking for members of the language communities. Thus we quickly come to a situation where a complex conceptual space is being created since the method as such is very attractive to describe culture specific conceptualization and concept relations, but where the relations are all of the same type "of course these two concepts are related" and where it seems inappropriate to associate reduced relation types. It is a matter of a posteriori curation by researchers to introduce some classification to introduce typing and thus to allow navigating easily in different semantic layers of conceptual spaces.

Concepts can be represented by "words" as indicated, often in multilingual conceptual spaces. Concepts, if they are concrete, are also represented by photos or fragments of photos. If they are represented by "words" they may have a lexical representation as well, if they are concrete they may be represented or explained by audio or video fragments, etc. Thus also conceptual spaces are part of the complex integrated knowledge representing cultures and languages, thus building and maintaining conceptual spaces means maintaining a large amount of relations to all other knowledge elements. Creating such a complex semantic weaving domain is a time consuming task. Until now attempts to facilitate this task with the help of semi-automatic methods failed. One issue is how to elicit new relations to increase the speed of conceptual space building and to help people considering new aspects. Due to completely different conceptualizations, existing semantic knowledge such as incorporated in Wordnet [13] and Cyc [14] cannot be applied. Simply analyzing the entry for "coconut" in Wordnet indicates how useless such attempts will be. Modern statistical methods extracting collocations from corpora will be tuned to small corpora to see in how far they can be used.

2.4. Metadata

Metadata descriptions that describe each object and each collection of objects encode part of the context in which a resource can be interpreted in a useful way. They describe objects with the help of a number of well-defined and widely shared elements. One type of collection certainly is the close relation of all objects sharing the same time axis (recordings and annotations of the same event). Another type of collection will cover all objects that describe a language (and a culture), other collections can include resources of a specific genre or speakers of a certain age, etc. Obviously each object can be member of several collections and the metadata is the glue that bundles all related objects in a structured way.

2.5. Concept Registries

All these knowledge components have a specific structure. Various attempts have been made to standardize these structures. For example, for metadata descriptions, the 15 Dublin Core elements have been very successful. However, in due time it also became clear that forcing everyone to use the same structures lead to misuse of elementary descriptors and thus to, for example, poor metadata. New initiatives take a more flexible approach, where a core meta model can be populated with elementary descriptors taken from a shared registry. The ISO Lexical Markup Framework (LMF; ISO 24613:2008) [15] is such a new initiative. The core LMF meta model consists of a relatively small UML class diagram for a lexical resource. This core model can be extended with additional classes where needed. And to create a project or even language specific model the completed UML diagram is populated with data categories taken from a Data Category Registry (DCR) [16]. The DCR provides a large collection of, sometimes standardized but also often private, data categories. Data categories are elementary descriptors to be used in a linguistic resource, e.g., */languageID/*, */noun/* or */partOfSpeech/*. Each data category has an elaborate specification, containing definitions, examples, value domain, etc. in and for various languages. The idea of shared meta models and shared data categories is currently being expanded to include other registries, i.e., a relation registry to store (ontological) relationships among data categories and other concept systems and a schema registry to store the project or language specific model instances.

This level of knowledge is more on a meta level than the previously discussed knowledge components. While a (small) group of linguists could construct and populate these components for their specific project or language, this level of knowledge aims to document the structures of the resources created and make them semantically interoperable. It is a need to document the use of categories, however, it is not yet clear which category definitions are of a type that is valid for all linguistic systems.

3. Tools Supporting Knowledge Management

In this chapter we want to discuss the tools that have been built and the formats that are being used to create and capture the knowledge as described above.

3.1. ELAN/ANNEX/TROVA

ELAN [17] is the framework that allows users to create any form of exactly aligned annotation as is required to conserve knowledge about the linguistic systems of a large variety of languages. Basis is a meta-schema based on the ideas of Annotation Graphs [18] and in addition allowing symbolic references and hierarchies that has shown in the last decade that it is flexible and powerful enough. Also the user interface has been trimmed to allow efficient manual operation; nevertheless continuously new insights about researchers' workflows need to be incorporated. This flexible schema supports thus the annotation of multi-streams as they occur in multimodal annotations and hierarchical encodings as they often occur on top of transcriptions and gesture phases. The interface allows users to define and store their annotation system (tiers and value sets) and share it with others. Categories can be taken from registries such as ISOcat (a DCR implementation, see section 3.5). Special options serve to include a large variety of time series such as EEG, eye tracking signals, etc.

ANNEX [17] is the web-based variant of ELAN allowing the visualization of the full complexity of annotated media streams as they are stored in an organized archive. TROVA [18] is the search engine that supports queries as they are typical for complex annotations: patterns (which can be regular expressions) from several tiers and at various distances can be combined and the results can be visualized in various forms amongst which are statistics and concordances. TROVA is integrated as well in ELAN as in ANNEX requiring different strategies to build fast indexes.

Since manual annotation is so time-consuming, the AVATech project [19] is devoted to start with building a large number of partly simple detectors by using adaptive statistical methods all creating annotations automatically and to use advanced selection and lattice processing techniques to facilitate linguistic analysis. ELAN can invoke these audio/video recognizers as services.

3.2. LEXUS

LEXUS [17] is a lexicon tool for linguists documenting languages. Linguists can create complex lexica, structuring them to fit the language as described above. It is possible to add recordings, images and video and create links to assets in archives. LEXUS has support for the ISO LMF standard for lexica. To be LMF compliant attributes should be taken from data category registries like ISOcat. To fit with fieldwork reality LEXUS allows stepping away from strict LMF compliance. Also other category registries such as MDF (Multi-Dictionary

Formatter) [20] or user defined attributes can be integrated.

3.3. VICOS

VICOS [17] is the tool that allows users to create conceptual spaces based on lexicalized concepts and typed relations and to navigate in such spaces. Therefore it is closely coupled with LEXUS supporting easy graphical operations. For relation drawing standard types such as "is_part_of" are offered, but users can create their own even allowing creating genealogies. For easy visualization types can be associated with a set of attributes (color, line type). Conceptual spaces can also be used as an attractive semantically defined view on a collection of resources since from every concept arbitrary references can point to content in other knowledge sources such as media fragments, annotations, lexical attributes, metadata elements, etc. The main challenge for such knowledge components is its suitable visualization and the easiness of navigation. Integrating smart new methods will remain a major task.

3.4. IMDI Metadata

Metadata for language resources needs to support structure, categories and terminology that are meaningful for linguists to describe their objects. This was the reason why in 2000 the IMDI metadata infrastructure [21] was designed and why we did not adopt Dublin Core [22]. IMDI is based on a structured XML schema where the chosen elements have fixed contexts. To support flexibility it is possible to add key-value pairs at many places, however, these are specific for the user or the project. IMDI is designed to not only support searches, but to establish organized collections and therefore to support browsing and management. Editors, browsers and search tools were provided.

Also in the metadata domain fixed schema approaches, although easy to handle, were seen as too limiting. By relying on the definitions created by the broad Athens Core group of experts and registered in the metadata profile in ISOcat we have specified a component based metadata infrastructure (CMDI) [23] which is much more flexible allowing researchers to describe resources as it is required.

3.5. ISOcat

ISOcat [24] is an implementation of the ISO 12620 standard (ISO 12620:2009) [25] for a Data Category Registry. And the MPI as the Registration Authority hosts an official instance as the registry for ISO TC 37 "Terminology and other language and content resources". This registry takes a grass roots approach and welcomes everyone. This means that anyone can register and lookup, create, share and submit data categories. Linguists can thus use the system to create specific data categories they require for the resources they are creating. But next to privately owned data categories ISO TC 37 is working towards a standardized core, i.e., this core will consist of a coherent collection of data categories which have been extensively reviewed by international experts. This core set of data categories is actually created by a number of Thematic Domain Groups (TDGs) in which the experts work together to create, review and maintain standardized categories. Individual users or groups of users can submit their new data categories or change requests for existing data categories to these TDGs.

A companion registry for ISOcat called RELcat is currently under construction. This registry will allow storing (ontological) relationships among data categories from ISOcat, but also among data categories and concepts from other registries. The semantic network thus created can be used by search algorithms to broaden the scope of the search by taking semantically nearby data categories and concepts into account. First experiments with these facilities are in preparation at the moment.

4. Conclusions

Language speaking skills are acquired in childhood and for speakers they seem to be a holistic entity allowing communicating. At school we learn that languages can be decomposed into various components obviously supported by many more or less crude classifications and modularization steps. In language documentation we follow this component wise description, but by adding many references at a large variety of places we can easily jump between the fragments in these components creating the illusion of the documentation representing the full complexity of a language.

Technologically this raises a number of questions: (1) Obviously such documentation is very fragile when references are broken to any software or hardware based measure. Therefor the conviction is widely being shared now that systematically persistent identifiers need to be used which are registered explicitly. Where possible we are using Handles [26] for this purpose. (2) We could ask whether there is not ONE integrated knowledge representation mechanism to represent the complexity. Assertions such as RDF triples have been suggested being the most elementary form of describing semantics. However, structure leading to compact representations and easy interpretations is completely given up resulting in a huge heap of semantics difficult to understand and to process. Already the introduction of meta-models - more abstract representations - can form interpretation problems. (3) Also we could ask how to design appropriate software that can be maintained. We have chosen for separate software components that support typical functions and are using programming interfaces to allow users to interact. This restricts the types of possible interactions, but makes software maintenance feasible. This modular structure is supported by using different models to represent the different types of knowledge.

5. References

[1] Crystal D. Language Death. Cambridge: Cambridge University Press; 2000.

[2] Evans N, Levinson SC. The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and Brain Sciences; 2009; 32: 429–492.

[3] Dokumentation Bedrohter Sprachen. mpi.nl/dobes/ [1/23/2011]

[4] The Hans Rausing Endangered Languages Project. hrelp.org [1/23/2011]

[5].The Archive of the Indigenous Languages of Latin America. ailla.utexas.org [1/23/2011]

[6] Electronic Metastructure for Endangered Languages Data. emeld.org [1/23/2011]

[7] Pacific and Regional Archive for Digital Sources in Endangered Cultures. paradisec.org.au [1/23/2011]

[8] The endangered language fund. endangeredlanguagefund.org [1/23/2011]

[9] Wittenburg P, Peters W. Drude S. Analysis of lexical structures from field linguistics and language engineering. Proceedings of the International Workshop on Resources and Tools in Field Linguistics; 2002 May 26-27; Las Palmas, Canary Islands, 2002. mpi.nl/lrec/2002/ [1/23/2011]

[10] Zinn C. Conceptual Spaces in ViCoS. In: Bechhofer S, Hauswirth M, Hoffmann J, Koubarakis M, editors. The Semantic Web: Research and Applications. Springer Berlin / Heidelberg; 2008. p. 890–894.

[11] Cassin B, editor. Le Vocabulaire Européen des Philosophies. Seuil / Le Robert; 2004.

[12] World Wide Web Consortium. OWL Web Ontology. w3.org/standards/techs/owl [1/23/2011]

[13] WordNet a lexical database for English. wordnet.princeton.edu [1/23/2011]

[14] OpenCyc.org. opencyc.org [1/23/2011]

[15] ISO 24613. Language resource management — Lexical markup framework (LMF). Geneva: International Organization for Standardization; 2008.

[16] Kemps-Snijders M, Windhouwer MA, Wittenburg P, Wright SE. ISOcat: Corralling Data Categories in the Wild. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Tapias D, editors. Proceedings of the 6th International Conference on Language Resources and Evaluation; 2008 May 28-30; Marrakech, Morocco. 2008.

[17] The Language Archive. The Language Archiving Technology portal. mpi.nl/lat/ [1/23/2011]

[18] Bird S, Liberman M. A Formal Framework for Linguistic Annotation. MS-CIS-99-01. Penn Engineering; 1999.

[19].Advancing Video Audio Technology in Humanities Research. mpi.nl/avatech/ [1/23/2011]

[20] Multi-Dictionary Formatter. sil.org/computing/shoebox/mdf.html [1/23/2011]

[21] ISLE Meta Data Initiative. mpi.nl/IMDI/ [1/23/2011]

[22] The Dublin Core Metadata Initiative. dublincore.org [1/23/2011]

[23] Component MetaData Infrastructure. clarin.eu/cmdi/ [1/23/2011]

[24] ISOcat. isocat.org [1/23/2011]

[25] ISO 12620. Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources. Geneva: International Organization for Standardization; 2009.

[26] The Handle System. handle.net [1/23/2011]