Constructing Knowledge Spaces From Linguistic Resources

C. Zinn, G. Cablitz, J. Ringersma, M. Kemps-Snijders, P. Wittenburg
Max Planck Institute for Psycholinguistics, University of Kiel
claus.zinn@mpi.nl

Language documentation aims at the creation of a representative and long lasting, multipurpose record of natural languages [1]. It contributes to the maintenance, consolidation and revitalizing of endangered languages and thus safeguards the full range of their uses. Such language documentation also contributes to the description of cultural practices of a speech community. Our aim is to enrich this cultural documentation by allowing users to link linguistic information of lexica and annotated media recordings with ontological information in a multimedia web-based lexicon tool. Our approach is centered around the creation of *knowledge spaces* (KS), where users model a world of concepts and their interrelations for which the organisation of lexical and cultural data is based on categorisation patterns made by the speech community members.

Resources. The DOBES archive for endangered languages hosts a rich set of primary resources (audio and video recordings) and annotations for about 35 languages [2]. For *Marquesan* and *Tuamotuan* trilingual lexica have been created comprising approximately 3000 and 850 lexical entries each. A lexical entry typically contains linguistic information about form, meaning, part of speech, definitions, sample sentences, usage (dialectal, register, etc.), synonyms, antonyms as well as sub-entries of derived forms of the headword. The lexical entry may be linked to multimedia files that display audio- or audio-visual representations of a lexeme. No links to ontological resources for those lexica exist so far.

Main Requirement. In conventional dictionaries relations between words are only cross-referenced within lexical entry articles, whereas ontological resources make the semantic network between headwords visible beyond lexical entry articles; the headword serves as a key to a multi-layered network of semantic relations that combines the linguistic properties of the word with the cultural meaning and usage of the concept [3]. By and large, there are two main user groups: scientists such as linguists and anthropologists, and members of the speech community. While scientists may contribute to and exploit resources to study the language and culture of a community, or compare them to other languages and cultures (etymology, usage of plants, *etc.*), the speech community members shall be motivated to actively participate in describing their language and culture and to *learn* from such resources. For community members, words are keys to access and describe relevant parts of their life and cultural traditions

such as food preparation, house building, medicine, ceremonies, legends, *etc.*. The organisation of relations between words is based on indigenous categorisation alone. The main requirement for the construction of knowledge spaces is thus to enable community members to anchor the words of a linguistic resource according to their classifications, hence creating a kind of "ethno-ontology". It seems clear that existing ontologies (*e.g.*, CYC or SUMO) have only limited relevance here. Their proper and effective use requires considerable expertise or training, and they also induce a significant and usually Westernized bias of how the world should be modeled. A community-based project needs a different approach: simple but effective tools that empower a broad base of users to describe those concepts. As language and culture are interlinked, knowledge engineering tools should interact with language resources; and existing language resources should be used whenever possible for the bootstrapping of ontological resources. In fact, when lexical space and ontological space remain connected with each other, language learning and cultural investigation go hand in hand.

ViCoS. The *knowledge space software* ViCoS [7] aims at providing a simple interface between the lexical space (giving access to lexical resources) and the ontological space (where users relate concepts with each other). ViCoS' user interface is divided into three main areas: a *wordlist view* where each item displays a lexical entry's head category (usually the *lexeme*) in some defined order (usually, lexicographical); a *lex entry view* that displays a configurable view of the lexical entry (including entry points to the multimedia archive); and a *knowledge space view* that depicts a graphical representation of the ontological space. When users click on an item in the wordlist view, its full lexical entry is displayed. Selected parts of the lexical entry can be entered (via drag&drop) into the knowledge space to either highlight its conceptual counterpart, if existing, or create a new concept node, now carrying the label of the item dropped. Similar to a drawing program, nodes can be connected to each other; and after two nodes have been connected the user is prompted to specify the relation type that exists between them. There are predefined relation types for the expression of hyponymy, hypernymy, meronymy, holonymy *etc.*, mirroring those available in the Wordnet [4]. However, users are encouraged to define new relation types.

Discussion. Language documentation requires an active involvement of speech community members. To overcome the limitations of a purely linguistic approach to language documentation, we use knowledge engineering methods that allow members of indigenous communities to play an active role in the documentation process. This emphasizes that a language is so much more than a list of lexical entries based on scientific linguistic descriptions. Our approach turns words into culturally relevant concepts and places them in relation to other con-

cepts. It attempts to engage and inspire community members to explore and to extend the resulting knowledge space. Because our design preserves the relationship between lexical and ontological space, users can browse them more or less simultaneously and can thus gain a richer experience of the language and culture being documented. In a way, our approach bridges scientific resources (lexica constructed by linguists; multimedia assets annotated by experts) with indigenous knowledge resources (KSs constructed by community members).

The challenges that we have to address are threefold: (1) developing a software environment that allows users to easily manipulate KSs with easy mechanisms to create concepts, link them together, and anchor them to existing linguistic and multimedia resources; (2) devise elicitation methods to extract knowledge from community members to help bootstrapping and enriching emerging KSs; and (3) ensuring that the emerging KSs stay easily manageable instead of becoming chaotic and hard to interpret. So far, we can address these three issues as follows. For (1), ViCoS can build on and interface to our existing tools for lexicon management (LEXUS, [5]) and multimedia resources and their description (ANNEX, [6]); For (2), we are anticipating elicitation scenarios that are in line with the current trend in community-based tagging and categorizing of so-called Web2.0 sites (*e.g.* the tagging of our photo, video and sound archive). For (3), one could hope that the users themselves organize the spaces they build. However, we anticipate the need for moderators, probably trained in knowledge engineering, to drive this task. The definition of relation types will be an issue as it is unclear whether the relation types currently suggested will be accepted and properly used by community members. Note however that KSs are built for human consumption rather than any sophisticated machine reasoning; and that humans have little problems in interpreting and exploiting them.

References

[1] Gippert J., Himmelmann N.P. and U. Mosel (eds.), 2006. Essentials of language documentation. Mouton de Gruyter. Berlin.

[2] DOBES: Documentation of endangered languages, *http://www.mpi.nl/dobes*.

[3] Haviland, J.B., 2006. Documenting lexical knowledge. In [1].

[4] Christiane Fellbaum, editor. 1998. WordNet An Electronic Lexical Database. The MIT Press.

[5] Cablitz, G., Ringersma, J., and Kemps-Snijders, M., 2007. Visualizing endangered indigenous languages of French Polynesia with LEXUS. In Proc. of the 11th Int'l Conf. on Information Visualization. IEEE Computer Society.

[6] Berck, P. and Russel, A., 2006. ANNEX - a web-based Framework for Exploiting Annotated Media Resources. In Proc. of LREC.

[7] ViCoS Reference website: http://www.lat-mpi.eu/tools/vicos.