LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 4 (2010), pp. 75-77
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/4469

Reaction to the LEXUS review in the LD&C, Vol.3, No 2

Jacquelijn Ringersma and Marc Kemps-Snijders
Max Planck Institute for Psycholinguistics, Netherlands

We would like to thank Kristina Kotcheva (2009) for her interest in and review of the LEXUS tool. We are pleased that she has brought the tool to the attention of the readers of LD&C and will consider her useful suggestions for improvements on the tool. To supplement Kotcheva's review, we would like to present some additional information useful for potential LEXUS users:

1. IMF and DCR. Lexus is based on ISO TC 37/4 recommendations for (digital) lexical resources. The recommendations, the aim of which is to enhance interoperability, are two-fold: (1) a standard for lexicon structures: Lexical Markup Framework (LMF, ISO DIS 24613:2007) and (2) a Data Category Registry (DCR, ISO 12620:2009).

LMF is an abstract lexical resource model which structures a lexical entry into separated Form and Sense components. The Form component is the holder of information on the form of the entry, either in a list of data categories or into subcomponents. Examples of form related data categories are orthography, phonetic form, etc. Under Sense, categories related to the meaning of the lexical entry can be listed, e.g., definition. More elaborate information on LMF and examples of lexicon structures can be found on www.lexicalmarkupframework.org/.

The DCR is a registry for data categories, i.e., elementary descriptors in a linguistic structure or annotation scheme (www.isocat.org). LEXUS users can consult the DCR and use and refer to data categories to identify the elements used in their linguistic models. By referring to well-defined linguistic concepts, the user can achieve semantic interoperability. An illustrative example of the heterogeneity of concept-naming is the concept for "part-of-speech" and the value set for this data category. In the lexicon data which have been developed in the framework of the DoBeS projects (Documentation of endangered languages, www.mpi.nl/dobes), we find the following variations, e.g., "part-of-speech": [ps], [pos], [part of speech], [part-of-speech], [pos-tag]. For the "part-of-speech" value "noun" we find [n], [N], [noun], [Noun]. For the human eye and cognition all these variations are recognized and understood as being one and the same. However, since we are dealing with digital lexica, mapping ontologies for each of these concepts need to be created before the resources become interoperable. The recommended use of the ISO DCR will, at least, facilitate this.

Because the LEXUS ideas are based on LMF and DCR, lexica created in LEXUS can be made semantically interoperable. This facilitates searching across and merging lexica. For researchers who do not wish to make use of the ideas of LMF or the DCR, LEXUS also allows the creation of non-LMF lexicon structures and the use of data categories not linked to the DCR.

E-ISSN 1934-5275

2. DOCUMENTATION.. Kotcheva mentions in her review that the LEXUS manual is on the tools website: www.lat-mpi.eu/tools/lexus. On this website is also a short guide (an A4 guide, as we call it) for easy start-up. The manual and A4 guide are always related to a version of the tool. At the moment of writing the version of both the LEXUS tool and the manual is version 2.0 beta 1. The tool version can be checked in the "about" under the LEXUS icon in the workspace.

3. SYNCHRONIZATION WITH TOOLBOX. We do realize that interoperability of programs with Toolbox is important to most linguists, especially since LEXUS does not yet provide interlinearization and interaction with ELAN (but we are working on this). However, Kotcheva already remarks that Toolbox does not prescribe its lexical entries to be consistent with the lexicon structure defined in the Toolbox type file. We have written a special manual on how to curate Toolbox data for initial import into LEXUS. Merging new lexical entries, created in Toolbox, with existing LEXUS lexica has similar consistency problems. The Toolbox developers (SIL) also realize the problem they have created by allowing inconsistent content, and like us, SIL is working on a method of data curation, based on chunking and parsing (Aumann and Bird 2009). As long as Toolbox does not force lexical entries to be consistent with a defined structure, merging of new lexical entries will remain a difficult and specialized task, but we are happy to assist users with this.

4. SEARCHING AND FILTERING. The search functions have been extended with a "filter" option, which can be used to create a filtered word list of one lexicon, e.g., to create a wordlist of lexical entries of one specific semantic domain only. Filters must be created and applied at the lexicon level.

5. EXPORT. Previous versions of LEXUS contained an XML export function based on the format suggested by the LMF standard. This export method suffers from a number of shortcomings that must be resolved in order to achieve interoperability at all levels of LMF. These are currently being addressed by some of our projects.

6. USER INTERFACE. The suggestion made by Kotcheva to create a simple user interface for read-only users has been implemented with the newest version of LEXUS (2.0 beta 1). A stripped user interface, with all editing options removed, is presented to users with read-only access to a lexicon. We would like to get feedback from users on this read-only user interface.

7. VICOS. It is unfortunate that ViCoS was not included in the review of LEXUS. ViCoS is one of the newly added strong points of LEXUS and should not be considered to be a separate tool. ViCoS has been developed at the request of members of the speech communities of the DoBeS projects. It was observed that, particularly for language community members, access to lexical data was not considered to be appealing if made available only through wordlists, but that conceptual spaces, where concepts are related to other concepts, based on culturally defined relation types and associations, provide much easier access to the underlying lexicon content. Such conceptual spaces can be created and browsed using ViCoS. Lexical information stored in LEXUS is at the basis of the conceptual space: a concept can be a complete LEXUS lexical entry or a part of it. From the conceptual space, users can move "back" to LEXUS to get more detailed lexical information or contextual multimedia displays. From ViCoS, users can also move forward to external resources on the web, e.g., Wikipedia. In ViCoS, relations between concepts can have universal relation types like synonym or antonym, but in addition, culturally relevant relation types can be created at the LEXUS workspace level. In the next version of the user interface, ViCoS will appear under an extra tab in the LEXUS Lexicon Editor, and ViCoS conceptual spaces will be made visible directly when viewing lexical entries.

8. CONCLUSION. We would like to point out that LEXUS and ViCoS are indeed tools still under development. This presents a challenging opportunity for you researchers in the language documentation field. We wish to create a tool which is useful for you, and we therefore invite all of you to send us your remarks and suggestions. We have one software developer working on the tools full-time, as well as a student assistant trained in LEXUS who provides intensive support for researchers who wish to integrate their data into LEXUS.

References

Aumann, Greg, & Steven Bird. 2009. Curating lexical databases for minority languages. Paper presented at the 1st International Conference on Language Documentation and Conservation (ICLDC), March 2009. Honolulu, Hawaii. (http://scholarspace.manoa.hawaii.edu/handle/10125/5094)

Kotcheva, Kristina. 2009. LEXUS, from Max Planck Institute for Psycholinguistics Nijmegen. *Language Documentation and Conservation* 3(2). 241–246.

Jacquelijn Ringersma
Jacquelijn.Ringersma@mpi.nl

Marc Kemps-Snijders
Marc.Kemps-Snijders@mpi.nl