abstract 138

Typological NNC-Database: Storage of Cross-Linguistic Data
Category: Oral/Poster
Theme Session: Linked Data in Linguistic Typology

The paper presents a cross-linguistic database of numeral-noun constructions (NNCs). The database under development will be freely accessible online and open-ended, allowing users to perform queries and to add their own research material. A special user interface will be provided for easiness of data-management.

The database is aimed to hold descriptions of a certain type of syntactic constructions in terms of a large number of grammatical features (morphological, morphosyntactic and morphosemantic). The use of the constructions in each language can be illustrated with linguistic examples supplied with interlinear glosses (following Leipzig glossing rules). These data and analyses will be described with basic metadata, such as information about languages, sources and contributors.

By now the majority of databases (not only those constructed for linguistic purposes) are developed as relational databases (RDB) and managed with some kind of SQL. It is the most popular and well-tried way to organize data storage and retrieval. However, in the last years data engineers give preference to Resource Description Framework (RDF) ([Lassila and Swick, 1999]) in combination with suitable RDF query languages, especially SPARQL. Using this approach for our purposes has a number of advantages over using RDB with SQL.

For each language, the database will hold an amount of highly structured data. An NNC usually consists of two or three elements: noun, numeral and (in some languages) classifier, and up to 20 features (depending on the language) are used to describe relations between them. Furthermore, each language can have several types of NNCs, differing in word order, syntactic relations, meaning, etc. To this we must add glossed examples and metadata. Multiplying by the number of languages, we get a highly complex data structure compared to a relatively moderate amount of data. Moreover, one might want to rethink this structure as new typological data come into consideration. RDF offers a way for uniform treatment of all relations as “subject-predicate-object” triplets. Also, in an RDF database we need not create additional tables or use foreign keys to create a new relation, as we do it in RDBs. One can merely add new triples by defining new subjects, objects or predicates, as shown in [Moran 2012]. This feature of RDF-storage is especially important for open-ended projects.

Second, RDF-storage is usually queried through web-interface, so the user need not install special tools to work with it, saving time and space.

Finally, the integration with other databases is much more seamless for an RDF storage (see e.g. [Chiarcos et al. 2012]). In addition to differences in data model between RDBs, there exist many kinds of SQL, which may impede integration or migration of data. RDF, on the contrary, has higher level of standardization, which makes changing storage a simple operation.

References:

Chiarcos, C., Hellmann, S., Nordhoff, S.: Towards a linguistic linked open data cloud: The Open Linguistics Working Group. TAL 52(3), 245–275 (2011).

Lassila O, Swick RR (1999) Resource Description Framework (RDF): Model and syntax specification (recommendation). http://www.w3.org/TR/REC-rdf-syntax

Moran S., “Using Linked Data to create a typological knowledge base”, in Chiarcos et al. (2012), p. 129-138, 2012. Companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany