Fine-grained typological investigation of grammatical constructions using Linked Data

Overview: This talk discusses the use of Linked Data to create a database of grammatical constructions, emphasizing the potential advantages of Linked Data in such a context and clarifying the extent to which the technology is ready to be utilized more generally for typological work. Unlike many prominent projects making use of Linked Data (e.g., Nordhoff 2012), the technology was not chosen specifically to facilitate interoperation with other datasets but, rather, because it was well suited to model grammatical objects of theoretical interest. The details of this work, therefore, should be of value both to those interested in data interoperation and to those engaged in more traditional typological investigation.

A database of templates: Linearization templates—that is, grammatical devices describing unexpected patterns of linear stipulation—do not lend themselves to straightforward typological classification due to their heterogenous nature, ranging from Semitic CV-skeletons, to Athabaskan position-class morphology, and beyond (Good 2011). Describing templates in ways which allow them to be rigorously compared, therefore, requires a flexible and easily extendable database system.

Bickel (2010) deals with comparable concerns for clause-linkage constructions by adopting a “multivariate” approach. This involves the development of an extensible database that allows for the coding of an open-ended range of patterns of variation as they are discovered. Bickel’s specific implementation is effectively based on a model where a construction is conceptualized as a “bag” of feature-value pairs. This approach is appropriate for his dataset but requires refinement to be applied to a database of templatic constructions. This is because, in addition to encoding a template’s holistic properties, it is clearly also desirable to encode fine-grained structural relations among its subcomponents.

The role of Linked Data: Formal syntactic approaches making use of feature structures like HPSG (Sag et al. 2003) provide well-developed solutions for encoding structural relations in the form of nested attribute-value matrices. However, associated implemented systems (e.g., Copestake 2002) are designed to describe grammars of individual languages rather than for cross-linguistic investigation. Linked Data, by contrast, can both readily encode nested attribute-value matrices and allow for data to be straightforwardly embedded within an ontology, facilitating rigorous comparison. It is, therefore, an excellent tool for describing templatic constructions typologically and, indeed, for any type of grammatical construction that can be modeled using sets of nested attribute-value pairings.

The Linked Data database of templatic constructions forming the basis of this talk was created using the readily available Protégé tool and is stored in an XML format which can be queried for typological purposes using existing code libraries. A fortunate byproduct of the use of Linked Data is that the database is immediately available in an open, interoperable format, lowering technical barriers to data sharing. Thus, a technology originally chosen for its expressive power also has important advantages with respect to data reusability. While the skillset required to exploit the relevant technologies is not yet common among typologists, it is not obviously more complex than what is needed to use the popular R statistical programming language, meaning the barriers are not as high as they may otherwise seem.

Bickel, B. 2010. Capturing particulars and universals in clause linkage: A multivariate analysis. In *Clause-hierarchy and clause-linking: The syntax and pragmatics interface,* 51–102. Amsterdam: Benjamins.

Copestake, A. 2002. *Implementing typed feature structure grammars*. Stanford: CSLI.

Good, J. 2011. The typology of templates. *Language and Linguistics Compass* 5:731–747.

Nordhoff, S. 2012. Linked data for linguistic diversity research: Glottolog/Langdoc and ASJP Online. In *Linked data in linguistics: Representing and connecting language data and language metadata*, 191–200. Berlin: Springer.

Sag, I. A. et al. 2003. *Syntactic Theory: A formal introduction*. Stanford: CSLI.