Linguistic Annotations and Knowledge Representation

In recent years linguists have become more interested in data-oriented research. They use corpora and they manage primary data. In their efforts they are helped by modern linguistic tools which promote standardisation and the use of metadata. In this way persistence and interoperability of primary linguistic data is gradually increasing.

As a timely initiative, Linked Open Data is of relevance for different kinds of linguistic online resources, including specialised encyclopedic knowledge banks such as the WALS, endangered-language archives, and federative linguistic online databases such as TypeCraft and the SSWL. Given new web technologies, it has become possible to search for embedded objects and in particular for classes and properties defined in ontologies. Applied to linguistics this means that indexes, in the form of linguistic glosses, become central as links between primary language data and more abstract linguistic knowledge.

Present attempts to make linguistic ontologies operable are still too weak. The online system, TypeCraft for example, provides URI-links between system tags and GOLD to allow the look-up of linguistic notions. Yet, in its present form the relation is not interactive and not informative enough to be useful for the linguistic glossing process. Although development within the Digital Humanities has made linguistic ontologies more framework independent and more comprehensive, ontologies are still not used to their full potential.

In our presentation we will discuss annotation and ontology integration, building on work by Chiarcos (2008). We will describe our own annotation model which consists of relations between morphemes, strings of tags (rather than individual ones) and tag classes, to suggest a design beyond the simple 1-1 mapping from tag to grammatical concept. We are particularly interested in the annotation of multi-lingual data from less-documented languages. We furthermore would like to reflect the incremental character of the linguistic annotation process (Mosel 2006a) by promoting a more dynamic integration of ontological knowledge.

In our presentation we would like to suggest a design which allows us to supplement subclassOf relations with disjointness and n-ary relations, as well as the flagging of certainty.

In order to discuss design questions on a fairly concrete level, we have acquired in-depth hand annotated data of 4 less-documented languages from typecraft.org. These languages feature between 56186 and 3079 annotated morphemes, and we will discuss this data in our presentation.

References

Chiarcos,C. 2008. An ontology of linguistic annotations. LDV-Forum 2008 – Band 23 (1) – 1-16.
Farrar, S. and D. T. Langendoen. 2003. A linguistic ontology for the Semantic Web. GLOT International 7 (3), 97 - 100.

Mosel, U. 2006a. The art and craft of writing grammars. in Felix Ameka, Alan Dench and Nicholas Evans. *Catching language*. The standing challenge of grammar writing, 41-68.
Syntactic Structures of the World's Languages. http://sswl.railsplayground.net/,
Accessed on 15/01/2013.
TypeCraft. The Natural Language Database. http://typecraft.org
Accessed on 15/01/2013.
Windhouwer, M.A.. Towards standardized descriptions of linguistic features: ISOcat and procedures for using common data categories. At the KONVENS 2012 workshop Standards for Language Resources. Vienna, Austria.