

Toward the Interoperability of Language Resources

July 13-15, in conjunction with the 2007 LSA Summer Institute



Codifying Linguistics: LingRDF as an interchange format

Michael Cysouw

Max Planck Institute for Evolutionary Anthropology, Leipzig

Of all the knowledge available about the world's languages, only a very limited amount can be processed (semi-)automatically. Any future interchange format for language resources should also allow for the codification of all those important details that are currently only stored in books and articles - or, in many cases, only in the linguist's heads. Semantic web technologies (especially RDF) seems to be the logical way to go, and I would like to discuss some basic aspects of a LingRDF (some draft proposals will be made available in time before the workshop to enhance discussion).

From the perspective of language corpora or other large bodies of data, RDF is mostly not the optimal format for quick processing of information. However, RDF is ideal for data exchange because all relations between all information has to be explicitly described into minute details. In contrast, from the perspective of the individual linguist, RDF is ideal because of its open and non-centralised nature. Every linguist will be able to publish his/her own information by providing it in the appropriate RDF format. Such information can even minimally be the simple correction of a (perceived) error. Publishing this information will be alike to providing information via a web-page, but with better integration and searchability. By having everybody self-publish (with a possible addition of an online publication service for less computer-savvy colleagues) the provenance of all information is implicitly provided.

The structure of the LingRDF should be completely based on the kind of entities that linguists have learned to care about through a few centuries of experience. This will probably often result in an extremely verbose format, because numerous "dummy" entities have to be introduced. Dummy nodes are required in an RDF graph to maintain consistency between the encoding of some piece of data and an ontology. For example, consider the simple case of codifying the language name "German". It will probably be something like this: the language name "German" is a <Wordform> in the language English, linked to a dummy node <LanguageNameVariant>, linked to a dummy node <LanguageName>, linked to an ISO 639/3 code (and other information). This layered structure is necessary because there are different <LanguageNameVariants> for the same <LanguageName>, like "Deutsch" or "Aleman", and each of these <LanguageNameVariants> have possibly different <Wordforms>, like "Deutsche" or "Deutschen".



I expect the formulation of the details of the lingRDF will go through numerous revisions, because it will have to suit both linguists interested in large bodies of data and those interested in the tiniest details of the structure of individual languages. Nevertheless this process will not only allow linguists to fully exploit the potential of the semantic web, but it is also likely to give them insights into structural aspects of their data that have hitherto gone largely unnoticed.

