Annotation for Descriptive Studies on Languages of the World

In language typology, i.e., the systematic study or the unity and variation of the languages of the world, there are various infrastructural needs which are not yet in place. A central such need, which is feasible, is a database of bibliographical references to descriptive materials on languages of the world. (In contrast, a bibliography of all research articles relevant in linguistics is much too large to be feasible.) Language documentation and description is, and has been, an extremely decentralised activity, wherefore tracking such information is very difficult without specialised tools. Throughout the past century, well-over 500 bibliographies of this kind have been published in book form[1] by *different* authors, who, (in addition to updating) essentially re-do the prior work done by others.

We are engaged in an endeavor to list all bibliographic references to descriptive data for all lesser-known languages. (If there are some 7 000 languages in total, about 100 would count as well-known, and the rest as lesser-known.) Let's call a bibliographic reference to a publication with descriptive/documentational data on a lesser-known language BDP for short. We currently have about 14 000 BDP:s, and it is anticipated that a complete catalogue would number at around 20 000. Empirically, BDP:s data turn out to be of two prototypical kinds: individual descriptions, e.g. *Grundriss einer Grammatik der Konde-Sprache* and group descriptions, e.g. *Languages of the Lumi Subdistrict*. Perhaps surprisingly, ca 28% is of the group kind overall (though number varies a lot across areas). Therefore, we propose that BDP:s should be annotated as to *focus* with:

- **Language-id** for individual BDP:s
- **Group-id** for group BDP:s

There is already an iso-639-3 draft standard (ISO 639-3:2007) for languages of the world, so the language-id can simply be the three-letter code id, with which location, speaker number etc. can be retrieved separately. Group-id:s could be any name with geographical, genealogical or other inspiration which is *equated with a set of language-id:s* separately from the annotation of the language entry. This annotation is slightly flatter than the

[1] In fact, there are even bibliographies of bibliographies of the languages of the world.

one proposed by Cysouw and Good (2007), but has much better automatizability propeties (see below).

In addition, BDP:s should be annotated as to type, according to the following relatively uncontroversial hieracharchy:

- (full-length) descriptive grammar
- grammar sketch
- description of some element of grammar (i.e. noun class system, verb morphology etc)
- phonological description
- dictionary
- text (collection)
- wordlist
- document with meta-information about the language (i.e., where spoken, non-intelligibility to other languages etc.)
- note on unpublished manuscripts or people engaged in studying the language

We will argue that this form of annotation is a good trade-off between user needs and annotation automatizatibility – in fact, the bulk of the annotation can be done computationally (Hammarström, 2008). This ensures that this infrastructural endeavor will actually lead to a finished product, i.e., a freely available annotated database.[2]

References

Cysouw, M. and Good, J. (2007). Towards a comprehensive languoid catalogue. Presentation at the *Towards a Comprehensive Language Catalogue* workshop at the MPI for Evolutionary Anthropology, Leipzig, 28 June 2007, available at `http://email.eva.mpg.de/~haspelmt/cat.html`.

[2] The exact nature and maintenance of the end product, e.g., a wiki-style resource or a benevolent dictator run resource, will not be the topic of the present presentation.

Hammarström, H. (2008). Automatic annotation of bibliographical references with target language. In *Proceedings of MMIES-2: Wokshop on Multi-source, Multilingual Information Extraction and Summarization*, pages 57–64. ACL.