

Toward the Interoperability of Language Resources

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



Filtering lexical noise to resolve issues of field linguistic lexicons

Thorsten Trippel & Dafydd Gibbon
Universität Bielefeld

The 2006 charette on lexicography at DTSL revealed a major problem in field linguists lexicons created by tools such as Toolbox or Shoebox: lexicons show the more inconsistencies the larger they get. We term these inconsistencies in lexicon structure and content “the lexical noise problem”. In this contribution we outline a method for “lexical noise filtering” and demonstrate its feasibility based on the EMELD showcase lexicon for the Ega language (Gibbon, Bow, Hughes & Bird, 2004).

Our approach starts with the charette outcome report, which essentially advocated the following procedure: (1) Convert to XML; (2) Fix data structures; (3) Map data categories to model; (4) Fix data structures; (5) Fix data contents; (6) Convert to desired format. The specification is plausible but still leaves much to be desired, and the devil lies in the detail. The main point is that “convert to XML” will not work “as is”: XML is a generic tree formalism, and on the one hand more complex graphs are needed, and on the other hand, non-tree-structured constraints over these general graphs need to be specified.

The more detailed procedure we propose is as follows:

INPUT SPECIFICATION: set of legacy lexicons in arbitrary formats with no formal specification, possibly Shoebox/Toolbox databases, but possibly spreadsheets, word processor generated or other documents.

OUTPUT SPECIFICATION: set of lexicons with well-defined macrostructures (standard semasiological; thesaurus-like onomasiological; multilingual) and microstructures with well defined data categories and typed values).

METHOD SPECIFICATION:

1. Manual or automatic “learning” of the specification of a formal model for the macrostructure and microstructure of the legacy lexicon in terms of a generic formalism such as the lexicon graph (Trippel 2006) or feature structures, and implementation of this as XML structure. Cf. previous work in this area by Bird & Bell (2000), Gibbon, Bow, Hughes & Bird (2004), and previous graph models by Polguère (2006) for lexicons and Bird and Liberman (2001) for annotations.
2. Classification of the “lexical noise” types in relation to the model (metasyntax: incomplete microstructure, re-ordered microstructure, invalid data values;



- metasemantics: incorrect cross-reference link targets or link types). For metasyntactic problems, noise filters can often be generated automatically; metasemantic problems in general require interactive intervention.
3. Instantiation of the formal model with the legacy lexicon to the XML version of the formal model, with application of noise filters (in general: partly automatic, partly interactive semiautomatic).
 4. Transformation of the XML version (e.g. with XSLT) into the required output format, which may be the original legacy format or a standard, generic format which permits re-transformation into sets of lexicons with different macrostructural views on the lexical data.

An example: If transitive and intransitive lexical items are linked as synonyms, and this is classified as an inconsistency to be resolved, the resolution requires additional knowledge of the following kind: a word A and a word B can be synonyms if they belong to the same word class, have the same valence properties, same grammatical gender, etc. This information is language dependent and requires a formal analysis of the interrelations. This, and other information such as microstructure gaps, can be found by defining querying strategies over a lexicon graph as search space.

We suggest that a complete cure for lexical noise will not be automatic, but will require considerable interactive user intervention for selection of candidate inconsistencies.

This approach is being applied in ongoing work on the Ega dictionary:

1. the transformation of samples from a legacy Shoebox/Toolbox lexicon into a graph format;
2. a demonstration that the transformation is lossless by regenerating the Toolbox lexicon entries and comparing it with the original samples;
3. specification of procedures to find the selected noise types, using standard XML procedures.

The querying is based on the standard compliant XML Database Tamino by Software AG, using W3C query language XQuery.

References

- Bell, John and Steven Bird (2000), Preliminary Study of the Structure of Lexicon Entries. Proceedings of the Workshop on Web-Based Language Documentation and Description, Philadelphia, December.
- Bird, Steven and Mark Liberman (2001). A formal framework for linguistic annotation. *Speech Communication*, 33 (1,2):23-60.
- Gibbon, Dafydd (2001). On lexical objects and their properties: A contribution to the 'MetaLex' requirements specification for spoken language lexicon documentation.
- Gibbon, Dafydd (2004). Catherine Bow, Steven Bird and Baden Hughes. Securing Interpretability: The Case of Ega Language Documentation. Proceedings of the 4th International Conference on Language Resources and Evaluation, pp 1369-1372. Lisbon, Portugal.
- Polguère, Alain (2006). Structural properties of lexical systems: Monolingual and multilingual perspectives. In Proceedings of the Post COLING/ACL-2006 Workshop: Multilingual Language Resources and Interoperability, Sydney.
- Trippel, Thorsten (2006). The Lexicon Graph Model: A generic model for multimodal lexicon development. AQ Verlag, Saarbrücken.

