

Issue statement : Can we build a universal lexicographic platform?

Alain Polguère

OLST — Department of Linguistics and Translation
University of Montreal
alain.polguere@umontreal.ca

In the old days, linguists working on lexical models were marginalized as lexicography was not considered a priority, neither in R&D nor in academic circles. Things have changed. Lexicography has “come in the open” thanks to major projects such as WordNet and FrameNet, and the realization that large-scale lexical resources were a bottleneck in NLP as well as many other areas of applied linguistics. We know what the R&D and NLP communities want from lexicographers: large-scale lexical databases. But what do lexicographers need from the computer science community?

Writing rich lexical descriptions —descriptions that account for all the main properties of lexical units— is an extremely complex activity, involving multitasking, trial and error strategies, processing of multiple sources of information, simultaneous access to all components of the database under development, updating/refining/correcting processes, etc. Strangely enough, very little has been done in order to develop lexicographic platforms for the benefit of the whole community of developers of lexical databases. It is high time to tackle this problem.

The construction of such platforms poses many theoretical and technological challenges; for instance:

- They should be compatible with various approaches to lexical modeling. This poses problems similar to those encountered in the process of developing an ISO standard for lexical database structuring (ISO TC 37/SC4 initiative).
- They should cope equally well with all sorts of writing systems, both for lexicographic description proper and for the consultation of integrated corpus resources.
- They should allow lexicographers to add their own flavor to the working environment. Lexicographers should be able to make use of their own descriptive concepts for modeling such varied lexical properties as semantic content, combinatorial behavior (at lexical, syntactic and morphological levels), paradigmatic links, written and spoken idiosyncrasies, etc.
- They should permit the addition of inference modules that control, guide and optimize the lexicographic activity.
- They should be able to function on different types of data storage, whether dictionary-like (huge collections of individual entries) or net-like (huge graphs of lexical entities and links).

We propose to discuss all major issues related to lexicographic platforms, building on experience acquired by researchers in the field. It will be important to identify which aspects of the design of a universal lexicographic platform (ULexP) lead to a consensus and which call for deeper understanding of how various approaches to lexical modeling can be made compatible. Ultimately, the group should arrive at a proposal for the architecture of a ULexP.