Building the Global WordNet Grid

Adam Pease[1], Christiane Fellbaum[2], Piek Vossen[3]
[1] Articulate Software, [2] Princeton University, [1] Vrije Universiteit Amsterdam
[1] apease@articulatesoftware.com, [2] fellbaum@priceton.edu. [3] p.vossen@let.vu.nl

Lexical databases are built for an ever-increasing number of languages (Singh 2002, Sojka et al 2004, Vossen 1998), inspired by the development of the Princeton WordNet lexical database (Fellbaum 1998). The Global WordNet consortium was formed to encourage this growth and to maximize sharing and interoperability of tools and methods; another major goal is to ensure the interlinking of the databases' lexical contents. But efforts towards interconnection and reusability have progressed unevenly and opportunistically within the community of wordnet developers. Attempts to link lexical elements have primarily relied on Princeton's English WordNet as a "hub". This presents obvious problems, as English does not lexicalize the same concepts as other languages; as a result, wordnets following this approach might well be biased towards the structure and lexical coverage of English. Another, language-independent approach is needed. At the Third Global WordNet Association conference (Sojka et al. 2006), a proposal was made for building a comprehensive worldwide wordnet Grid, free of language-specific biases.

A semantic network like WordNet shows structural gaps, where the geometry of the arcs expressing semantic relations would require a word, yet where the language does not have one. For example, Fellbaum (1998) argues for the existence of specific “accidental” gaps in the English verb lexicon on the basis of syntactic evidence. Cross-linguistic differences in lexicalizaton patterns abound; a well-known case are kinship relations (e.g. Kroeber 1917).

Arguably, the concept-word mappings of any given language are to some extent accidental; existing words do not fully reflect the inventory of concepts available to speakers. That inventory can be represented in the non-lexical ontology of the Suggested Upper Merged Ontology (SUMO) (Niles and Pease 2001), which is an open-source, formal ontology stated in first-order logic. SUMO consists of an upper level of roughly 1000 terms and 4000 axioms, a mid-level ontology that has several thousand more terms and definitions, and domain ontologies that cover over several dozen specific areas including world government, finance and economics, altogether totaling 20,000 terms and 70,000 axioms. SUMO has been mapped by hand to all of the WordNet 3.0 noun and verb synsets (Niles&Pease 2003).

The construction of the Grid will initially focus on three different sets of concepts:

1. The Common Base Concepts (Vossen 1998): a set of concepts that play a major role **[Piek: can we elaborate wht "major role" means and thus justify the choice of person and vehicle?]** in the building of wordnets in various languages, e.g. the concepts of *person* and *vehicle*.
2. The Basic Level Concepts (Rosch 1977): a set of concepts that are frequent and salient; they are neither overly general nor too specific. "Sister" Basic Level Concepts elaborate their shared superordinate concept in a way that maximally distinguishes the Basic Level Concepts from one another, whereas concepts subordinate to the Basic Level Concepts do not add highly distinguishing features. Basic Level Concepts include *house, apple*, *car.*
3. Other concepts lexicalized in languages that **somehow [can this be specified?]** depend on the first two sets, e.g. verbs like *sell* and *buy* that represent different perspectives of the same process.

The Common Base Concepts (CBCs) typically are found at the top of a wordnet hierarchy and subsume many hyponyms. The semantic implication of the wordnet structure is preserved, but the ways in which wordnets are structured across languages vary. SUMO provides a common semantic representation for different wordnets. For example, although *container* is a CBC, it is not lexicalized in Dutch and therefore the Dutch wordnet will have a mapping of more specific *containers* to this CBC.

The Roschian Basic Level Concepts (BLCs) are expected to be more universally lexicalized. These concepts are well represented in SUMO and that the mappings from wordnets is relatively straightforward, i.e., the appropriate lexemes are labels for equivalent concepts in SUMO (following Fellbaum and Vossen 2007 and Vossen and Fellbaum forthcoming.).

Other areas of the lexicon are likely exhibit language-specific idiosyncracies. Some words may refer to notions that are unique to a given culture; these need to be represented in the ontology. Others idiosyncracies are linguistic in nature, such as gender lexicalization (e.g., male and female professions), aspect lexicalization (different phases of a process), lexicalization of perspectives of the same process, etc.. From an ontological point of view, such lexicalizations do not warrant extensions to the interlingua, but the Grid needs to provide a formal semantic grounding as well as the possibility to relate these lexicalizations cross-linguistically. For example, *teacher* in English is underspecified for gender, whereas some other languages make a gender distinction explicit via distinct lexicalizations. The Grid must provide a mechanism for expressing such lexicalizations in the interlingua through complex mappings to the set of concepts (Fellbaum and Vossen 2007, Vossen and Fellbaum forthcoming.).

The procedure for constructing the Grid will initially focus on the CBCs and the BLCs, which will be expressed in terms of SUMO definitions. We invite people from all language communities to upload synsets from their language to the Grid, so that it will gradually come to be representative for many languages. The Grid will be freely and publicly available, in the spirit of Princeton WordNet. In a later phase, the mapping of other types of concepts to the initial core structure will be examined.

References

Fellbaum, C. (ed.) . 1998. WordNet: An Electronic Lexical Database. MIT Press.

Fellbaum C. and P. Vossen. 2007. "Connecting the Universal to the Specific: Towards the Global Grid", In: Proceedings of The First International Workshop on Intercultural Collaboration (IWIC 2007), Kyoto, Japan, January 25-26, 2007

Kroeber, Alfred. 1917. California Kinship Systems. University of California: Coyote Press.

Niles, I., & Pease, A. 2001. Toward a Standard Upper Ontology, in Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith (eds.). See also http://www.ontologyportal.org

Niles, I., and Pease, A. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, Proceedings of the IEEE International Conference on Information and Knowledge Engineering, pp 412-416.

Rosch, E. 1977. Human Categorisation. Ed. N. Warren Studies in Cross-Cultural Psychology, Vol. I, pp. 1-49. Academic Press. London.

Sinha, M., Reddy, M., and Bhattacharyya, P. 2006. An Approach towards Construction and Application of Multilingual Indo-WordNet, 3rd Global Wordnet Conference, Jeju Island, Korea.

Singh, U. N. (ed.). 2002. Proceedings of the First Global WordNet Conference. Central Institute for Indian Languages, Mysore, India.

Sojka, Petr, Pala, Karel, Smrz, Pavel, Fellbaum, Christiane, and Vossen, Piek (eds.) 2004. ``Proceedings of the Second International WordNet Conference." Masaryk University, Brno, Czech Republic.

Sojka, Petr, Choi, Key-Sun, Fellbaum, Christiane, and Vossen, Piek (eds.). 2006. Proceedings of the Third Global WordNet Conference. Masaryk University, Brno, Czech Republic.

Vossen, P. (ed.). 1998. EuroWordNet. Dordrecht, Holland: Kluwer.

Vossen P., and C. Fellbaum. Forthcoming. "Universals and Idiosyncracies in Multilingual WordNets" In: Handbook Multilingual Lexicography, Oxford University Press.