Towards a Polish WordNet

Tadeusz Piotrowski (Opole University), Aleksandra Skrzypczyk and Adrianna Szafruga

WordNet is a is a lexical database of English, developed at Princeton University by a team headed by George Miller. The database is supposed to be organized according to the way lexical items are stored in the mind. Lexical items such as nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept, and synsets are interrelated by means of conceptual-semantic and lexical links. As WordNet is freely and publicly available, and its data can be downloaded, it is widely used as a lexicographic resource for computational linguistics and natural language processing, but also for building dictionaries for the human user. Because of its success there have been efforts to produce localized WordNet databases for other languages, for example EuroWordNet, including Dutch, Italian, Spanish, German, French, Czech and Estonian. A number of researchers in Poland have expressed their interest in producing a Polish WordNet, usually linguistics engineering specialist, and submitted research proposals.

To evaluate the feasibility of various methods of producing WordNets we carried out an intensive linguistic-cum-lexicographic study, in which two synsets for Polish, one for nouns and one for verbs, were actually produced and compared with the English counterparts. The nominal synset focused on the *emotion* synset, the verbal synset – on the sensory perception verb synset. The choice of particular synsets was completely arbitrary/random. The method used to build the synsets consisted in taking over data from monolingual dictionaries of Polish (so-called 'merge' method) and, after a careful study, rearranging them into synsets modelled of the English prototype.

There are several conclusions. One group of them relates to the quality of English WordNet and its suitability as the basis for WordNets of other languages. A very important one is that the present version of English WordNet is an unreliable source for comparison of organization of concepts in different languages, as it does not consistently follow theoretical principles on which it is based. The second conclusion is that the concepts, as reflected in the synsets, are strongly culture specific. These observations give additional support for the decision of applying the "merge" model for building new WordNets, as this method prevents copying inconsistencies of English WordNet.

The other group of conclusions relates to the linguistics differences between English and Polish, and it is interesting that it was possible to form some hypothesis by studying completely de-contextualized items. From our analyses it follows that, first, Polish and English do not match as far as the conceptual hierarchy is concerned. Second, there is a different coding between form and meaning in both languages. While in Polish a given form is usually unambiguously related to a specific meaning, in English one lexical form can be used in a variety of meanings. What is crucial in English is the use of the given lexical form in a wider syntactic form (a phrase) which identifies the intended meaning. Therefore one can say that one lexical item in Polish often corresponds semantically to several lexical items in English. This finding is in agreement with what we know from linguistics typology, from translation studies or from research on bilingual lexicography.