Towards the Global Wordnet

Cvetana Krstev[1], Svetla Koeva[2], Duško Vitas[3]
[1]Faculty of Philology
University of Belgrade
Studentski trg 3
11000 Belgrade
cvetana@matf.bg.ac.yu

[2]Department of Computational Linguistics – IBL, BAS
52 Shipchenski prohod, Bl. 17
Sofia 1113
svetla@ibl.bas.bg

[3]Faculty of Mathematics, University of Belgrade
Studentski trg 16
11000 Belgrade
vitas@matf.bg.ac.yu

The global wordnet is an extensive lexical-semantic network that constitutes of synonymous sets (synsets) linked with the semantic relations existing between them. The cross-lingual nature of the global wordnet is provided by the establishing of relations of equivalents between synsets that express the same meaning in different languages. The global wordnet offers not only the extensive data for the comparative analysis over lexical densities and levels of lexicalization but furthermore presupposes the successful implementation in different application areas such as cross-lingual information and knowledge management, cross-lingual content management and text data mining, cross-lingual information extraction and retrieval, multilingual summarization, machine translation, etc. Therefore the proper maintaining of the completeness and consistency of the global wordnet is an important prerequisite for any type of text processing to which it is intended.

The EuroWordNet (EWN) extended the Princeton wordnet (PWN) with cross-lingual relations [Vossen, 1999], which were further adopted by BalkaNet (BWN) [Stamou, 2002]. The languages covered by the EWN are Czech, Dutch, Estonian, French, German, Italian, and Spanish, respectively, and those covered by the BWN are Bulgarian, Greek, Romanian, Serbian and Turkish. The equivalent synsets in different languages are linked to the same Inter-Lingual Index (ILI) thus connecting monolingual wordnets in a global lexical-semantic network. The Inter-Lingual Index is based on the PWN (ILI is consecutively synchronized with the PWN versions), the synsets of which are considered as language independent concepts. Thus a distinction between the language-specific modules (English among them) and the language-independent module (the ILI repository) has to be focused. The ILI is considered as an unstructured list of meanings, where each ILI-record consists of a synset (if the language is not English, a proper translation or at least transliteration must be ensured), an English gloss specifying the meaning and a reference to its source.

Both EWN and BWN adopted the hierarchy of concepts and relations' structure of the English wordnet as a model to be followed in the development of each language-specific wordnet. For the monolingual wordnets a strong rule is observed – strictly to preserve the structure of the PWN because via the ILI a proper cross-lingual navigation is ensured. It is natural, that some of the concepts stored in ILI are not

lexicalized in all languages and there are language specific concepts that might have no ILI equivalent. In the first case, the empty synsets were created (called non-lexicalized synsets) in the wordnets for the languages that do not lexicalize the respective concepts. The non-lexicalized synsets preserve the hierarchy and their purpose is to cover the proper cross-lingual relations. Regarding the second case, the ILI is further extended both in EWN and BWN with some language specific concepts. The language specific concepts that are shared between Balkan languages are linked via a BILI (BalkaNet ILI) index [Tufis, 2004]. The initial set of common Balkan specific concepts consisted mainly of concepts reflecting the cultural specifics of the Balkans (family relations, religious objects and practices, traditional food, clothes, occupations, arts, important events, measures, etc).

There are four morpho-semantic relations included in PWN and mirrored in EWN and BWN, *Be in state*, *Derivative*, *Derived* and *Participle* [Koeva, 2004]. Those relations semantically linked synsets although they can actually be applied to the literals only (graphic and compound lemmas). Consider the following examples:
Be in state is an asymmetric inverse intransitive relation that links derivationally and semantically related adjectives and nouns. The English synset {attractive:3, magnetic:5} with a definition 'having the properties of a magnet; the ability to draw or pull' is in a *Be in state* relation with the synset {magnetism:1, magnetic attraction:1, magnetic force:1} with a definition 'attraction for iron; associated with electric currents as well as magnets; characterized by fields of force'; also the synset {attractive:1} with a definition 'pleasing to the eye or mind especially through beauty or charm' is in a *Be in state* relation with {attractiveness:2} denoting 'a beauty that appeals to the senses'.

Derivative is an asymmetric inverse intransitive relation between derivationally and semantically related noun and verb synsets. For example the English synset {rouge:1, paint:3, blusher:2} with a definition 'makeup consisting of a pink or red powder applied to the cheeks' is in *Derivative* relation with two synsets: {rouge:1} with a meaning 'redden by applying rouge to' and {blush:1, crimson:1, flush:1, redden:1} denoting 'turn red, as if in embarrassment or shame'.

Derived is an asymmetric inverse intransitive relation between derivationally and semantically related adjective and noun synsets. For example the synset {Cuban:1} with a definition 'of or relating to or characteristic of Cuba or the people of Cuba' is in a *Derived* relation with {Cuba:1, Republic of Cuba:1}.

Participle is an asymmetric inverse intransitive relation between derivationally and semantically related an adjective synset denoting result of an action or process and the verb synset denoting the respective action or process. Consider {produced:1} with a definition 'that is caused by' which is in a *Participle* relation with {produce:3, bring about:4, give rise:1} denoting 'cause to occur or exist'.

As can be seen by the examples, although the synsets are semantically linked, the actual derivational relations are established between particular literals. For the best performance of the multilingual data base in different text processing tasks a specification of the derivational links must to be kept at the level of literal notes (LNotes).

There are systematic morpho-semantic differences between English and Slavic languages – namely derivational processes for building relative adjectives, gender pairs and diminutives. The Slavic

languages possess rich derivational morphology which has to be involved into the strict one-to-one mapping with the ILI.

A vivid derivational process rely Slavic nouns with respective relative adjectives with general meaning 'of or related to the noun'. For example, the Bulgarian relative adjective {*стоманен*:1} defined as 'of or related to steel' has the Serbian equivalent {*čelični*:1} with exactly the same definition. Actually in English this relation is expressed by the respective nouns used with an adjectival function (rarely at the derivational level, consider wooden↔wood, golden↔gold), thus the concepts exist in English and the mirror nodes have to be envisaged.

The gender pairing is systematic phenomenon in Slavic languages that display binary morpho-semantic opposition: male↔female, and as a general rule there is no corresponding concept lexicalized in English. The derivation is applied mainly to nouns expressing professional occupations. For example, Bulgarian synset {преподавател:2, учител:1, инструктор:1} and Serbian synset {predavač:1} that correspond to the English {teacher:1, instructor:1} with a definition: 'a person whose occupation is teaching' have their female gender counterparts {преподавателка, учителка, инструктурка} and {predavačica} with a feasible definition 'a female person whose occupation is teaching'. There are some exceptions where like in English one and the same word is used both for masculine and feminine in Bulgarian and Serbian, for example {президент:1} which corresponds to the English synset {president:3} with a definition: 'the chief executive of a republic', and as a tendency the masculine noun can be used referring to females.

Diminutives are standard derivational class for expressing concepts that relate to small things. The diminutives display a sort of morpho-semantic opposition: big ↔ small, however sometimes they may express an emotional attitude too. Thus the following cases can be found with diminutives: standard relation big ↔ small thing, consider {стол:1} corresponding to English {chair:1} with a meaning 'a seat for one person, with a support for the back' and {столче} with an feasible meaning 'a little seat for one person, with a support for the back'; small thing to which an emotional attitude is expressed. Also, Serbian synset {lutka:1} that corresponds to the English {doll:1, dolly:3} with a meaning 'with a replica of a person, used as a toy' is related to {lutkica} which has both diminutive and hypocoristic meaning. There might be some occasional cases of the expression of that kind of concepts in English, {foal:1} with a definition: 'a young horse', {filly:1} with a definition: 'a young female horse under the age of four', but in general these concepts are expressed by phrases.

There are several possible approaches for covering different lexicalization at different languages [Vitas & Krstev, 2005]:

- treat them as denoting specific concepts and define appropriate synsets;
- include them in the synset with the word they were derived from;
- omit their explicit mentioning, but rather let the flexion-derivation description encompass these phenomena as well.

Treating morpho-semantic relations, relative adjectives, gender pairs and diminutives, in Slavic languages as relations that involve language specific concepts requires an ILI addition for the languages where the concepts are presented (respectively lexical gaps in the rest). This solution takes grounds from the following observations:

- relative adjectives, feminine gender pairs and diminutives denote an unique concept;

- relative adjectives, feminine gender pairs and diminutives are lexicalized with a single word in Bulgarian, Serbian, Czech and other Slavonic languages;
- relative adjectives, feminine gender pairs and diminutives in most of the cases belong to different word class comparing to the word from which they are derived (there are some exceptions, like diminutives that are derived from neuter nouns in Bulgarian) .
Moreover, as with the other morpho-semantic relations, a special attribute assigned at the LNotes must provide information for one-to-one derivational relations.

Although PWN's coverage does not compare yet with new wordnets, the latter are continuously extended and improved so that a balanced global multilingual wordnet is foreseen, thus the task of the proper encoding of different level of lexicalization if different languages is in a great importance regarding the Natural Language Processing.

[Koeva at al., 2004] S. Koeva, T. Tinchev and S. Mihov Bulgarian Wordnet-Structure and Validation in: Romanian Journal of Information Science and Technology, Volume 7, No. 1-2, 2004: 61-78.

[Stamou, 2002] Stamou S., K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit, M. Grigoriadou, BALKANET: A Multilingual Semantic Network for the Balkan Languages, Proceedings of the International Wordnet Conference, Mysore, India, 21-25 January 2002, 12-14.

[Tufis, 2004] D. Tufis, D. Cristea, S. Stamou BalkaNet: Aims, Methods, Results and Perspectives. A General Overview in: Romanian Journal of Information Science and Technology, Volume 7, No. 1-2, 2004: 1-32.

[Vitas & Krstev, 2005] Duško Vitas, Cvetana Krstev (2005) Derivational Morphology in an E-Dictionary of Serbian in Proceedings of 2nd Language & Technology Conference, April 21-23, 2005, Poznań, Poland, ed. Zygmunt Vetulani, pp. 139-143, Wydawnictwo Poznańskie Sp. z o.o., Poznań, 2005.

[Vossen, 1999] Vossen P. (ed.) EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer Academic Publishers, Dordrecht. 1999.