Corpora Issues in Validation of Serbian Wordnet

Cvetana Krstev[1], Gordana Pavlović-Lažetić[2], Ivan Obradović[3], and Duško Vitas[2]

[1] Faculty of Philology, cvetana@matf.bg.ac.yu
[2] Faculty of Mathematics, {gordana,vitas}@matf.bg.ac.yu
[3] Faculty of Mining and Geology, ivano@afrodita.rcub.bg.ac.yu
University of Belgrade

Abstract. In this paper we describe how the existing monolingual Serbian corpus, the bilingual Serbian/English (S/E) and Serbian/French (S/F) aligned corpora, and the appropriate morphological e-dictionaries, have been used in validation, development, and refinement of Serbian WordNet. The influence of different derivational processes, e.g. derivation of augmentatives/diminutives and possessive adjectives from nouns, to the structure of Serbian synsets is examined. A part of the experimental results that justify the applied approach is given.

1 Introduction

The Serbian wordnet (SWN) is being developed within the scope of BalkaNet, the Balkan wordnet project (BWN) aimed at producing a multilingual database with wordnets for a set of Balkan languages [4], based on the model of the EuroWordNet (EWN) [7]. Both BWN and EWN are structured as the Princeton WordNet (PWN) [1] in terms of synsets, sets of synonymous words representing a concept, with basic semantic relations between them forming a semantic network. In addition, BWN and EWN feature an Inter-Lingual-Index (ILI), which relates concepts in different languages.

The BWN, as a multilingual database, in its initial phase followed the basic pattern set by EWN. The formation of databases started from a translation of a common set of concepts named Base Concepts in EWN. Further development of the BWN databases was aimed at maintaining a common set of concepts to be shared among languages, while at the same time, giving liberty to each WN to introduce specific concepts for its own language on an as needed basis.

The SWN to date comprises 3314 synset entries, with 5821 (literal string, sense) pairs. Each literal string is accompanied by a unique sense number denoting its meaning within a particular synset. As in most BWN databases, the SWN synsets were developed by translation of their English counterparts in EWN 1.5, while preserving the EWN hypernym/hyponym structure. Existing E/S dictionaries in paper form were used, and literal senses, where existing, were taken from the six volume explanatory Serbian dictionary by Matica Srpska (RMS). The problems generated by the addopted approach essentially originate from the inherent differences of the Serbian and English languages, and they pinpointed the question of validation of Serbian synsets.

2 Applied Tools and Resources

The corpus of contemporary Serbian developed for NLP purposes contains about 30MW and is constantly being enlarged. It consists of texts from various sources: newspapers, agency news, literature, and textbooks. Parallel S/F and S/E corpora are also being developed and their size is now close to 1MW [5]. Texts in parallel corpora are aligned at sentence level.

Parts of these corpora have been used for the validation of synsets from the Serbian wordnet. From the monolingual corpus 15 issues of daily newspaper "Politika" were used (582KW). The E/S corpus consists of the novel "1984" and 5 issues of the monthly "JAT Review" (130KW in the Serbian and 150KW in the English part). The F/S corpus consists of three novels: "Candide", "Bouvard et Pécuchet", and "Le tour du monde en 80 jours", and one issue of the international monthly "Le Monde Diplomatique" (173KW in the Serbian and 201KW in the French part).

For corpora pre-processing the Intex system has been used [3]. The standard distribution of this system incorporates morphological e-dictionaries for French and English. The development of the morphological e-dictionaries in Intex format for Serbian is described in [6]. The system of morphological e-dictionaries of simple words in Intex format consists primarily of three parts: the dictionary of lemmas DELAS, the dictionary of word forms DELAF and regular expressions implemented by finite transducers that describe the inflectional properties of entries in DELAS. In text processing DELAF is used for word recognition and lemmatization, while DELAS and the regular expressions are used for word form generation. The actual size of the Serbian DELAF is about 800.000 word forms.

3 The Validation Process

The validation process starts with the search for the occurrences of literal strings from Serbian synsets in Serbian monolingual corpus and Serbian part of the multilingual corpora. For all occurrences it is checked whether they conform to the synsets to which the literal strings belong. This process can confirm the inclusion of a literal string into a synset or lead to its exclusion and possible move to some other synset. For instance, the verb boraviti has been originally placed in the synset (stanovati:1b, zxiveti:4, boraviti:1, prebivati:1) that corresponds to (dwell:2, inhabit:1, live:6, make one's home:1, ...) from PWN. However, concordances produced by Intex showed that this verb has the exclusive meaning of a temporary stay and that it was misplaced in this synset.

Bilingual corpora can be used for synset validation in a more fruitful way, especially having in mind the request that all synsets from a wordnet for a language other than English have to be associated, if possible, to a corresponding English synset via ILI. Thus between synsets in the English (or French) wordnet and SWN a 1-1 correspondence is established on the basis of the `eq-synonyms` relation. For instance, a 1-1 correspondence exists between the following synsets: (*glava:1*) ↔ (head:8); (*glava:2*,*um:1a*)↔(brain:2, head:9, mind:1, nous:1, psyche:1, chief:1); (*glava:5*, *odgovorno lice:1*)↔(chief:2, head:19, top dog:1).

Between the literal strings from the English (or French) wordnet and the SWN, however, a many-to-many correspondence exists. The purpose of the validation process is to investigate the nature of this many-to-many correspondence and confirm or reject its appropriateness. The ideal result is one which confirms the initial 1-1 mapping.

Table 1. Concordances of noun glava from English/Serbian corpus

...monolog koji mu se doslovno godinama odvijao u glavi. ...monologue that had been running inside his head, literally for years.
...u glavi osvetlila jedna sasvim razlicyita uspomena,... ...a totally different memory had clarified itself in his mind,...

The validation process proceeds in two steps. In the first step one literal string from the Serbian wordnet is searched for in the Serbian part of the bilingual corpus and then its corresponding terms are looked for in the English (or French) part of the corpus. In the second, all literal strings in the English (or French) wordnet that are in a many-many correspondence with the chosen Serbian literal string are looked for in the English (or French) part of the corpus; all corresponding Serbian terms are then searched for in the Serbian part of corpus.

The nature of correspondence is then analyzed. This analysis can either remove some links from the initial correspondence or add new Serbian literal strings and links. Table 1 shows an excerpt from the concordances of the aligned corpus. Using this procedure the adequacy of the English and French wordnet can also be analyzed: however, that was not our goal.

4 Illustrative Example

The validation process will be illustrated by the noun glava. This literal string in SWN to date occurs in 3 synsets. These synsets, together with their corresponding synsets in the English and French wordnets are shown in Table 2. In the first step of the validation process, noun glava was searched for in all subcorpora, concordances were produced, and the meanings of all the occurrences were analyzed. Out of 404 occurrences, the noun appeared in its first meaning 197 times, in the second meaning 52 times, and in its third meaning only twice.

After that, all the translation equivalents of the noun glava, belonging to the same part of speech, were looked for in the concordances of aligned texts (Table 3). The obtained results show that the translation of the noun glava, as expected, is itself not always a noun. For instance, glava in sense 1, which occurs 67 times in the E/S corpus, has an English equivalent in one out of four different nouns only 52 times. In the remaining cases, the equivalent is either missing (e.g. for English "...he nodded..." the Serbian equivalent is "...on klimnu glavom..."), or is differently formulated (e.g. for English "...without looking..." the equivalent is "...ne dizxucxi glave...").

Table 2. Three synsets to which noun glava belongs

SWN	EWN	FWN
glava:1	head:8	tête:2
spolxasxnxi deo tela:X	external body part:1	membre:5
deo tela:1	body part:1	partie du corps:1
deo:y, komad:1b	part:10,piece:10	morceau:1,partie:3
entitet:1,objekat:1	entity:1	entité:1
um:1a,glava:2	brain:2,head:9,mind:1,. . .	cerveau:5,esprit:1
saznanxe:1	cognition:1,knowledge:1	savoir:1,connaissance:5
psihicyko svojstvo:X,. . .	psychological feature:1	chose du psychisme:1
odgovorno lice:1,glava:5	chief:2,head:19,top dog:1	chef:1
vodxa:1	leader:2	leader:1,guide:5,chef:6
lxudsko bicxe:1,. . .	human:1,individual:1,. . .	homme:7,mortel:1, . . .
bicxe:1,stvor:1a,. . .	being:1,life form:1,. . .	organism:1,être:2,. . .
enitet:1,objekat:1	entity:1	entité:1

Table 3. The translation equivalents of the noun glava analyzed by its senses. The instances that were in the original correspondence are in italic

	English	French
glava:1	*head* (46), face (2), bleat (1), skull (3)	*tête* (95), cervelle (1), chef (2), profil (1), menton (1), hure (1), caboche (1)
glava:2	*head* (5), *mind* (2), *brain* (2), consciousness (1)	tête (3), mémoire (1)
glava:5	no occurences	père (1)

In the second step of the validation process, concordances were produced for all literal strings that were initially in a many-to-one correspondence with the noun glava. The concordances of the aligned texts enabled the extraction of all equivalent literal strings being of the same PoS (Table 4). The literal strings *nous*, *psyche* and *top dog* have not been found in the E/S corpora. In a certain number of cases such equivalences cannot be found because they are either missing (e.g. from English "...She had not a thought in her head that..." glava is missing in Serbian text "...Nije imala nijednu misao koja...") or have been differently formulated (e.g. instance, "...always in the back of the head, without warning,..." is in the Serbian text "...pucali su s ledxa, bez upozorenxa,..."). Finally, the analysis of results obtained on the E/S corpora yielded new synsets: (head:8) ↔ (*glava:1*, glavica, potilxak, lice); (brain:2, head:9, mind:1, nous:1, psyche:1) ↔ (*glava:2*, *um*, secxanxe, pamet, mozak, misao (pl.), svest, duh, intelekt); (chief:2, head:19, top dog:1) ↔ (sxef, prvi cyovek, domacxin, poglavar, nacyelnik).

The second step of the validation process was than repeated on the F/S corpora (Table 4). The same phenomenon occurred as in the E/S corpora, that is, that for certain searched literal strings the equivalents are missing (e.g., the equivalent of "...Au-dessus de sa tête se déployaient .." is in Serbian "...Iznad nxe su se sxirila...") or are formulated using different phrases: the adverbial phrase "la tête la première" has as equivalents the adverbs glavacyke and strmoglavce.

Table 4. The equivalents of the literal strings that are in a many-to-one correspondence with glava in E/S and F/S aligned texts, accompanied by the numbers representing their total frequency in the coprora vs. their frequency in the chosen 3 senses

English equivalents	Serbian equivalents with number of occurrences
head (77/73)	*glava:1* (46), *glava:2* (5), glavica (1), potiljak (4), lice (1), sxef (3),secxanxe (1), pamet (1), prvi cyovek (1), domacxin (1)
brain (17/3)	*glava:2* (2), mozak (1)
mind (111/99)	*glava:2* (21), *um* (3), misao (6), svest (20), duh (22), secxanxe (2), mozak (2), intelekt (1)
chief (17/3)	prvi cyovek (1), poglavar (1), nacyelnik (1)
French equivalents	Serbian equivalents with number of occurrences
tête (135/109)	*glava:1* (95), *glava:2* (3), cyelo (1), duh (1), znanxe (1)
cerveau (14/6)	*um* (1), mozak (4)
esprit (72/17)	duh (14), duhovni (1)
chef (32/18)	*glava:5* (3), sxef (6), vodxa (3), vodx (1), poglavica (1), staresxina (1), rukovodilac (1), premijer (1), sxeficx (1)

New synsets obtained after the analysis of the F/S subcorpora are: (tête:2) ↔ (*glava:1*, cyelo); (cerveau:5, esprit:1) ↔ (*um*, duh, pazxnxa, mozak); (chef:1) ↔ (*glava:5*, sxef, vodxa, vodx, poglavica, staresxina, rukovodilac, premijer, sxeficx).

The union of initial synsets with synsets obtained after the analysis of the E/S and F/S subcorpora led to the following result: (*glava:1*, glavica, potilxak, lice, cyelo); (*glava:2*, *um*, secxanxe, pamet, mozak, misao (pl.), svest, duh, intelekt, pazxnxa); (*glava:5*, *odgovorno lice*, sxef, prvi cyovek, domacxin, poglavar, nacyelnik, vodxa, vodx, poglavica, staresxina, rukovodilac, premijer, sxeficx).

At this point it seems that the original synsets have been significantly augmented. However, hypernyms, hyponyms and, sometimes, coordinates often occur as translations, and, in that light, some of the new synset elements have to be reconsidered. As the Serbian equivalent for the synset (chief:2, head:19, top dog:1) has gained the greatest number of new elements, we will take it as an example and show which of the new elements actually exist in the hypernym/hyponym tree in its vicinity, and should therefore not be added:

- vodx, vodxa: (leader:2) is hypernym of (chief:2, head:19, top dog:1);
- poglavica: (chieftain:2, headman:1, tribal chief:1) is hyponym of (chief:2, head:19, top dog:1);
- poglavar, staresxina: (higher-up:1, superior:3, superordinate:2) is hyponym of (leader:2) (that is coordinate of (chief:2, head:19, top dog:1);
- domacxin: (head of household:1, paterfamilias:1, patriarch:3) is hyponym of (chief:2, head:19, top dog:1)
- rukovodilac: (executive:3, executive director:1) is hyponym of (administrator:3, decision maker:1) that is hyponym of (chief:2, head:19, top dog:1)
- premijer: (chancellor:2, premier:2, prime minister:1) is hyponym of (chief of state:1, head of state:1) that is hyponym of (leader:2).

By repeating the same process for the other two synsets we get the final result: (*glava:1*, glavica); (*glava:2*, *um*, pamet, mozak, misao (pl.), svest, duh); (*glava:5*, *odgovorno lice*, sxef, sxeficx, prvi cyovek, nacyelnik).

In these final synsets two diminutives were added: glavica from glava, and sxeficx from sxef. Two notes should be added here. Sxeficx was detected in the aligned corpus as an equivalent of "le petit chef", and, as such, its synonymous equivalence with glava:5 and other elements of the synset is doubtful. On the other hand, glavica occurred as an equivalent of "...the head of a child...", and, as such, seems to be a valid synonym of glava:1. A second point is that it seems that the diminutive glavica could be a synonym only of the first sense of glava, and not of the others. The validation of such a statement can be established on the corpora only, as there is no valid derivational dictionary of Serbian.

5 Conclusion

The results obtained by validating the 3 synsets presented in this paper, and many more that have been processed, fully approve the usability of the corpora approach to the validation of wordnet synsets. Besides the reestablishment of synsets themselves, this approach enables the establishment of relations between various derivatives, either by including them in the same synset, if they have same PoS, or, by setting up a cross-PoS relation [2]. In this respect the corpora approach is particularly useful in detecting the derived forms in connection to the senses.

Iterating the procedure can further refine the validation process. For instance, one of the equivalents of glava:2 in the E/S subcorpora was *consciousness*, and its equivalence in the F/S subcorpora was *mémoire*. These new equivalencies can be included in the second step of the validation process.

References

1. Fellbaum C. (ed.): *WordNet: An Electronic Lexical Database*, The MIT Press (1998)
2. Pala, K.; Sedlaek, R., Veber, M.: Relations between Inflectional and Derivation Patterns, Proc. of Workshop *Morphological Processing of Slavic languages*, EACL'03, Budapest (2003) 1–8
3. Silberztein, M. D.: *Le dictionaire électronique et analyse automatique de textes: Le systeme INTEX*, Paris: Masson (1993)
4. Stamou S., et al.: BALKANET: A Multilingual Semantic Network for Balkan Languages, Proc of *1st International Wordnet Conference,* Mysore, India (2002)
5. Vitas, D.; Krstev, C.: Structural derivation and meaning extraction: a comparative study on French-Serbo-Coratian parallel texts. In: Barnbrook, G. (ed.): *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora.* The University of Birmingham Press (2002)
6. Vitas, D.; Krstev, C.; Pavlović-Lažetić, G.: The Flexible Entry. In: Zybatow, G. et al. (eds.): *Current Issues in Formal Slavic Linguistics.* Peter Lang (2001)
7. Vossen, P. (ed.): EuroWordNet: *A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic Publishers (1998)