ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume 7, Numbers 1–2, 2004, 147–161

Using Textual and Lexical Resources in Developing Serbian Wordnet

Cvetana KRSTEV
Faculty of Philology, University of Belgrade
E-mail: cvetana@matf.bg.ac.yu

Gordana PAVLOVIĆ-LAŽETIĆ, Duško VITAS
Faculty of Mathematics, University of Belgrade
E-mail: {gordana, vitas}@matf.bg.ac.yu

Ivan OBRADOVIĆ
Faculty of Mining and Geology, University of Belgrade
E-mail: ivano@afrodita.rcub.bg.ac.yu

Abstract. In this paper we present two techniques for using textual and lexical resources, such as corpora and dictionaries, in validation and refinement of Serbian wordnet. We first 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 can be used in validation and enhancement of Serbian wordnet synsets. Then we present a quantitative technique based on a set of frequency parameters which indicate the coverage of a corpus by wordnet literals, the significance of one sense of a literal relative to the others, as well as the significance of one literal in a synset compared to other literals in the same synset. Experimental results justifying the applied techniques are given.

1. Introduction

The Serbian wordnet (SWN) is being developed in the scope of BalkaNet, the Balkan wordnet project (BWN) aimed at producing a multilingual database with wordnets for five Balkan languages (Greek, Turkish, Bulgarian, Romanian and Serbian) as well as Czech [10]. BWN is based on the model of the EuroWordNet (EWN),

a multilingual database with wordnets for Dutch, Italian, Spanish, German, French, Czech and Estonian [13]. The structures of both BWN and EWN wordnets are basically the same as the structure of American wordnet for English, the Princeton WordNet (PWN) [3], in terms of synsets, that is the sets of synonymous words representing a concept, with basic semantic relations between them forming a semantic network. Although each wordnet represents a unique language-internal system of lexicalization, the BWN and EWN wordnets are all linked to an Inter-Lingual-Index (ILI), which makes it possible to relate similar concepts between languages, a feature that can be used, among others, for cross-language information retrieval.

Since EWN tackled many problems that PWN as a monolingual database did not face, the databases of the BWN, in their initial phase, followed the basic pattern set by EWN. Although an alternative approach has been inspected, namely, a buildup of wordnets for Balkan languages from scratch, the approach that took prevalence was the initial formation of databases starting from a common set of concepts named Base Concepts in EWN. The synsets corresponding to these concepts in BWN were then generated mainly by translation from their equivalents in English, French and other languages included in EWN. Further development of the BWN databases aimed at keeping a common set of approximately 8 500 concepts (Base Concepts 1, 2, and 3) to be shared among languages, at the same time allowing each wordnet to introduce specific concepts for its own language on an as needed basis. However, in the course of the BalkaNet development the relation with EWN has been abandoned as the common decision has been reached to follow the development of the PWN and thus maintain a relation to its version 2.0 through the ILI. The common format for the exchange and linkage of data is XML. Although all BWNs use the same basic XML schema, each wordnet is free to enhance the schema for some particular purposes.

The Serbian wordnet has been developed according to this common approach. In the absence of both an explanatory dictionary and an English/Serbian dictionary in electronic form, the translation of English synsets from PWN was done manually, while preserving the PWN semantic structure. The fact that a Serbian dictionary of synonyms does not exist even in paper form made this task even more difficult. Literal senses in SWN, where existing, were taken from the six volume explanatory Serbian dictionary of Matica Srpska (RMS).

Since the RMS dictionary was published in 1971, new senses had to be attributed in SWN to some of the existing literals but also new literals had to be added. Another reason for refinement of senses defined by RMS is due to the fact that concepts, and hence literal senses in PWN are far more fine grained than the ones in RMS.

The main problems the adopted approach to development of the SWN generated originate from the inherent differences between the Serbian and English languages. Thus the question of validation of Serbian synsets on corpora was brought up.

In this paper we describe how existing monolingual Serbian, and aligned bilingual Serbian/English (S/E) and Serbian/French (S/F) corpora have been used in this process. Also, in order to establish more precise criteria for synset validation a set of numerical parameters related to literal-sense pair frequency in corpora has been developed.

2. Serbian Wordnet

In order to describe the structure of SWN, let us first stress the difference between terms "literal string" and "literal" as used in this paper. The former refers to a character string corresponding to a dictionary entry (of a traditional paper dictionary), while the later refers to a specific sense of the literal string. We could also say that a literal represents a pair: (literal string, sense). It is obvious that the number of literal strings is less than the number of literals in any natural language dictionary, and the same is true for a wordnet.

The SWN to date comprises of 6 290 synsets, with 10 583 literals. Out of 8 470 BWN common concepts, 5 381 have been included in SWN, or 63.5%. There are also 911 concepts in SWN that are not in the common set established by BalkaNet project team. The language specific concepts, not already in the PWN, have not yet been included in SWN.

Since literal string senses in SWN in general correspond to the ones in the RMS dictionary, the SWN turned out to be somewhat specific with respect to the sense tag, which does not contain only numbers but also combinations of numbers and letters and even letters only.

The average number of literals per synset for SWN is 1.68. This ratio, however, significantly differs for different PoS (Part of Speech). Table 1 shows the PoS related distribution of synsets and literals (l), the literal/synset (l/s) ratio, the number of synsets with only one, two or three literals, and the maximum number of literals per synset(max).

Table 1. Distribution of literals per synsets

PoS	synsets	percentage	literals	l/s	1	2	3	max
nouns	4 562	72.5%	7 308	1.60	2 508	1 539	395	10
verbs	1 494	23.7%	2 974	1.99	546	574	275	10
adjectives	227	3.6%	294	1.30	175	39	12	4
adverbs	7	0.1%	7	1.00	1			1
total	6 290	100%	10 583	1.68	3 230	2 152	682	

The distribution of literals per literal strings also differs for different PoS. Table 2 represents the number of literals and literal strings per specific PoS, the ratio between the two, the number of literal strings with only one, two or three senses. The last column in the table shows the literal strings that have the greatest number of senses in certain PoS categories.

Table 2. Distribution of literals per literal strings

PoS	literal(l)	literal string(ls)	l/ls	1	2	3	max
nouns	7 308	6 162	1.19	5 434	495	142	*mesto, vreme* (11)
verbs	2 974	2 200	1.35	1 716	332	87	*drzxati* (13)
adjectives	294	262	1.12	242	16	1	*velik* (8)

The semantic relations between synsets in SWN are summarized in Table 3. These relations have automatically been inherited from the PWN, but then they were all manually checked.

Table 3. Distribution of relations between synsets in SWN

relation	number of occurrence
hypernym	5 816
near_antonym	415
holo_part	302
verb_group	133
holo_member	498
be_in_state	105
subevent	56
causes	44
derived	97
particle	9

The problems encountered in developing the SWN following the adopted methodology are many, and as we have already mentioned are due to differences between the Serbian and English language. To name only a few: augmentatives and diminutives of nouns, possessive adjectives, verb form aspect and transitiveness, as well as the cross-PoS problem, which occurs when a literal that belongs to one PoS in English corresponds to a literal belonging to some other PoS in Serbian. An example is "peer:1", an English noun which is usually translated into an adjective "ravan" in Serbian: for instance "he is his peer" can be translated as "on je njemu ravan". Also, English noun "sort:2" is usually translated with the indefinite pronoun "nekakav". For instance, the most natural translation of the example of usage from English WordNet "she served a creamy sort of dessert thing" would be "posluxila je nekakvo kremasto zasladxenxe". These and other problems related to the fact that SWN has been developed based on the English WordNet pinpointed the question of validation of Serbian synsets using the existing textual and lexical resources, mainly corpora and e-dictionaries.

3. Textual and Lexical Resources

3.1. Monolingual and Multilingual Corpora

The corpus of contemporary Serbian developed for NLP purposes at the Faculty of Mathematics has now about 23 Mw and is constantly being enlarged. It consists of texts from various sources: newspapers, agency news, literature, and textbooks. It is available on-line for authorized users at `http://korpus.matf.bg.ac.yu/korpus` under IMS. Parallel Serbian/French and Serbian/English corpora are also being developed and their size is now close to a million words [11]. The Serbian/French parallel corpus consists of a dozen mostly classical novels and their translations (mainly French to Serbian but also Serbian to French) and texts from the French magazine "Le Monde Diplomatique". The Serbian/English corpus is of a smaller size and incorporates, besides novels, texts from the commercial magazine "JAT Review" published both in Serbian and English. Texts in parallel corpora are aligned on the sentence level using

two different alignment programs. One of them is the Vanilla aligner [2] based on the Church and Gale algorithm [4] and the other is XAlign, developed at LORIA in France [1] with some additional features built in by the NLP team at the Faculty of Mathematics. Excerpts from these multilingual corpora can be seen at the same address. Parts of these corpora have been used for the validation of synsets from the Serbian wordnet, and their size (in words) is shown in Table 4.

Table 4. The size of subcorpora used for validation of synsets from SWN

English/Serbian	JAT (Sr)	40 039	French/Serbian	Kandid (Sr)	33 857
	JAT (En)	46 751		Kandid (Fr)	36 435
	1 984 (Sr)	89 542		Flober (Sr)	79 085
	1 984 (En)	103 813		Flober (Fr)	94 693
	Winnie (Sr)	20 837		Vern (Sr)	50 421
	Winnie (En)	22 748		Vern (Fr)	59 413
	Hobit (Sr)	84 238		Monde (Sr)	9 745
	Hobit (En)	95 020		Monde (Fr)	10 173
Total	Serbian	198 626	Total	Serbian	173 108
	English	268 332		French	200 714

3.2. Serbian E-dictionary

For corpora pre-processing the Intex system, based on appropriate e-dictionaries and finite state transducers, has been used [9]. The standard distribution of this system includes morphological e-dictionaries for French and English. Morphological e-dictionaries in Intex format for Serbian are being developed at the Faculty of Mathematics [12]. The system of morphological e-dictionaries of simple words in Intex format consists primarily of three parts: dictionary of lemmas (DELAS), dictionary of word forms (DELAF) and a set of regular expressions implemented by finite transducers that describe the inflectional properties of entries in DELAS. An excerpt from Serbian DELAS dictionary is the following:

```
pasti,V682+Perf+It+Iref
pasti,V7+Imperf+Tr+It+Iref
glava,N600
zxenxen,A1+PP
zxersejski,A2+PosQ
krisxom,ADV
krivudavo,ADV+Adj
```

A lemma is followed by a PoS mark (V for verb, N for noun etc.) with a number that identifies the regular expression describing the inflectional properties of the corresponding lemma (for instance, noun glava 'head' belongs to the inflectional class 600), and various syntactic and semantic properties (for instance +Perf denotes that the verb pasti 'to fall' is perfective). The regular expression describing the inflectional properties of the nouns from the class 600 is:

```
<E>/:fs1q:fp2q + 1e/:fs2q:fp1q:fp4q:fp5q +
1i/:fs3q:fs7q + 1u/:fs4q + 1o/:fs5q + 1om/:fs6q +
1ama/:fp3q:fp6q:fp7q
```

Based on DELAS dictionaries and the regular expressions for all inflectional classes the DELAF dictionary is generated. An entry in this dictionary has the following form for Serbian:

```
glavom,glava.N600:fs6q
glavo,glava.N600:fs5q
glavama,glava.N600:fp3q:fp6q:fp7q
glave,glava.N600:fs2q:fp1q:fp4q:fp5q
```

In text processing, the DELAF dictionary is generally used for word recognition and lemmatization, while the DELAS dictionary and regular expressions are used for word form generation. The actual size of Serbian DELAS/DELAF dictionary is given in Table 5.

Table 5. The size of Serbian e-dictionaries

	DELAS	**DELAF**	**DELAF/DELAS**
Nouns	27 053	159 525	5.90
Adjectives	21 163	341 737	16.15
Verbs	14 511	431 800	29.76
Adverbs	2 947	2 947	1.00
Other	650	2 853	4.39
Total	64 324	938 862	14.60

Various lexical transducers have also been constructed that are used for the recognition of the running text words not covered with the e-dictionaries.

4. The Validation Process

The aim of the validation process is to establish the validity of the synsets designed and relations established among them. More specifically, it is:

- to justify the inclusion of a literal in a synset (section 4.1.)
- to detect other literal strings that can be added to a synset (section 4.2.)
- to calculate the significance of a certain literal string in a synset (section 4.3.)

4.1. Justification for Literal String / Synset Pairing

The validation process starts with the search for the occurrences of literal strings from Serbian synsets in Serbian monolingual corpus. The aim of this process is to establish whether the literal string belonging to a given synset actually lexicalize the corresponding concept. This process can confirm the inclusion of a literal string into

a synsets or lead to its exclusion and possible move to some other synset, possibly up or down the hypernym/hyponym branch. For instance, the noun instrumentarijum has been originally placed in the synset (instrumentarijum:1a) that corresponds to the synset (instrumentality:3, instrumentation:1) in the English WordNet. This noun occurs 6 times in the Serbian corpus, but not once in the required meaning (Table 6). Its inclusion in the corresponding Serbian synset was obviously a consequence of its English lexicalization. As the analysis of dictionary definitions for all the established hyponyms of this concept did not suggest another solution, this synset has, at least for the time being, been placed in the category of non-lexicalized concepts.

Table 6. Concordances for the noun instrumentarijum in Serbian corpus

teorijski okvir i metodološki <instrumentarijum>, gube u moru činjenica, tako da
beleži 475 godina rada i ima vredan <instrumentarijum>. Pravi vagnerijanci i
130 muzičara, i ima veoma vredan <instrumentarijum>. Tu postoji kontrabas iz 17.
izvan kontrole svesti. Frojdov <instrumentarijum> U svom eseju manjeg obima
svog postepeno izgrađivanog <instrumentarijuma>. Zatim , antropološki pristup
služeći se psihoanalitičkim <instrumentarijumom>, pokušao da pronikne u tajnu

In many other cases, all literal strings of a synset have been confirmed on the corpus, as is the case, for example, for synsets (poslovna zgrada:1, poslovni objekat:1) ↔ (office building:1, office block:1), and (prepustiti:2, ostaviti:11) ↔ (entrust:2, leave:9). In addition to that, examples including literal strings extracted from corpus have been included in the Serbian wordnet, as the content of the element <usage>. Presently, 319 synsets have been checked against corpus, and as a result 386 <usage> tags have been added in Serbian wordnet. Some of the checked literals, precisely 8 of them, were not found in corpus of contemporary Serbian, and the content of their <usage> tag has hence been set to "not confirmed".

4.2. The Enhancement of Serbian Synsets

Bilingual corpora can be used for synset validation in a more demanding but also more fruitful way, especially having in mind that all the synsets from a wordnet for a language other than English are associated, if possible, to a corresponding English synset via ILI. Thus between synsets in English (or French) WordNet and SWN a one-to-one correspondence is established based on the EQ-SYNONYMS relation[1].

For instance, in Serbian wordnet there are five synsets to which the literal string oblik belongs. A 1-1 correspondence exists between the set of these synsets and the English and French WordNets:

1. (oblik:1, forma:1x) ↔ (form:7, shape:5, cast:4) ↔ (forme:3)
2. (oblik:4, forma:y) ↔ (form:3, shape:8, pattern:1) ↔ (forme:5)
3. (oblik:5, oblik recyi:X) ↔ (form:1, word form:1, signifier:1, descriptor:1) ↔ (forme:8)

[1]The French WordNet used is the one delivered by the EuroWordNet project.

4. (oblik:8a, forma:1) ↔ (shape:2, form:6) ↔ (forme:10)
5. (vrsta:1ax, oblik:3b, forma:x) ↔ (kind:1, sort:1, form:2, variety:5) ↔ (forme:2, espèce:3, sorte:3, variété:2, ganre:11)

In the bilingual corpus, however, a many-to-many correspondence exists between the literal strings associated to the Serbian synsets and those associated to the corresponding English synsets. The purpose of the validation process is to investigate the nature of this many-to-many correspondence and confirm or decline its appropriateness. In addition to that, it detects other literals both Serbian and English that could be added to the initial synsets.

Table 7. Concordances of the noun oblik from E/S corpus

nina bogatih sxumom, kao i najlepsxi deo obale Jadranskog mora omogucxavaju Jugoslaviji razvoj svih oblika letnjeg i zimskog turizma.</seg> .EOS <seg id='JAT0209-e-21.1.21.2.5'>Favorable climate (four seasons) and qood natural and geographic position, a qreat number of rivers, lakes and wooded mountains, as well as the most beautiful coast on the Adriatic Sea have made it possible for Yugoslavia to develop all *forms* of summer and winter tourism.</seg> .EOS
Ove prvobitne figurice nisu bojene, a izradjivane su u obliku bebe, majke sa detetom u narucyju, Hrista u kolevci...</seg> .EOS <seg id='JAT0201-e-07.1.7.3.2'> These earliest figurines were not painted and were *shaped* as a baby, a mother holding baby in her arms, as Christ

The validation process proceeds in two steps ([6]):

1. The literal strings from one Serbian synset are searched for in the Serbian part of the bilingual corpus using the Serbian e-dictionaries and then their corresponding terms are looked for in the English (or French) part of the corpus.
2. All literal strings in a corresponding English (or French) synset together with the terms detected in the step 1 are looked for in English (or French) part of the corpus using the English (or French) e-dictionaries; all corresponding Serbian terms are then searched for in the Serbian part of the corpus.

The nature of the correspondence is then analyzed. This analysis can either remove some links from the initial correspondence or add new Serbian literal strings and links. Table 7 shows an excerpt from the concordances of aligned corpus. Using this procedure the adequacy of English and French WordNets can also be analyzed; however, that was not our goal.

The application of this procedure to the chosen set of five synsets and the E/S corpus showed that three of the concepts occur in the corpus represented by literals from appropriate Serbian synsets. The corresponding English terms were in most cases literals already attributed to these concepts in the English WordNet, but there

were some other realizations too (Table 8). Naturally, in a number of cases a direct literal-to-literal correspondence does not exist due to translation modifications (for instance, a noun phrase is translated by a verb phrase, as is the case with the second example in Table 7.

Table 8. Realization of chosen set of Serbian synsets in E/S corpus

Synset	**Frequency**	**Realized literal strings**	**Non-realized literal strings**	**Corresponding English terms**	**New candidates for English synsets**
1.	20	*oblik* (16), *forma* (4)		*shape, form*	*shapliness* (1), *appearance* (1)
2.	2	*forma* (2)	*oblik*	*form*	
3.	0		*oblik, forma*		
4.	0		*oblik, forma*		
5.	28	*oblik* (2), *vrsta* (26)	*forma*	*sort, kind, variety*	*type* (8), *assortment* (1)

In the second step, the same procedure is applied using literal strings from the English counterparts of the chosen set of Serbian synsets, enhanced by the literal strings detected in step one as possible candidates for the English synsets (the last column in Table 8 as keywords for Intex "locate pattern" function. Again, three out of five synsets were found, and some new candidates for the Serbian synsets were obtained. However, only one of them occured with the frequency greater that one (Table 9). This is the literal tip:1a (in English WordNet type:1). In SWN it is a hyponym of the synset 5, which justifies its appearance in the searched context.

Table 9. Realization of English synsets linked to the chosen set of Serbian synsets enhanced by literals detected in step one of the procedure

Synset	**Freq.**	**Realized literal strings**	**Non-realized literal strings**	**Corresponding Serbian terms**	**New candidates for Serbian synsets**
1.	16	*form, shape, shapeliness, appearance*	*cast*	*oblik, forma*	*linija* (1), *gomila* (1)
2.	5	*Form*	*shape, pattern*	*oblik, forma*	
3.	0		*form, wordform, signifier, descriptor*		
4.	0		*shape, form*		
5.	33	*kind, sort, form, variety, type*		*oblik, vrsta*	*tip* (17), *vid* (1),

The same procedure has been applied to the F/S corpus, but in different order. In the first step the literals from the French synsets have been searched for using Intex with French e-dictionaries. The results did not differ much from those demonstrated in Table 9, except that the Serbian literal tip did not occur as an equivalent for literals representing the concept 5, while the literal linija occured again as the equivalent for a literal representing the concept 1. The second step has then been applied to the same corpus, using Serbian e-dictionaries, and literals from SWN, with the addition of literals tip and linija. This step did not yield new candidates: French terms

corresponding to Serbian keywords were, almost without exception those already in specific synset. However, linija appeared twice representing concept 1 ("...trgovacyki brod lepih linija..." translated from "...navire de commerce à hélice, de formes fines...", and "... skladne linije nxenih oblih krsta..." translated from "...l'élégante cambrure de ses reins arrondis..."). The check in Serbian RMS dictionary shows that linija in sense 7 represents concept 1, which is therefore enhanced by the literal linija:7.

This need not always be the case. For instance, the application of the described procedure to the following set of synsets yielded somewhat different result.

1. (operacija:2) ↔ (operation:6) ↔ (opération:2)
2. (operacija:3a) ↔ (operation:1) ↔ (opération:8)
3. (matematicyka operacija:x, operacija:3bx) ↔ (mathematical operation:1, mathematical process:1, operation:5) ↔ (opération mathématique:1, opération)
4. (operacija:3by) ↔ (operation:8) ↔ (opération:7)
5. (operacija:4, postupak:1x) ↔ (operation:3, procedure:1) ↔ (opération:6, procédure:4)

The noun operacija occurred in the Serbian part of E/S subcorpora 4 times (3 times in chosen senses), and in the Serbian part of F/S subcorpora 5 times (in chosen senses 2 times). The equivalents of this Serbian noun were in both English and French subcorpora always in the scope of the associated English and French synsets: moreover, the equivalence was always the most direct one: operacija (Sr.) ↔ operation (Engl.) ↔ opération (Fr.), which is not surprising having in mind their Latin root. The reverse procedure that was looking for all the literal strings from the chosen English and French synsets showed a little more variety. The equivalents of English *operation* were operacija, in senses 2, 3a and 4, and rad, while French *operation* equaled operacija:2, posao, and stvar. English *procedure* occurred once but its equivalent in Serbian is missing, while French *procédure* did not occur at all. In the Serbian part of the English subcorpora the possessive adjective operativni (in the sense of operacija:2) appeared as the equivalent of English *operation* (for instance, "... that could be activated in war operations..." equals to "... brodove koji su mogli da posluxxe za operativne svrhe..."). After this validation process, one synset had to be augmented by new elements, and that was synset 5:

(operacija:4, postupak:1x, rad, posao, stvar)

The check of the hypernym/hyponym tree of synset 5 showed that rad, posao is related to (work:1) which is a hypernym of (operation:3, procedure:1). The noun stvar occurring only once is however, too general to be included in a synset on such a weak evidence. Therefore, this synset remained unchanged.

The results obtained by validation of synsets using the described procedure fully approve the usability of corpora 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 the same PoS, or by setting up a cross-PoS relation [7]. In this respect the corpora approach is particularly useful in detecting derived forms in connection to their senses.

The other useful issue here is the detection of phrases and their translation equivalents. For instance, in the E/S subcorpora the phrase *made up one's mind* occurred twice with Serbian verb equivalent resxiti, while the phrase *change one's mind* occurred six times with Serbian verb equivalent predomisliti se. Both these phrases are already in the English WordNet. However, the English phrase *focus one's mind* is not in the English WordNet, and it appeared twice in the E/S subcorpora with the Serbian verb equivalent koncentrisati se. Also, the English adverbial phrase *in one's own mind* occurred twice with the Serbian adverbial equivalent u sebi – neither of them was in respective wordnets.

Iterating the procedure can further refine the validation process. For example, the whole validation process can be repeated with the literal string tip that was included in the synset 5 of the synset set for the noun oblik, in search for some other possible synonyms.

4.3. Quantification of Literal / Synset Pairing

As already stated in section 4.2. the many-to-many correspondence exists between Serbian literal strings and concepts in the Serbian WN. For the purpose of sense disambiguation of Serbian text it would be useful if the nature of this correspondence could be quantified, in order to establish:

- for a given literal string, what is its most prominent sense;
- for a given concept, what is its most prominent lexicalization.

Intuitively it seems obvious that some senses of a literal string are more frequent, and also, that some literal strings are more often then the other used for the lexicalization of a particular concept. For instance, the occurrence of the literal string oblik in the S/E corpus shows that, at least in this rather small sample, the most prominent sense of this literal string is the one represented by the concept 1 (16 occurrences out of 18), and at the same time, it could be expected that this same concept would be expressed using this same literal string rather than some other possible literal strings (16 out of 20). On the other hand, the most prominent lexicalization of the concept 5 is not oblik but vrsta (26 out of 28). The similar results follow from the occurrence of the literal string oblik in the S/F aligned corpus.

The question is whether this intuitive notion can be more precisely expressed. One possibility to establish the most prominent sense of a literal word is to rely on the ordering of the senses in the Serbian explanatory dictionary RMS. For example, the sense of the literal string oblik that was established on the S/E aligned corpus as the most prominent one is also the first one listed in the RMS dictionary. This, however, is not always the case.

The set of most polysemous literal strings from the SWN were chosen and checked against a subcorpus of contemporary texts from the daily newspaper "Politika" comprising of approximately 1.7 Mw. This set consisted of 34 nouns and 22 verbs. For

each of the chosen literal strings the concordances were produced that enabled manual sense disambiguation of all of its occurrences. The obtained results showed, for instance, that by far the most frequent sense of the noun kraj is 5a (ending:4, conclusion:4, finish:7) (676 out of 715) compared to the sense 1 (end:1 and end:10) (39 out of 715). The same is the case for the verb pokazati whose sense 3 (prove:2, demonstrate:2, establish:3, show:2, shew:1) is more freqent then the sense 1 (uncover:1, bring out:1, unveil:2, reveal:1), that is 243 vs. 25 out of 409. This leads to the conclusion that the sense ordering in the dictionary RMS does not always give the reliable answer to the question: what is the most prominent sense of a literal string?

In order to try to provide a more reliable basis for answering the two posed questions, a set of indices were introduced that were based on the occurrence of the keywords and their senses in the subcorpus:

1. The *overall synset relevance index* of a literal (k) is defined as the ratio of the number of times the literal string has been used in a specific sense (i) and the total number of its occurrences in the corpus, namely: $I_{ik}^{C} = LS_{ik}^{C}/L_{k}^{C}$. This index range is $0 < I_{ik}^{C} \leq 1$, where $I_{ik}^{C} = 1$ means that the literal string L_k is used in one and only one sense, while the value $I_{ik}^{C} = 0$ means that the particular sense did not occur in the subcorpus.

2. Since the SWN is still under development, its coverage of the senses of a literal string is not always complete. Thus, we define the *wordnet synset relevance index* as the relevance of a particular sense of a literal string within a more restricted part of the corpus, that is, the part already covered by the SWN. This index is defined as the ratio of the number of times this literal string has been used in a specific sense and the total number of its occurrences within the corpus denoting concepts represented in SWN, namely: $I_{ik}^{WN} = LS_{ik}^{C}/L_{k}^{WN}$. As is the case with I_{ik}^{C}, the index range is $0 < I_{ik}^{WN} \leq 1$, where $I_{ik}^{WN} = 1$ means that the literal string L_k is used in one and only one sense. Since $L_{k}^{WN} \leq L_{k}^{C}$, then $I_{ik}^{WN} \geq I_{ik}^{C}$. The ideal case $I_{ik}^{WN} = I_{ik}^{C}$ means that all the detected senses of a literal string were already represented in WN, that is $L_{k}^{WN} = L_{k}^{C}$. In order to establish how close a particular litreal string k is to the ideal case when all its possible senses are covered by the wordnet, we compare the number of its occurrences within the corpus denoting concepts represented in the wordnet L_{k}^{WN} to the total number of its occurrences within the corpus L_{k}^{C}. We therefore define the *wordnet coverage index* of a literal string, namely $I_{k}^{WNC} = L_{k}^{WN}/L_{k}^{C}$. The index ranges between 0 and 1, and in case of full coverage it is equal to 1.

3. In order to compare the relevance of a literal to a particular synset in comparison to other literals designating the same concept we define the *local synset relevance index* of the literal string k as the ratio of the number of occurrences of this literal string in the corpus denoting the concept represented by the synset i, and the number of occurrences of all literals denoting this same concept (i.e. belonging to synset S_i): $I_{ik}^{L} = LS_{ik}^{C}/S_{i}^{C}$, where $S_{i}^{C} = \Sigma_{j=1}^{n_i} LS_{ij}^{C}$ represents the frequence of the concept in the corpus measured by the frequency of the literals appearing in the corresponding synset. It should be noted that the range of the

index is $0 < I^L_{ik} \leq 1$ where $I^L_{ik} = 1$, holds when either the synset has only one literal, or other literals from that synset have not appeared in the corpus.

In order to calculate the introduced indices concordances were produced for all the literal strings (we call them *supporting literals*) that occur beside the chosen literal string in the synsets. The main and supporting literal strings form the "lexical sample" as defined by the SENSEVAL project [5]. In Table 10 we present the values of the introduced indices for the literal string oblik on the bases of its occurrences in the subcorpus Politika.

Table 10. The frequency indices for the keyword *oblik* obtained on newspaper corpus

Synset	**oblik**	LS^C_{ik}	**forma:1x**	**vrsta:1ax**	**forma:x**	**forma:y**	**forma:1**	S^C_i	I^C_{ik}	I^{WN}_{ik}	I^L_{ik}
form:7, shape:5, cast:4	1	5	8	*	*	*	*	13	.040	.161	.385
form:3, shape:8, pattern:1	4	4	*	*	*	7	*	11	.031	.129	.364
form:1, word form:1, ...	5	0	*	*	*	*	*	0	.000	.000	*
shape:2, form:6	8a	1	*	*	*	*	4	5	.008	.032	.200
kind:1, sort:1, form:2, ...	3b	21	*	91	19	*	*	131	.167	.677	.160
	L^{WN}_k	31									
	other	95							.753	*	*
	L^C_k	126	65	225	65	65	65				
	I^{WNC}_k	.246	I^C_{ik} .123	I^C_{ik} .404	I^C_{ik} .292	I^L_{ik} .108	I^L_{ik} .062				
			I^L_{ik} .615	I^L_{ik} .364	I^L_{ik} .145	I^L_{ik} .636	I^L_{ik} .800				

The results presented in this table show that the most prominent sense of the literal string oblik is sense 3a (the greatest values of the indices I^C_{ik} and I^{WN}_{ik}). On the other hand, the same string is not the most representative for any of the synsets to which it is assigned, namely, for every synset i its index I^L_{ik} is less than the same index of some supportive literal of the same synset. The ordering of literals among chosen synsets in SWN would thus be:

1. (forma:1x, oblik:1) ↔ (form:7, shape:5, cast:4)
2. (forma:y, oblik:4) ↔ (form:3, shape:8, pattern:1)
3. (oblik:5, oblik recyi:X) ↔ (form:1, word form:1, signifier:1, descriptor:1)
4. (forma:1, oblik:8a) ↔ (shape:2, form:6) ↔ (forme:10)
5. (vrsta:1ax, oblik:3b, forma:x) ↔ (kind:1, sort:1, form:2, variety:5)

The results obtained for the *Politika* corpus significantly differ from those obtained for the same synsets on the S/E aligned corpus. For the literal string oblik, the indicies computed on data obtained on the aligned corpus were $I_{ik}^{WN} = 16/18 = 0.889$ (as compared to 0.161 on the first corpus), and $I_{ik}^{L} = 16/20 = 0.8$ (as compared to 0.385) for synset 1 and $I_{ik}^{L} = 2/28 = 0.071$ (as compared to 0.160) for synset 5. The same thing has been happened with some other literal string (see [8]). It suggests that with the change of the nature of texts the value of these indices changes as well. In this particular example, for instance, the literal forma was first in the literal ordering in most of the synsets, on basis of the data from the *Politika* corpus, which would not be the case if the results obtained on the S/E corpus were used. This literal, with a Latin origin, is not so readily used in literary texts that comprise the largest part of the S/E corpus.

The applied procedure confirmed the importance of the validation of synsets on a corpus. The frequency indices can serve as useful numerical indicators of sense and synset relevance, as well as of the relevance of (literal, synset) pairing. However, to get a fair estimate of a literal in terms of these parameters, the procedure needs to be applied on a large and balanced corpus which would be a very time-consuming task requiring a loto of man-power, especially having in mind both the number of literal strings and and the number of synsets appearing in wordnet. Data represented in Tables 1 and 2 show however that most of the literal strings appear only once in SWN (5 434 out of 6 162 for nouns), and also that most of the synsets have just one literal (2 508 out of 4 562 for nouns). Thus, the value 1 of the index I_{ik}^{C} attached to a literal would suggest that its sense can be unambiguously detected, and that would be the case for 74% of all literal strings represented in the current Serbian WN. On the other hand the value 1 of the index I_{ik}^{L} attached to a literal would suggest that the concept i can be expressed only by using that literal, and that would be the case for the 51% of all the synsets in the current SWN.

The development of the Serbian wordnet is in its initial stage, so it is to be expected that with the addition of new synsets the number of senses per literal string will increase, and it is also to be expected that by addition of new literal strings the number of literals with only one sense will increase as well.

5. Conclusion

The results obtained by validating a number of synsets, presented in this paper, confirm the importance of synsets validation on corpora. Two complementary techniques for synsets validation using both monolingual and bilingual corpora were presented and illustrated.

The first technique is qualitative in its nature and provides for justification of inclusion of a literal string in a synset, as well as for detection of other literal strings that can be added to the synset. It also enables the establishment of relations between various derivatives, setting up a cross-PoS relation if necessary. The procedure can be iterated, further refining the validation process.

The second technique is a quantitative one, based on frequency indices, providing for numerical indicators of how adequately a literal and its sense have been placed

in a particular synset. Still, in order to get reliable indice values for a literal it is necessary to calculate them on a large and balanced corpus. But manual concordance analysis is a highly time-consuming task, which implies the necessity of developing an automatic (or at least semiautomatic) procedure for it.

References

[1] BONHOMME, P., NGUYEN, T.M.H., O'ROURKE, S., *XAlign : l'aligneur de Langue & Dialogue*, 2001, `http://www.loria.fr/equipes/led/outils/ALIGN/align.html`

[2] DANIELSSON, P., RIDINGS, D., *Practical presentation of a "vanilla" aligner*, in *TELRI Workshop on Alignment and Exploitation of Texts*, Institute Jožef Stefan, Ljubljana, 1997, `http://svenska.gu.se/PEDANT/workshop/workshop.html`

[3] FELLBAUM, C. (ed.), *WordNet: An Electronic Lexical Database*, The MIT Press, 1998.

[4] GALE, W.A., CHURCH, K.W., *A program for aligning sentences in bilingual corpora*, Computational Linguistics, **19** (3), 75–102, 1993.

[5] KILGARRIFF, A., ROSENZWEIG, J., *English SENSEVAL: Report and Results*, in *Proc. of LREC*, Athens, May–June 2000.

[6] KRSTEV, C., PAVLOVIĆ-LAŽETIĆ, G., OBRADOVIĆ, I., VITAS, D., *Corpora Issues in Validation of Serbian Wordnet*, in Matoušek, V., Mautner, P. (eds.), *Text, Speech and Dialogue,* LNAI 2807, Springer, 132–137, 2003.

[7] PALA, K., SEDLACEK, R., VEBER, M., *Relations between Inflectional and Derivation Patterns*, in *Proc. of Workshop "Morphological Processing of Slavic languages"*, EACL'03, Budapest, 1–8, 2003.

[8] OBRADOVIĆ, I., KRSTEV, C., PAVLOVIĆ-LAŽETIĆ, VITAS, D., *Corpora Based Validation of WordNet Using Frequency Parameters*, in Sojka, P. et al. (eds.), *Proceedings of the Second International WordNet Conference GWC-2004*, Brno, Czech Republic, Masaryk University, 181–186, 2004.

[9] SILBERZTEIN, Max D., *Le dictionaire électronique et analyse automatique de textes : Le système INTEX*, Paris, Masson, 1993.

[10] STAMOU, S., OFLAZER, K., PALA, K., CHRISTODOULAKIS, D., CRISTEA, D., TUFIŞ, D., KOEVA, S., TOTKOV, G., DUTOIT, D., GRIGORIADOU, M., *BALKANET: A Multilingual Semantic Network for Balkan Languages*, in *1st International Wordnet Conference*, Mysore, India, January 2002, `http://www.ceid.upatras.gr/Balkanet/files/balkanet-elsnet-ko-accept.pdf`

[11] VITAS, D., KRSTEV, C., *Structural derivation and meaning extraction: a comparative study on French-Serbo-Coratian parallel texts*, in Geoff Barnbrook, Pernilla Danielsson, Michaela Mahlberg (Eds.), *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, Birmingham, The University of Birmingham Press [in print], 2003.

[12] VITAS, D., et al., *An Overview of Resources and Basic Tools for Processing of Serbian Written Texts*, in *Proc. of the Workshop on Balkan Language Resources*, 1st Balkan Conference in Informatics, 2003.

[13] VOSSEN, P. (ed.), *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Dordrecht, Kluwer Academic Publishers, 1998.