Resource-based WordNet Augmentation and Enrichment

Ranka Stanković, Miljana Mladenović, Ivan Obradović, Marko Vitas, Cvetana Krstev

Resource based WordNet augmentation and enrichment

Ranka Stanković and **Ivan Obradović**
Faculty of Mining and Geology
University of Belgrade, Serbia
ranka@rgf.bg.ac.rs
ivano@rgf.bg.ac.rs

Miljana Mladenović
College for
Preschool Teachers
Bujanovac, Serbia
ml.miljana@gmail.com

Cvetana Krstev and **Marko Vitas**
Faculty of Philology
University of Belgrade, Serbia
cvetana@matf.bg.ac.rs
vitas.marko@gmail.com

Abstract

In this paper we present an approach to support production of synsets for Serbian WordNet (SerWN) by adjusting Princeton WordNet (PWN) synsets using several bilingual English-Serbian resources. PWN synset definitions were automatically translated and post-edited, if needed, while candidate literals for Serbian synsets were obtained automatically from a list of translational equivalents compiled form bilingual resources. Preliminary results obtained from a set of 1248 selected PWN synsets show that the produced Serbian synsets contain 4024 literals, out of which 2278 were offered by the system we present in this paper, whereas experts added the remaining 1746. Approximately one half of synset definitions obtained automatically were accepted with no or minor corrections. These first results are encouraging, since the efficiency of synset production for SerWN was increased. There is also space for further improvement of this approach to wordnet enrichment.

1. Introduction

Semantic networks, such as wordnets, are among the most important resources in Human Language Technologies. Thus, for example, the Princeton WordNet - PWN (Fellbaum, 1998), has been in use for more than two decades as the standard lexical database for English. Several projects inspired by PWN for the development of wordnets for clusters of other languages have subsequently emerged, such as EuroWordNet - EWN (Vossen, 1998) for a cluster of European languages, and BalkaNet -BWN (Tufis et al., 2004) for a cluster of languages from the Balkans.

The EuroWordnet and BalkaNet projects extended the PWN approach by introducing an Inter-Lingual Index (ILI), used to align synsets in different languages within each cluster on basis of PWN structure. In BWN, the structure of PWN was extended to accommodate concepts specific to Balkan languages that were not found in PWN.

Manual development and maintenance of wordnets has proved to be a demanding task, requiring a considerable amount of human resources. As a consequence, some wordnets developed within the EWN and BWN projects, as well as some others wordnets, developed independently, have been stalled or frozen. In order to overcome this issue, methods for semi-automatic and automatic development and enhancement of wordnets have been proposed.

Even PWN, despite being one of the most comprehensive wordnets, is facing the development problem, as many concepts, especially domain specific, or newly emerging concepts generated by technological developments, need to be added. Targeting specifically PWN WordNet 3.0, SemEval-2016 proposed Task 14 for development of systems for automatic enrichment of taxonomies with new word senses, drawn from other lexicographic resources. The task provided prospective systems with a set of word senses that were not in PWN, where each word sense comprised a lemma, a part of speech, and a definition. The systems were to identify the most plausible point for placing each of the word senses from this set in PWN, either by merging it into an existing synset, or adding it as a new hyponym synset.

Keywords: WordNet, bilingual resources, term alignment, parallel lists

Five teams submitted 13 systems, with all teams performing better than chance, but only one team surpassing a simple baseline, thus pointing towards a need for further efforts aimed at improving taxonomy enrichment (Jurgens and Pilehvar, 2016).

When non-English wordnets are concerned, one of the common approaches to wordnet enhancement is the use of a developed wordnet (most commonly PWN) as the basic source for enrichment, and a method for aligning the synsets of the two wordnets, such as the ILI. Automatic alignment of synsets belonging to different languages is closely related to the task of pairing their word senses. This approach was followed by Matuschek and Gurevych (2013) who solved the word sense alignment (WSA) task by pairing senses with the same meaning from different lexical-semantic resources.

Besides alignment with a developed wordnet, the use of other available resources for development and enrichment of wordnets have also been proposed. Thus, Oliver and Climent (2014) used parallel corpora for five European languages to produce aligned wordnets. The English part of each corpus was semantically tagged, after which the process of wordnet creation was transformed into a word alignment problem, where wordnet synsets in the English part of the corpus were aligned with in the target language part of the corpus. The obtained precision was satisfactory, but the overall number of extracted synset pairs was too low, resulting in poor recall.

In methods used for automatically enriching wordnets using other available lexical resources, the successfulness of the method is strongly correlated with the comprehensiveness of the resource used in the alignment process (Hristea, 2007). Different methods and resources can be used for alignment. One of the common approaches is to take PWN as the source for alignment, and a bilingual dictionary of English and the target language.

There are, however, several other approaches. In (Chugur et al., 2001) a monolingual and a bilingual Spanish-English dictionary were used for the enrichment of the Spanish WordNet with adjective synsets, where the bilingual dictionary was used for alignment with EWN. Bentivogli and Pianta (2003) proposed a method for extending MultiWordNet[1] with phrases, by extracting them from bilingual dictionaries and corpora with techniques similar to those used for collocation extraction. In (Bhingardive et al., 2014) a method for Sanskrit WordNet extension using a Sanskrit English dictionary is presented. The dictionary entries were automatically extracted and linked to PWN. In (Simões et al., 2016) an approach is given to enrichment of Portuguese WordNet with synonyms available in a standard monolingual Portuguese dictionary. Dictionary word senses were automatically identified, annotated and extracted, independently of definitions, and synsets were created. These synsets were then aligned with Portuguese WordNet synsets. The process resulted in both the addition of new synonyms to existing synsets, and in the creation of new synsets.

There were also some interesting approaches related to Slavic languages. Vintar and Fišer (2017) proposed a method to enrich the Slovenian WordNet (sloWNet) with domain-specific single and multiword expressions. They used a large monolingual Slovene corpus of texts to extract terminology from the domain of informatics, and a parallel English-Slovene corpus and an online dictionary as bilingual resources to facilitate the addition of new terms to sloWNet, based on the semantic structure of the PWN. The Croatian Wordnet (CroWN) was enlarged with the WN-Toolkit and CroDeriV, a derivational database for Croatian (Oliver et al., 2015), from 10,026 synsets and 31,367 synset-lemma pairs to 23,137 synsets and 47,931 synset-lemma pairs. The WN-Toolkit, a set of Python programs for wordnet creation and expansion uses a dictionary, Babelnet and parallel-corpora based strategies.

In this paper we present a method to speed up the enrichment of the Serbian WordNet (SerWN) using PWN, several bilingual resources and the Google Translate API. The paper is organized as follows: in Section 2. we present wordnets and bilingual resources that are the base of our solution described in Section 3., while in Section 4. we discuss results of preliminary automatic and manual evaluation. Finally, in Section 5. we give some concluding remarks and plans for future research and implementation of results.

[1] http://multiwordnet.fbk.eu/english/home.php

2. Resources used

2.1. WordNets

Serbian WordNet was initially developed within the scope of the BalkaNet project and subsequently manually enhanced and upgraded. At present it contains 22,530 synsets, out of which 18,248 are nouns, 2,249 verbs, 1,907 adjectives and 126 adverbs. There are 12,248 synsets with only one literal, 7,204 with 2 literals, 2,130 with 3 literals, 612 with 4 literals, 215 with five literals, and 121 synsets with more than five literals. The majority of SerWN synsets are aligned with corresponding PWN 3.0 synsets via the Interlingual Index, with the exception of a little over 1,000 Serbian specific synsets that do not exist in PWN.

In our research we used 20,221 aligned synsets (from a previous version of SerWN), coupled with a number of bilingual resources, to speed up the development of SerWN, by selecting and adjusting PWN synsets from a set of approximately 95,000 PWN synsets without an equivalent in SerWN (further on referred to as non-adjusted PWN synsets).

We analyzed at the outset the number of literals in the 20,221 aligned synsets of PWN and SerWN, and found out that 11,091 synsets (55%) had the same number of literals, that the number of literals in 6,107 synsets (30%) differed by 1, in 1,834 synsets (9%) by 2, in 682 for (3%) by 3 literals, whereas an even greater difference in the number of literals was found in the remaining 507 (3%) synsets. For the 11,091 synsets with same number of literals in English and Serbian the distribution according to the number of literals was as follows: there were 7063 (64%) synsets with one literal, 3288 (30%) with two, 561 (5%) with tree, and 179 (1%) with more that tree literals. As for the remaining synsets with different number of literals, the smaller number of literals was mainly found in SerWN. Namely, the average number of literals per synset in PWN is 2.01, whereas the average number of literals per synset in SerWN is 1.66. After the initial analysis, we used the 20,221 aligned synsets to produce a parallel list of 72,262 literals for the purpose of this research. For example, we used the aligned synset pair:

```
building, edifice -- zgrada, kuća
```

to produce the following list:

```
building -- zgrada
building -- kuća
edifice --  zgrada
edifice -- kuća
```

We have produced another list from non-adjusted PWN synsets, where each item in the list consisted of an ID, a definition and list of literals. The items were then translated using Google Translation API. We used the *LanguageApp* service in Google Apps Script[2] to create our own version of Language Translation API, which, unlike the official Google Language Translation API, produces text translated into Serbian in Latin script, instead of Cyrillic, and serializes it into a plain text file.[3]

An example of a list item is:

```
ENG30-08331011-n | a court with jurisdiction in equity | chancery; court of chancery
```

which is converted, after translation by our Translation API based on Google Language Translation API, into:

```
ENG30-08331011-n | sud koji je nadležan za pravičnost | kancelarija ; sudski ured
```

If a Serbian translational equivalent for a PWN literal obtained by Google Translation is not offered by any other bilingual resource, it is taken into account only if the term is found in the Serbian morphological e-dictionary and if the POS of the term matched the synset POS.

2.2. Bilingual lists

In addition to the list of English-Serbian (en-sr) term pairs extracted from aligned PWN and SerWN synsets, we have used thirteen other bilingual resources to produce additional lists of term pairs, as

[2]Google Apps Script is a scripting language based on JavaScript that provides easy ways to automate tasks across Google products and third party services and build web applications – https://developers.google.com/apps-script/overview

[3]https://cloud.google.com/translate/docs/translating-text

a means to speed up the process of adjusting PWN synsets that still do not have their counterpart in SerWN. A brief description of these resources follows.

Parallel list is a simple bilingual parallel list, developed gradually from various resources and used as an auxiliary resource in WS4LR (later upgraded and dubbed LeXimir), a workstation for lexical resources we have developed (Krstev et al., 2006), to produce SerWN synsets on basis of PWN and to support multilingual queries in English and Serbian. The structure of bilingual parallel lists has in general the following form:

```
TermI[,TermI]*; TermII[,TermII]*
```

where TermI represents a word in one language, and TermII a corresponding word in another. A few examples from the en-sr parallel list are:

```
grandmother;baka,baba
soft drink;bezalkoholno piće
safe,secure;bezbedan,siguran
```

From the en-sr parallel list we produced a simple list of translation pairs to be used in our approach. For example, the four items from the en-sr parallel list listed above generated 9 items in the simple list of translation pairs:

```
grandmother;baka
grandmother;baba
soft drink;bezalkoholno piće
secure;bezbedan
secure;siguran
safe;bezbedan
safe;siguran
```

The Dictionary of Library and Information Sciences[4] is a terminology dictionary that covers the library and information sciences and related disciplines in Serbian, English and German, developed at the National Library of Serbia (Kovačević et al., 2004) (further referred to as Biblio). It contains 12,060 aligned lexical entries, from which a list of 17,761 aligned term pairs was produced.

GeolISSTerm (Stanković et al., 2011) and RudOnto (Kolonja et al., 2016) are bilingual resources developed at the Faculty of Mining and Geology, University of Belgrade (FMG). GeolISSTerm[5] is a thesaurus of geological terms with entries in Serbian and English, containing 4,788 concepts in Serbian and English. From this resource, a list of 6,213 aligned term pairs was produced, both from preferred terms and their synonyms. RudOnto[6] is another complex bilingual terminological resource developed at FMG with the aim of becoming the reference resource in Serbian for mining terminology. It contains 1,041 concepts in Serbian and English, from which a list of 1,273 aligned term pairs was produced, both from preferred terms and their synonyms.

Termi application[7] was also developed at the FMG, with the support of Tempus BAEKTEL project[8]. It has been aimed to support the development of terminological dictionaries in various domains within BAEKTEL. It contains 870 concepts in Serbian and English, which were used to produce 1,290 aligned term pairs.

The structure of lexical entries in GeolISSTerm, Rudonto and Termi is similar. Each term comes with a name, definition, an optional list of synonyms, abbreviations and a bibliographic source. Each term, except the top term in the dictionary tree, has only one hyperonym term, but it can have an arbitrary number of hyponym terms. However, in our approach we have used only the relation between aligned terms and their synonyms to produce aligned term pairs.

Eurovoc[9] is a multilingual, multidisciplinary thesaurus covering the activities of the EU. Serbian is one of the 23 languages managed within this thesaurus. EuroVoc is managed by the Publications

[4] http://rbi.nb.rs/en/dict.html
[5] http://geoliss.mre.gov.rs/recnik/
[6] http://rudonto.rgf.bg.ac.rs/
[7] http://termi.rgf.bg.ac.rs/
[8] http://baektel.eu/
[9] http://eurovoc.europa.eu/

Office, which moved forward to ontology-based thesaurus management and semantic web technologies compliant to W3C recommendations, as well as latest trends in thesaurus standards. For this research we used the bilingual en-sr version 4.7 in xls format, with 6,939 term entries, and 6,971 aligned pairs of terms.

Microsoft language portal[10] has published Microsoft Terminology Collection data in the form of a .tbx (ISO 30042:2008) file containing: Concept ID, Definition, Source term, Source language identifier, Target term, Target language identifier. The number of terms differ from language to language, due to varying levels of localization. The Microsoft Terminology Collection is a set of standard technology terms used across Microsoft products, and comprises 13,147 aligned terms for Serbian.

Additional bilingual lexicons from domains of ecology, psychology, sport and mathematics were compiled from various on-line resources and textbooks and used for generating lists of translational equivalents. Finally, we compiled a list of aligned en-sr terms from the web site of the Serbian Institute for Standardization.

3. Production

The final parallel list of translation equivalents compiled from all of the abovementioned resources had 151,641 different en-sr term pairs, as summarized in Table 1. The first two columns show the bilingual resource the term pairs originated from, and its abbreviation. The numbers of en-sr pairs generated from each bilingual resource are shown in column 3.The total number of different English terms in the final parallel list was 90,563, and column 4 shows the number of pairs, per each resource, where the English word within the pair was found among English literals from non-adjusted PWN synsets, while the last column shows the corresponding percentage. It comes as no surprise that the Parallel list, as a general-domain lexicon has a considerably higher percentage than domain-specific resources. However, when it comes to particular domains, aligned terms from those resources tend to be more relevant. There were 124,887 literals within the non-adjusted PWN synsets, out of which 22,645 had at least one corresponding Serbian literal in the compiled parallel list.

Src	Source	No of pairs	Mapped with PWN	Percentage
L	Parallel list	10542	9071	86
B	Biblio	17761	4296	24
P	Psih	11898	3945	33
K	Eko	4900	2850	58
I	Iss	16423	2560	15
E	Eurovoc	6971	1324	18
M	Microsoft	13147	2398	18
G	Geoliss	6213	701	11
R	Rudonto	1273	376	29
T	Termi	1290	346	26
S	Sport	634	194	30
H	Math	178	90	50
W	WordNet	72262	27600	38
O	Google	290639	290593	99

Table 1: Overview of term-pairs in the final parallel list and their mapping with PWN literals

Each term pair in the final parallel list is accompanied by information about POS (part of speech), number of occurrences (frequency) within all resources as well as information on the resources where it was found, as illustrated by examples in table 2. For example, the term pair `access;pristup` appeared a total of 9 times in six different resources: `BIMPKO`, namely B–Biblio, I–Iss, M–Microsoft, P–Psih, K–Eko, O –Google list.

[10]https://www.microsoft.com/en-us/language/Terminology

POS	English	Serbian	Frequency	Source
N	access	pristup	9	BIMPKO
N	base	baza	7	BILWO
N	base	osnova	6	BILWO
N	contact	kontakt	7	MPRWO
N	content	sadržaj	7	BMPWO
V	fall	padati	4	LWO
V	fall	pasti	4	LWO
V	enrich	obogatiti	3	MLO

Table 2: Examples of en-sr term pairs in the final parallel list

3.1. Feasibility assessment of the approach

In order to assess the feasibility of our approach, we have performed an experiment with the aligned PWN and SerWN synsets. Namely, we have used the compiled bilingual lists, excluding the list generated by the aligned synsets, to find the number of English literals for which the expected corresponding Serbian literal would be offered by the compiled bilingual list. For each synset, the number of literals for which the expected term was offered was recorded. In general, the expected corresponding literal was found for more than 25% of literals from all synsets. A correct corresponding literal for all literals was offered for 3,925 synsets, while 13,724 did not have any offered equivalent. In Figure 1 numbers of synsets with expected corresponding equivalents are shown in comparison to those with no equivalent offered, for synsets with 1, 2, 3, 4 and more than 4 literals, respectively. For example, 11,024 (54%) of all synsets had only one literal, and for 3390 (31%) of these synsets the expected Serbian equivalent was offered. Similarly, 6540 (32%) of synsets had two literals, and 1341 (21%) of them had the expected Serbian equivalent offered, and so on.

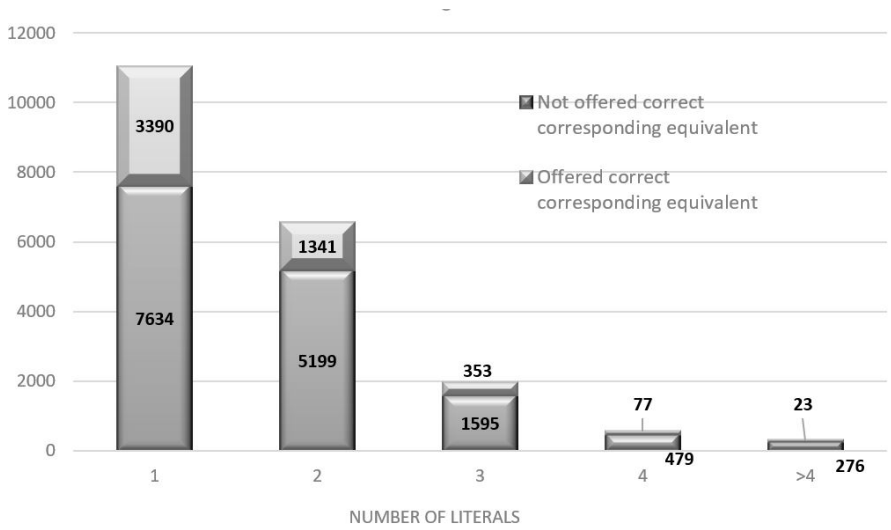

Figure 1: Number of synsets with and without expected corresponding equivalents

We found the results of this experiment encouraging, which motivated us to proceed with the procedure of semi-automatic adaptation of non-adapted PWN synsets, on basis of the final list of parallel terms. Given that the final parallel list was larger than the one we used in our experiment (as it included the list of parallel terms extracted from aligned synsets), we expected to get even better results than the ones in the experiment.

3.2. The semi-automatic adjustment procedure

In order to evaluate our approach for expanding SerWN the final list of parallel terms and a subset of non-adjusted PWN synsets were presented to experts, specialist for specific domains. Their task was to adjust the 1418 PWN synsets (categories) from the so called Base Concepts Set (BCS) that had no equivalents in SerWN. In addition to that, smaller sets of PWN synsets, for which corresponding terms in the parallel list of terms were available, were selected from the following specific domains: computer_science, earth, ecology, engineering, geology, a total of 803 synsets.The experts were expected to perform the following three tasks:

- post–edit Serbian version of synset definition produced by our Translation API;
- select appropriate literals for each Serbian synset from a list of translational equivalents offered by the system;
- add new literals, if adequate candidate(s) could not be found on the list of translational equivalents, but also in other cases if deemed appropriate.

Figure 2 presents an example of the document generated to support experts in their adjustment task. Column `ID` contains the interlingual index, used to bind definition and literals of a specific synset. Column `domain` shows the synset domain from PWN and is used to check whether the synset has been assigned to the appropriate domain expert. In order to reduce false alignments due to homographs, we introduced a constraint to check whether the POS mark for the PWN synset and the POS of the Serbian equivalent offered for a term in that synset are equal, where Serbian POS was available. Column `POS` shows the PWN synset POS, while the last column `SrpPOS` shows the POS for the Serbian equivalent, obtained from Serbian morphological dictionaries (Krstev et al., 2010), if this term was found in the dictionary.

Definition in English `DefEn` is intended to help the expert in correcting the definition in Serbian `DefSr` if needed, as the latter is generated using the Translation API, and is in general subject to post-editing. Pairs of literals from the original English synset and the final list of parallel terms are given in `LiteralsEn`, `LiteralsSr`. In the column `Mark` the expert marks if the offered literal is acceptable. The next column `Frek` shows the frequency of the candidate pair in the final list of parallel terms, whereas the resources in which the term pair was found are displayed in column `Sources`. In general, the higher the frequency and the number of resources, the greater the chance that the offered Serbian term is an appropriate candidate. Each resource, as mentioned earlier, is represented by one letter, e.g. `ML` next to the term pair `update - ažurirati` means that this pair was found in the list obtained from Microsoft Terminology Collection data (M) and the parallel list (L).

If the expert wants to add a new literal, he/she must insert a row, and copy the appropriate ID in the first column of the inserted row, in order to enable further automatic mapping of literals.

ID	domain	POS	DefEn	DefSr	LiteralsEn	LiteralsSr	Mark	Freq	Sources	Srp PC
ENG30-00170857-v	compute r_science	V	bring to the latest state of technology	dovesti do najnovijeg tehnološkog stanja	update;	ažuriranje		0		
ENG30-00170857-v								0		
ENG30-00170857-v					update	ažurirati	da	3	ML	V
ENG30-00170857-v					update	osavremeniti	da	2	LK	V
ENG30-00170857-v					update	ažurirana verzija		1	B	
ENG30-00170857-v					update	modernizovati		1	K	V
ENG30-00200242-v	compute r_science	V	locate and correct errors in a computer	pronaći i ispraviti greške u kodu računarskog programa	debug;	debug		0		
ENG30-00200242-v								0		
ENG30-00200242-v					debug	otkloniti grešku	da	1	M	
ENG30-00200242-v					debug	trebiti		1	P	V

Figure 2: An excerpt from the document to support the adjustment task

3.3. Integration into SerWN

At the beginning of its development, SerWN was maintained using VisDic (Horák and Smrž, 2004), a free tool for editing wordnets in XML format. A more powerful new web tool SWNE[11] was subsequently developed, inheriting all useful characteristics of VisDic and enriched by full XML support, distributed work, and advanced search, as presented in (Mladenović et al., 2014). Further improvements of this tool in developing and maintaining SerWN are presented in (Mladenović and Mitrović, 2014), the main improvement being the replacement of XML by a relational database. The tool provides monitoring of relevant statistics, such as the total number of synsets, the number of semantic relations, quality and speed of enriching SerWN (statistics listed by days and by authors), etc.

The current version of SerWN, thanks to the SWNE tool, integrated external resources – SUMO (Niles and Pease, 2003), SentiWordNet (Baccianella et al., 2010), WordNet Domains (Bentivogli et al., 2004) and PWN v3.0 (Fellbaum, 1998). The official mapping files[12] are used for mapping each synset with these resources. One of the goals of our approach is to further develop the semi-automatic procedure to enable automatic evaluation, periodical analysis and automatic extraction of XML synsets structure for further processing and integration into SWN with sumo, sentiments, relations etc.

In our approach, we used the procedure described in (Mladenović and Mitrović, 2014) for conversion from VisDic XML into MS SQL Server database format, and adjusted it for updating and inserting new synsets into SerWN. While inserting, the procedure checks if there are mapping entries for SUMO, SentiWordNet and Domains resources and generates the appropriate links.

4. Initial results

As the application of the procedure we have developed is a work in progress so far we can report the initial results for 1248 synsets. Results show that in the set of 1248 produced Serbian synsets having a total of 4024 literals, experts accepted 2278 literals offered by the system, and added the remaining 1746.

The average number of English literals in candidate en-sr pairs was 2.1, with the average number of 10.2 literals offered as candidates in Serbian. The average number of Serbian literals in produced Serbian synsets (both selected and added) was 3.2. Figure 3 presents the number of literals selected as correct corresponding literal equivalents for a specific concept, number of English literals in the corresponding PWN synset, and Serbian literals offered by the system. It can be seen that there was a large number of candidates that were rejected, which we plan to reduce by introducing additional constraints.

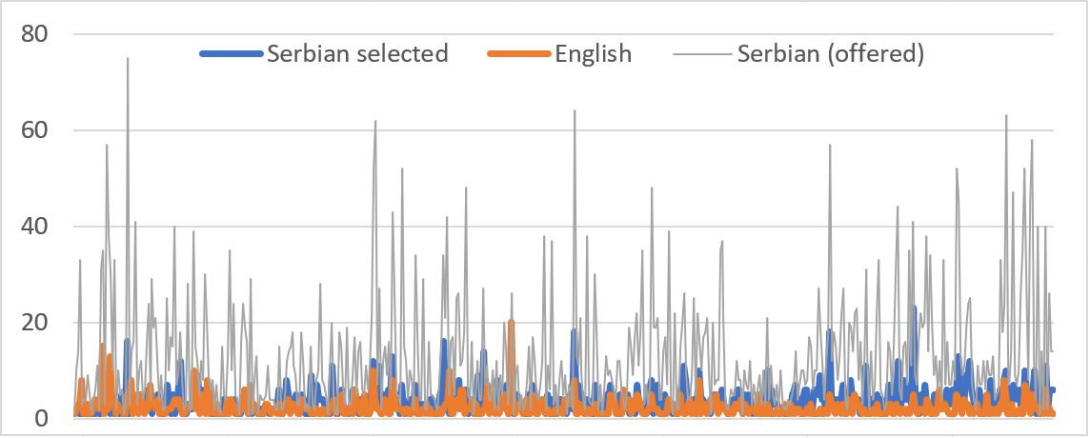

Figure 3: Number of literals per synset

We also evaluated the usability of the Translation API based on Google Language Translation API

[11] http://sm.jerteh.rs

[12] SentiWordNet – http://sentiwordnet.isti.cnr.it; SUMO - https://github.com/ontologyportal/sumo; WordNet Domains – http://wndomains.fbk.eu/; PWN – http://wordnet.princeton.edu/wordnet/download/current-version/

for automatic translation of synset definitions. All definitions for the analyzed synsets, automatically translated from English into Serbian by Google were checked and post-edited, if needed. Semantical similarity between automatically obtained and post-edited translations was measured by Levenshtein distance (LD). As to the number of corrections during post-editing, it was found that 21% of automatic definitions did not have corrections, 12% had correction up to 5 characters, 17% up to 10, etc. It is thus safe to say that for approximately half of the definitions obtained by automatic translation, there were no corrections or corrections were moderate. Only 15% definitions had corrections of over 30 characters (Figure 4).

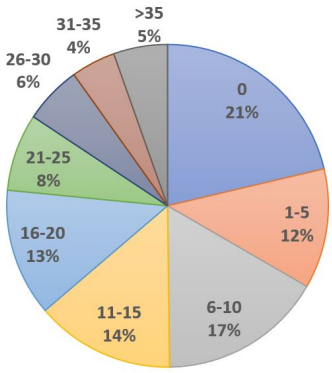

Figure 4: Number of corrections in definition during post-editing phase

5. Concluding remarks and future work

In this paper we presented an approach to speed up WordNet enlargement using bilingual resources and machine translation APIs. PWN synset definitions from a selected set of non-adjusted PWN synsets were translated to Serbian using Google API and manually post-edited. For synset literals, the approach used a compilation of a bilingual list of en-sr pairs of terms, which were aligned with literals from PWN synsets. Since our approach is aimed at semi-automatic production and evaluation of SerWN synsets, the procedure includes obligatory post-editing of automatically prepared datasets. In our first case-study we applied our approach to 1248 synsets, and found out that expert had that 56.6% of literals in produced Serbian synset were literals offered by the system. Also, approximately one half of synset definitions in Serbian offered by the system did not need corrections, or the corrections were moderate. We can thus conclude that the system developed in our approach offers considerable help in augmentation and enrichment of SerWN. However, there are several opportunities for improvement.

As the average number of literals per synset in SerWN (1.66) is about 20% lower than the average number of literals per synset in PWN (2.01), in the following period not only the addition of new synsets, but also the addition of literals to existing synsets will be considered, to ensure a better representation of concepts in SerWN. In addition to that, we plan to broaden the set of parallel resources, and search for new pairs of aligned literals for synsets, which will then be manually post-edited. We also plan to use parallel corpus based methodologies relying on two strategies proposed in ((Oliver et al., 2015)) for automatic construction of the required corpora: by machine translation of sense-tagged corpora and by automatic sense-tagging of English-Serbian parallel corpora. POS tag annotation of bilingual en-sr parallel list is also envisaged, with the aim of increasing precision of offered translational equivalents.

We will investigate possibilities for expanding the information offered to experts by including examples of usage and hypernyms, which would further clarify the concept context. In addition to the Translation API, we plan to include other APIs, like BING API. The bottleneck of the procedure, manual evaluation, will be further supported by crowdsourcing techniques and by the development of a user-friendly web application for post-editing of automatically adapted Serbian synsets. We also plan to include further restrictions using semantic and domain markers from Serbian Serbian morphological dictionaries and other resources, in order to improve precision of system.

Acknowledgements

This research was partially supported by Serbian Ministry of Education and Science under the grants #III 47003 and 178006.

References

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., Eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Bentivogli, L. and Pianta, E. (2003). Beyond lexical units: Enriching wordnets with phrasets. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 67–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. (2004). Revising the wordnet domains hierarchy: Semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, MLR '04, pages 101–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bhingardive, S., Ajotikar, T., Kulkarni, I., Kulkarni, M., and Bhattacharyya, P. (2014). Beyond lexical units: Enriching wordnets with phrasets. In *Global WordNet Conference (GWC)*.

Chugur, I., Peñas, A., Gonzalo, J., and Verdejo, F. (2001). Monolingual and bilingual dictionary approaches to the enrichment of the spanish wordnet with adjectives. In *Carnegie Mellon University, Pittsburgh*.

Fellbaum, C., Ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Horák, A. and Smrž, P. (2004). VisDic – Wordnet browsing and editing tool. In *Proceedings of the GWC 2004*, pages 136–141.

Hristea, F. (2007). Semiautomatic generation of wordnet type synsets and clusters using class methods. an overview. *Revue roumaine de linguistique*, 52:97–133.

Jurgens, D. and Pilehvar, M. T. (2016). Semeval-2016 task 14: Semantic taxonomy enrichment. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1092–1102, San Diego, California, June. Association for Computational Linguistics.

Kolonja, L., Stanković, R., Obradović, I., Kitanović, O., and Cvjetić, A. (2016). Development of terminological resources for expert knowledge: a case study in mining. *Knowledge Management Research & Practice*, 14(4):445–456.

Kovačević, L., Injac, V., and Begenišić, D. (2004). *Bibliotekarski terminološki rečnik: englesko-srpski, srpsko-engleski*. Narodna biblioteka Srbije.

Krstev, C., Stanković, R., Vitas, D., and Obradović, I. (2006). WS4LR: A Workstation for Lexical Resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*, pages 1692–1697.

Krstev, C., Stanković, R., and Vitas, D. (2010). A Description of Morphological Features of Serbian: a Revision using Feature System Declaration. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., and Tapias, D., Eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Matuschek, M. and Gurevych, I. (2013). Dijkstra-WSA: A graph-based approach to word sense alignment. *TACL*, 1:151–164.

Mladenović, M. and Mitrović, J. (2014). *Natural Language Processing for Serbian – Resources and Application*, chapter Semantic Networks for Serbian: New Functionalities of Developing and Maintaining a WordNet Tool. University of Belgrade, Mathematical Faculty.

Mladenović, M., Mitrović, J., and Krstev, C. (2014). Developing and maintaining a wordnet: Procedures and tools. In *Proceedings of the GWC 2014*, pages 55–62.

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*, pages 412–416.

Oliver, A. and Climent, S. (2014). Automatic creation of wordnets from parallel corpora. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., Eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Oliver, A., Šojat, K., and Srebačić, M. (2015). Enlarging the croatian wordnet with wn-toolkit and cro-deriv. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 480–487.

Simões, A., Gómez, X. G., and Almeida, J. J. (2016). Enriching a portuguese wordnet using synonyms from a monolingual dictionary. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., Eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Stanković, R., Trivić, B., Kitanović, O., Blagojević, B., and Nikolić, V. (2011). The Development of the GeolIS-STerm Terminological Dictionary. *INFOtheca*, 12(1):49a–63a.

Tufis, D., Cristea, D., and Stamou, S. (2004). Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.

Vintar, Š. and Fišer, D. (2017). Enriching Slovene wordnet with domain-specific terms. *Annotation, exploitation and evaluation of parallel corpora: TC3 I: TC3 I*, 3.

Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.