sloWNet: construction and corpus annotation

Darja Fišer
Department of Translation, Faculty of Arts
University of Ljubljana, Slovenia
darja.fiser@guest.arnes.si

Tomaž Erjavec
Department of Knowledge Technologies
Jožef Stefan Institute
tomaz.erjavec@ijs.si

Abstract

This paper presents a wordnet for Slovene which was created semi-automatically with a combination of approaches and multilingual resources, in particular a bilingual dictionary, a parallel corpus and Wikipedia. Analysis of the results shows that the dictionary approach yields a good core wordnet but requires substantial manual editing due to a lack of automatic word-sense disambiguation. This was successfully improved with the corpus approach which, however, was limited to single-word literals. The last approach, based on Wikipedia, was only used for domain-specific monosemous terms, and can deal with multi-word literals and therefore usefully complements the previous two approaches. The created sloWNet was then used to semantically annotate a corpus for Slovene: one hundred high frequency nouns were annotated in a corpus of 100,000 words. The paper reports on the method and results of this manual annotation. Both the Slovene wordnet and annotated corpus are to be publicly available.

1 Introduction

sloWNet is a lexico-semantic resource for Slovene, in which words that describe the same concept and therefore have the same meaning (literals) are organized into sets of synonyms (synsets). Synsets are linked into a semantic network with various lexical and semantic relations. Slovene wordnet is based on Princeton WordNet (Fellbaum 1998) and was built automatically following the expand model (Vossen 1998) according to which PWN concepts are rendered in the target language but the relations that hold among those concepts are preserved. Three different approaches and several bi- and multilingual resources were used to generate sloWNet which currently contains about 20,000 synsets and 24,000 literals, 17,000 of which are monosemous. It is aligned to all wordnets for other languages that use PWN synset ids.

The topic of this paper is a project in which frequent nouns from a corpus of Slovene were manually annotated with wordnet senses. The result of the annotation process is a list of concordances in which each nucleus word has an assigned sense called semantic concordances. Semantic concordances are a useful resource for a wide range of applications, such as automatic word sense disambiguation, or for corpus-based studies of sense frequency, distribution and co-occurrence. They are also invaluable as an aid for translation as well as for vocabulary acquisition in a foreign language.

The paper is organized as follows: Section 2 presents the approaches used to construct sloWNet and analyses the results, Section 3 presents manual annotation of a corpus with sloWNet synsets and Section 4 concludes the paper with a discussion and suggestions for future work.

2 Slovene wordnet

For the construction of Slovene wordnet we have leveraged the resources at our disposal, namely a bilingual dictionary, a multilingual parallel corpus and encyclopaedic resources from the Wikipedia family. Based on the assumption that the translation relation is a plausible source of semantics (Dyvik 1998) and that it will reveal words which can have more than one meaning on the one hand and different expressions that share the same meaning on the other, we have used these resources in combination with BalkaNet wordnets (Tufis et al. 2000) to extract semantically relevant information in three different approaches we describe below.

First, we used a bilingual dictionary to translate basic concepts into Slovene. At this stage of the project, our aim was to obtain a core wordnet, which is why we only included synsets from Base Concept Sets (see Tufis et al. 2000). The translations were checked and corrected by hand (see Erjavec and Fišer 2006).

With the second approach we wished to extend the core wordnet as well as to improve automatic disambiguation of polysemous words in order to avoid subsequent extensive manual editing of the generated synsets. A parallel corpus for five languages was word-aligned and the extracted multilingual lexicon was disambiguated with the existing wordnets for these languages from the BalkaNet family (see Fišer 2009). If there was an overlap between all possible synset ids for lexicon entries, the same id was assigned to their Slovene equivalent in the lexicon. All Slovene entries in the lexicon with the same assigned id were treated as synonymous and therefore added to the same synset (e.g. *armada* and *vojska* for *army*). On the other hand, if the same Slovene expression appeared in several lexicon entries and was assigned different synset ids in each case, it was treated as polysemous and therefore added to different synsets (e.g. *stranka1* for *political party* and *stranka2* for *client*).

In the last approach, our goal was to overcome a limitation of the corpus-based approach, which used a 1:1 word-alignment algorithm and could therefore only deal with single-word literals, and to enlarge sloWNet with domain-specific terminology. We used open-source resources, such as Wikipedia and Eurovoc from which we extracted Slovene equivalents for monosemous PWN literals. We also used Wikipedia articles to extract additional synonyms and definitions for synsets that were left in English in the previous approaches (see Fišer and Sagot 2008).

Synsets obtained from all three approaches were merged and filtered according to the reliability of the sources of translations. The structure of PWN synsets for which no translation could be found with any of the approaches was adopted from PWN based on the hierarchy preservation principle (Tufis 2000), only the literals were left empty. These synsets will be translated in the future. The entre network of synsets was then formatted in XML and loaded to the DEBVisDic editor for viewing and editing (Horak 2005).

An example of a Slovene synset with its corresponding English equivalent can be seen in Figure 1. The synset is marked with a Part-of-Speech label, a unique id and a Base-Concept-Set category. The Synonyms field, the most important one in the synset, contains all the literals that are used to describe the concept. They share a common definition which is in most cases still in English at this point of the project. What follows is domain information, mapping to the SUMO/MILO ontology and lexical and semantic relations, such as hypernymy, meronymy and hyponymy. The Stamp field contains information about when the synset was validated and who validated it.

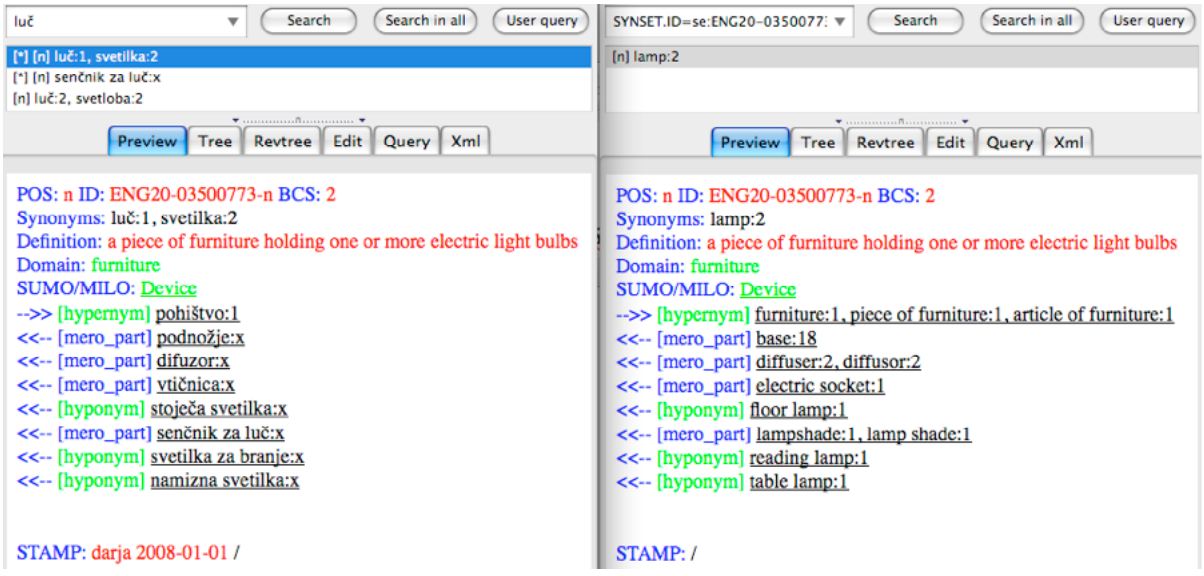

Figure 1. Slovene synset *{luč:1, svetilka:2}* with its English equivalent *{lamp:2}* in DEBVisDic.

2.1 Analysis of sloWNet

The latest version of sloWNet (2.1, 30/09/2009) contains about 24,000 literals, which are organized into almost 20,000 synsets, covering about 15% of PWN 2.1. 17,000 or about 71% of the literals in sloWNet are unique, i.e. appear in only one synset.

Base Concept Sets 1 & 2 are fully covered but there are also many specific synsets. The most frequent domain in sloWNet is Factotum (25%) which was mostly obtained from the dictionary and a parallel corpus while the following three are Zoology (17%), Botany (13%) and Biology (7%) and come from Wikipedia.

sloWNet mostly contains nominal synsets (91%), although there are some verbal and adjectival synsets as well. We have not been able to obtain adverbial synsets with the approaches described above. Apart from single word literals, there are also plenty of multi-word expressions (43%). These too mostly come from Wikipedia. Synsets in sloWNet are relatively short as 66% of them contain only one literal, average synset length being 1.16. The longest synset contains 16 literals (for verb *goljufati*, Eng. *to cheat*).

The most common relation in sloWNet is hypernymy, which represents almost half of all relations in wordnet (46%). Hypernymy is by far the most prevalent relation for nouns (91%). Nominal hypernymy chains tend to be quite long, the longest ones contain as much as 16 synsets. Since sloWNet does not cover the entire inventory of PWN concepts, there are some gaps (empty synsets) in the network. An investigation of nominal hierarchies revealed that all of the nine top nodes exist in Slovene and that almost half (46%) of the chains do not contain a single gap. What is more, only 2% of chains contain five or more gaps. These gaps will have to be filled in the future in order to obtain a denser hierarchy of nodes.

A comparison of nominal synsets from sloWNet and the jos100k corpus showed that sloWNet nouns cover 30% of the common nouns present in jos100k. Most frequent nouns in the corpus (freq. $\geq$ 30) have 91% coverage in sloWNet, medium-frequency nouns (freq. 4-29) have 65% coverage while infrequent nouns (freq. $\leq$ 3) only have 28% coverage in sloWNet (Fišer and Erjavec 2008). While coverage of the most frequent words is good, we will try to improve overall coverage of sloWNet in the future in order to make the resource more useful.

3 Semantic annotation with sloWNet

The main goal of our annotation process was to obtain the first semantically annotated corpus for Slovene which can be used in corpus-based linguistic research as well as a resource for HLT applications requiring training data. However, because sloWNet had been created automatically and had been based on a foreign-language resource, our secondary goal was to check coverage of the senses it contains compared to the senses represented in the corpus and thereby evaluate and improve the developed lexicon in a practical semantic task.

As opposed to sequential annotation, in which all the words in the corpus are annotated, we followed the targeted semantic annotation principle (Miller, et al. 1994) which aims at determining senses only for a selection of polysemous corpus words. Targeted annotation is preferred by many researchers (see Kilgarriff 1998) because this way the semantic characteristics of each word are taken into consideration only once, and the whole corpus achieves greater consistency.

In addition, we followed the joint approach of coordinated wordnet validation and corpus annotation as proposed by Agirre et al. (2006) because it ensures that word senses in the lexicon reflect real usage and guarantees a better fit between sense distinctions in the lexicon and the corpus.

3.1 The jos100k corpus

The corpus used for semantic annotation was jos100k, which is part of the JOS project that is developing annotated corpora and associated resources meant to facilitate developments in human language technologies for the Slovene language. At present, the JOS resources comprise morpho-syntactic specifications, two word-level annotated corpora, and two web services. The developed resources are available under the Creative Commons licenses.

The jos100k corpus (Erjavec and Krek 2008) is a 100,000 word Slovene corpus which contains sampled paragraphs from the Slovene reference corpus FidaPLUS. The corpus is annotated with manually validated morphosyntactic descriptions and lemmas. The corpus has been carefully composed and checked and is intended to serve as a gold-standard reference corpus. In the scope of the JOS project we are currently annotating it for syntactic structures, and for lexico-semantic information, which is the topic of this paper.

3.2 Annotation of the corpus

In the first attempt of semantically annotating Slovene, we limited the task to nouns only because sense assignment for nouns is the easiest and because their coverage in sloWNet is currently by far the best. We extracted 100 most frequent common nouns from the jos100k corpus which also exist in sloWNet. Multi-word expressions that already exist in sloWNet were not included in the annotation as multi-word expressions are typically monosemous and therefore do not require manual sense assignment; it is also harder to automatically identify them in the corpus.

The annotation procedure consisted of several stages: the annotators started from sloWNet in which they checked all the senses of the target word and corrected any errors they found. In the second step, the annotators turned their attention to the concordances and tried to assign a wordnet sense to each occurrence of the given word in the corpus. If they came across a meaning of a word or a phrase they could not find in sloWNet, they added it to wordnet. In the final stage, the annotations were tested for errors and consolidated.

Because no tailor-made annotation software was available, the annotation was performed in MS Excel. Annotators received xls files with the concordances containing the target word that were extracted from the jos100k corpus. After studying an occurrence of the target word in context they determined which synset was the most appropriate for it and annotated it with the corresponding synsed id from wordnet to column C. Any comments that were required for this target word were added to column D. An example of the annotation process can be seen in Figure 2 where a multi-word expression *zemljiška knjiga*/*cadaster* was identified and the appropriate synset id and comment were added for this line in columns C and D.

The goal of the annotation process was to assign a sense to all occurrences of the target words. If more than one sense seemed appropriate despite best efforts to disambiguate them, the annotators were instructed to choose the most basic sense.

If an occurrence of the target word belonged to a multi-word expression (MWE), it was annotated with that sense and marked as a MWE. In case the target word was (part of) a proper name that does not exist in wordnet, the word was flagged as (part of a) proper name. If the appropriate sense could not be found in either sloWNet or PWN, the word was left unannotated and flagged as an out-of-vocabulary item. Most of these senses are language-specific and should therefore be added as such to sloWNet at a later stage of wordnet development.

3.3 Analysis of the annotations

While the annotation is still undergoing some minor revision, we report here on the current state of the semantic concordances over jos100k. Table 1 gives the basic statistics over the annotation set. Each of the 100 nouns has, on average, 54 occurrences in the corpus, which range between 30 (e.g. *oče/father*) and almost 350 (*leto/year*) with the next most frequent being *dan/day* (150). The annotators assigned over 500 different synsets to this set, i.e. over 5 senses per noun. Five of the nouns were monosemous (e.g. *muzej/museum*), while the most polysemous noun annotated was *čas/time* for which a total of 15 senses were used. Finally, almost 50 tokens were proper names, or parts of proper names not present in PWN, and for a further 25 tokens (0.5%) no appropriate synset could be found, e.g. for *voda/water* in *voda na [nekogaršnji] mlin / water for [somebody's] mill*, a Slovene idiom.

tokens	5,384
literals	100
avg tokens/literal	53.8
min tokens/literal	30
max tokens/literal	346
synsets	502
avg synsets/literal	5.4
min synsets/literal	1
max synsets/literal	15
proper names	46
no synset	25

Table 1. Annotation statistics

A	C	D	E	F	G
n	pomen	Opomba	levi kontekst	beseda	desni kontekst
5			aje lenari in prebira	knjige	, predvsem tiste za osebnos
6			prebral prvo takšno	knjigo	, za njo so se zvrstile druge
7	ENG20-06100818-n	*zemljiške knjige	natizacija zemljiške	knjige	
8			ilelfu , da je napisal	knjigo	zgodb , ki so jih opisali kot "

Figure 2. Annotation of the target word *knjiga* (*book*) in MS Excel.

Although MWEs were not explicitly selected for annotation, a surprisingly large number of focus nouns turned out to be part of MWEs which had, or could sensibly have their own literals in the wordnet. Table 2 gives the number of instances tagged as MWEs, almost 10% of the overall tokens, of which almost half had to be annotated with an approximate synset. Altogether, MWEs were tagged with 170 synsets, a third of the overall total.

MWE tokens	471
MWE tokens with approx. synset	223
MWE tokens with approp. synset	248
MWE synsets	170

Table 2. Multi-word expressions

As expected, the complexity of sense assignment to the target nouns correspond to their level of polysemy in sloWNet. On the other hand, it turned out that the lexicon was still missing some senses for nouns which are frequent in the corpus but have very few senses or are even monosemouns in the initial version of sloWNet, which is why these nouns had to be carefully examined as well (e.g. *člen*, freq. 57 appeared in sloWNet only in the sense of *link* but not in the sense of *article in a legal document* or the *grammatical article*).

There were quite many synsets containing the target nouns that were not used by the annotators. There is a good reason for not using some of these synsets because the target nouns appeared in them only due to insufficient disambiguation during wordnet generation and were deleted by the annotators during wordnet revision. An example is the word *sodišče*/*court* which appears in some synsets because the English word court was wrongly translated into Slovene in three synsets:

(1) *a yard wholly or partly surrounded by walls or buildings* – the correct translation is *dvorišče*,
(2) *the sovereign and his advisers who are the governing power of a state* – the correct translation is *dvor* and
(3) *the family and retinue of a sovereign or prince* – the correct translation is *dvor*.

Other senses were not used because they did not appear in the corpus. However, they should not automatically be treated as irrelevant for Slovene because the 100.000 word corpus that was used is far too small for making such conclusions and it would do more harm than good if such senses were deleted from sloWNet at this stage. One such example is the noun *stran* (*page*) which has seven senses in sloWNet, four of which do not appear in the corpus not because they are not used in Slovene at all but because they simply did not appear in our corpus:

(1) *an extended outer surface of an object*,
(2) *a distinct feature or element in a problem*,
(3) *a sheet of any written or printed material (especially in a manuscript or book)* and
(4) *one side of one leaf (of a book or magazine or newspaper or letter etc.) or the written or pictorial matter it contains*.

A comparison of annotations for the same target word that were submitted by two different annotators shows that their annotations vary to a great extent: they chose the same synset id for only 60% of the annotated tokens. It has also been observed that target words differ substantially in the level of agreement between the annotators, which means that some words were much easier to annotate than others. Perfect agreement was reached only with the words that were assigned only one sense (e.g. *odstotek*/*percentage*). Words with a low number of assigned senses (3 or 4, such as *člen*/*article* or *oče*/*father*) have an agreement exceeding 90%. The level of agreement decreases with the increase of target word frequency in the corpus. This suggests that highly frequent and polysemous words were harder to annotate.

As the inter-annotator agreement was rather low, we checked whether annotators agreed on the most frequent sense for a given word. The most predominant sense is very useful for many HLT applications because it has been found that the predominant sense baseline is quite hard to beat by word sense disambiguation algorithms. It turns out that the distribution of senses of the annotated words are in favour of the predominant sense, and that non-predominant senses chosen are in the minority. Also, annotators agreed on the most frequent sense almost in all the cases.

One of the words in which the annotators disagreed even on the most frequent sense is *predstavnik*/*representative* for which the share of the most frequent sense is similar (56.7% and 46.7%) with both annotators but the synsets they used to annotate the most occurrences of this noun in the corpus are different. One annotator most frequently chose the synset *agent: a representative who acts on behalf of other persons or organizations* while the other one preferred the synset *representative: a person who represents others*. When we study both synsets in detail, we find that they are both very similar and it is indeed hard to distinguish between them. This shows that sense distinctions in wordnet are not clear-cut and are very fine-grained, which is a common criticism of the resource as a sense repository for practical applications.

4 Conclusions

The paper presented sloWNet, in particular the method of its construction and the annotation of selected high-frequency nouns with wordnet senses in the jos100k corpus.

The main findings are that automatic methods can lead to reasonably high-quality wordnet construction. The validation of sloWNet with corpus annotations has shown that most senses that were required to annotate the corpus had already been present in sloWNet whereas the same is not true for non-core senses and especially for multi-word expressions which had to be added by the annotators in many cases. Multi-word expressions were especially difficult, as in almost half of the cases no exactly appropriate sense could be found in wordnet. This suggests that sloWNet will have to be further extended in order to ensure a thorough coverage of the sense inventory relevant for Slovene.

Semantic annotation of a corpus, be it manual or automatic, is still one of the challenging annotation tasks. It is very different from e.g. morpho-syntactic annotation in which all the units are annotated with the same set of categories, whereas in determining the meaning of a word, different categories have to be used for each unit we wish to annotate. This is why inter-annotator agreement is typically lower for semantic annotation than other annotation tasks.

In our annotation, we encountered significant problems in determining the best sense for each token, often involving lengthily discussions, and inconclusive decisions.

One way of simplifying and improving the annotation process in the future is collapsing fine-grained hard-to-distinguish senses into more general categories, called supersenses. This had already been done manually by Palmer, Dand and Fellbaum (2007) and automatically by Bruce and Wiebe (1998) who achieved a 10% improvement on the results.

Notwithstanding the difficulties of the annotation, the result is the first Slovene corpus that is annotated at the semantic level. The corpus will be freely available for linguistic analysis or as a training set for applications in human language technologies on the project website: http://nl.ijs.si/jos/index-en.html, while sloWNet is already publicly available via the Creative Commons license: http://nl.ijs.si/slownet.

Acknowledgments

The work was supported in part by the Slovenian Research Agency grant J2-9180 “Linguistic annotation of Slovene language: methods and resources” and by the EU 6FP-033917 project SMART “Statistical Multilingual Analysis for Retrieval and Translation”.

References

Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K. & Quintian, M. (2006). A methodology for the joint development of the Basque WordNet and SemCor. Proceedings of LREC'06. Genoa.

Bruce, R., & Wiebe, J. M. (1998). Word sense distinguishability and inter-coder agreement. Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (pp. 53–60). Granada.

Dyvik, H. (1998): Translations as semantic mirrors. In: Proceedings of Workshop W13: Multilinguality in the lexicon II of the 13th biennial European Conference on Artyificial Intelligence, ECAI 1998, pp. 24-44, Brighton, Gret Britain.

Erjavec, T., & Krek, S. (2008). The JOS Morphosyntactically Tagged Corpus of Slovene. Proceedings of LREC'08. Marrakech.

Erjavec, T. & Fišer, D. (2006): Building Slovene WordNet, In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006. Genova, Italy, 24.-26. May 2006.

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

Fišer, D., & Erjavec, T. (2008). Predstavitev in analiza slovenskega wordneta. Proceedings of IS-LTC'08 (pp. 37–42). Ljubljana.

Fišer, D., & Sagot, B. (2008). Combining Multiple Resources to Build Reliable Wordnets. Proceedings of TSD'08. Brno.

Fišer, D. (2009). Laveraging parallel corpora and existing wordnets for automatic construction of the Slovene wordnet. In: Human language technology: challenges of the information society, (LNCS 5603). Berlin; Heidelberg: Springer, pp. 359-368.

Horak, A., Pala, K., Rambousek, A., & Povolny, M. (2005). DEBVisDic - First Version of New Client-Server Wordnet Browsing and Editing Tool. Proceedings of the GWA'05 (pp. 325–328). Brno.

Kilgarriff, A. (1998). Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs. Computer Speech and Language. Special Use on Evaluation , 12 (4), 453–472.

Mihalcea, R., Chklovski, T., & Kilgarriff, A. (2004). The Senseval-3 English lexical sample task. Proceedings of ACL/SIGLEX Senseval-3.

Miller, G. A., Chodorow, M., Landes, S., Leacock, C., & Thomas, R. G. (1994). Using a semantic concordance for sense identification. Proceedings of the workshop on Human Language Technology. Plainsboro, NJ.

Palmer, M., Dand, H. T., & Fellbaum, C. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. Natural Language Engineering (13), 137–163.

Tufis, Dan; Dan Cristea in Sofia Stamou (2000): BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. V: Dascalu, Dan (ur.): Romanian Journal of Information Science and Technology Special Issue. 7/1-2, 9-43.

Veronis, J. (1998). A study of polysemy judgements and inter-annotator agreement. Programme and advanced papers of the Senseval workshop. Herstmonceux Castle.