Generating Sets of Synonyms between Languages

Ondrej Dzurjuv[1], Ján Genči[1]
and Radovan Garabík[2]

[1] Department of Computers and Informatics, Technical University of Košice
[2] Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava

Abstract. Princeton WordNet is a lexical database that contain sets of synonyms for the English language together with their semantic relationship. In this paper we explore several methods of generating synsets in another language by using English language WordNet and a bilingual dictionary. The methods have been used to generate Slovak language synsets and to bootstrap a Slovak WordNet database.

1 Introduction

Communication is one of the most important things in the social life of every person. It is important to understand other people, learn or change experience. Nowadays, people travel all over the world and learn from books written in foreign languages, so the presentation of knowledge in more languages becomes a common fact. But learning new words is not enough to translate a sentence properly. The key is to know the meanings of a word, how to use it and where to use it.

There are many projects that offer multilingual synonym dictionaries with semantics included like BalkaNet [1], EuroWordNet [2] and Global WordNet [3]. All these projects have something in common. They are based on English Princeton WordNet [4]. English WordNet is a lexical database that connects sets of English synonyms into a semantic net. Chapter one of the paper deals with it in detail. Every project mentioned above connects similar meanings for several languages, but not for Slovak. This paper describes some methods on how to automatically generate Slovak synonyms using the WordNet database and online Slovak-English dictionaries. Generating groups of Slovak synonyms from English ones ensures that Slovak synonym sets will be properly connected to the English equivalents. The article contains statistics records about the quantity of results and an experiment that shows how a quantity can be influenced by the usability of words in an English synset.

Finally, after generating, a small Slovak WordNet was created by using generated Slovak synsets and English WordNet as a pivot. This paper also describes an approach to the building of Slovak WordNet.

The presented work is based on two previous projects [5] and [6]. The first one presents proof of concepts for the automation of Slovak synset generation by using online dictionaries and the second one focuses on building equivalent bilingual synsets from dictionary items only.

2 Introduction of WordNet

WordNet is a lexical database of English synonyms containing nouns, adjectives, verbs and adverbs. The development began in 1990 at Princeton University. WordNet has two main characteristics:

- Words with the same meaning are grouped into synsets – sets of synonyms.
- Synsets are connected by relations and create a synonymic net of synsets.

Today, the WordNet database is at version 3.0 and contains more than 117 000 synsets. Table 1 shows some statistical information about the WordNet database.

Part of speech	**Nouns**	**Adjectives**	**Verbs**	**Adverbs**	**Totals**
Unique strings	117798	21497	11529	4481	155287
Strings with one sence	101863	16503	6277	3748	128391
Word-meaning pairs	146312	30002	25047	5580	206941
Synsets	82115	18156	13767	3621	117659
Synsets of one word	42054	11353	8041	2400	63848

Table 1. WordNet 3.0 statistics

The basic relations between synsets in WordNet are:
Synonymy – is a relation between literals (synonyms) in one synset.
Hyponymy – is a relation of sense specification between synsets, a relation heads from general synset to a more specific synset (motor vehicle → car, automobile).
Hypernymy – is a relation of sense generalization between synsets, a relation heads from a specific synset to a more general synset (motor vehicle ← car, automobile).
Meronymy – is a relation between a term denoting the part and a term denoting the whole, leading from the whole to its part (car, automobile → engine).
Holonymy – is a relation between a term denoting the part and a term denoting the whole, (car, automobile ← engine).

3 Generating Slovak synsets

3.1 WordNet and Slovak-English dictionary

The process of building Slovak synonym sets uses the WordNet database as a source of English synsets. The most important WordNet relations used for generating Slovak synonym sets are synonymy, hypernymy and holonymy, where synonymy is the equivalence in meaning (for words in the same synonym set). Hypernymy and holonymy are described in section 2. Generated Slovak synsets are mapped to their English equivalents, so after the process of building Slovak synsets, WordNet relations should be valid also between Slovak synsets.

The second very important source of data is a good quality electronic English-Slovak dictionary, for example an online dictionary which is used for search of Slovak synsets. The size of a translator's database and quality of its translations are very important for the quantity and quality of created groups of Slovak synonyms.

3.2 Methods for generating Slovak synonyms

Method A

This method uses a synonym relation between words in one synset. When we translate English synonyms, we can expect that some translations will contain the same words. So the words that are in two or more translations constitute a synset in the Slovak language.

For example:
We have the English synset *{kind; sort; form; variety}*, which means "a category of things distinguished by some common characteristic or quality".
After translating using English-Slovak dictionary [7], we get these groups of Slovak words:

kind – druh, rod, kategória
sort – druh, akosť, trieda, typ, forma, chlap
form – forma, tvar, podoba, formulár, blanketa, formula
variety – rozmanitosť, odroda, výber, druh, rad, množstvo, mnohotvárnosť, rôznosť

After intersecting all the pairs of translations, the final group *{druh, forma}* represents the Slovak equivalent of the English synset *{kind; sort; form; variety}*.

Advantages of this method: sense accuracy

Disadvantages of this method: empty Slovak synsets for English synsets consisting of one word

Method B

The next method is based on an idea that English words with one sense should be translated into one group of synonyms. If a synset contains more words with one sense, these words should have similar translations. A slovak synset will be created by the union of these translations.

The synset *{kind; sort; form; variety}* in the previous example contains one univocal word:

kind – 1 sense, translation: druh, rod, kategória
sort – 4 senses
form – 16 senses
variety – 6 senses

After the translation of all univocal words and the union off all these translations we get the Slovak synset: *{druh, rod, kategória}*.

Advantages of this method: possibility of creating Slovak synsets from English synsets consisting of one word (this word must have only one sense in WordNet)

Disadvantages of this method: quality of Slovak synsets depends on translation accuracy, univocal words in English can have more senses in Slovak; empty synsets for English synsets with no univocal word

Method C

Method C alsoises hypernym and hyponym synsets in addition to the default English synset. There are small differences in sense between some English synsets that are in a hypernymic or hyponymic relationship. It is expected that after translation some words will be the same for more synsets. In this method two groups of words are created. A group of words belonging to the default synset and a group created from all its hypernyms and hyponyms. The next step is to translate these groups and then to intersect them.

For example, we have a synset *{kind; sort; form; variety}* (group 1). Its hypernyms and hyponyms will create one group: *{category, type, brand, genus, species}* (group 2).
Translated groups:

Group 1 – druh, rod, kategória, akosť, trieda, typ, forma, chlap, tvar, podoba, formulár, blanketa, formula, rozmanitosť, odroda, výber, rad, množstvo, mnohotvárnosť, rôznosť

Group 2 – kategória, skupina, trieda, typ, symbol, litera, druh, odroda, značka, označenie, známka, kvalita, akosť, ohorok, rod, forma, tvar

The final Slovak synset will be: *{druh; rod; kategória; akosť; trieda; typ; forma; tvar; odroda}*

Advantages of this method: possibility of creating Slovak synsets from English synsets consisting of one word; quantity

Disadvantages of this method: lower quality of Slovak synsets

Method D

This method is a modification and extension of method C. It is also based on small differences in sense between WordNet synsets. This method doesn't use a synset that we are using for generating its Slovak equivalent. The aim is to create an intersection between the translation of its hypernym synset and the translation of its holonym synsets, so ultimately the synset equal to the default English synset should be created. At first, two groups are created: the first group represents the hypernym synset (more general), the second group represents the union of all hyponym synsets (more specific). The Slovak synsets are created by translating and intersecting these groups.

For the synset *{kind; sort; form; variety}*:

Hypernym group: *{category}*

Hyponym group: *{type, brand, genus, species}*

Translated groups are:
Hypernym group: kategória, skupina, trieda
Hyponym group: typ, symbol, litera, druh, odroda, značka, označenie, známka, kvalita, akosť, ohorok, druh, rod, skupina, trieda, forma, tvar

The final Slovak synset is: *{skupina, trieda}*

Advantages of this method: possibility of creating Slovak synsets from English synsets consisting of one word

Disadvantages of this method: small quantity; lower quality of Slovak synsets

4 Statistics of results

All previous described methods were used for generating Slovak synonym sets. The whole process was divided into four steps:

1. translation of English words from WordNet
2. using methods A-D to build Slovak synsets from translations according to English words in English synsets
3. additional correction of created synsets (removing words with duplicate entries and words with incorrect parts of speech)
4. storing a new synset with reference to an English equivalent

After the process of generation, statistics of the results were created to evaluate the reliability of automatic generation for all presented methods.

4.1 Complete Results

The next table shows results for an attempt to generate Slovak synsets for a complete WordNet database.

	Totals	**Nouns**	**Adjectives**	**Verbs**	**Adverbs**
Synsets in WN	117659	82115	18156	13767	3621
Total EN synsets with Slovak synset	40521 (34.4%)	26787 (32.6%)	6859 (37.8%)	5839 (42.4%)	1036 (28.6%)
Method A	10267 (8.7%)	5705	2175	2109	278
Method B	30243 (25.7%)	20510	6059	2715	959
Method C	11533 (12%)	8192	-	3341	-
Method D	1917 (1.4%)	1348	-	569	-

Table 2. Comparison of synset generation methods

Slovak synsets were generated for 34% of all English synonym sets in WordNet. Method A generated less than 9%. It is because most of the synsets in WordNet contain only one word. A high amount of univocal words in WordNet caused the generation of more than 25% of synsets with method B. Method D generated much fewer synsets than other methods so it is not as effective as expected.

Adjective and adverb synsets have not a hypernymy and hyponymy relationship between them so methods C and D could not be used in this case.

4.2 Experiment

Low quantity of generation was caused by more factors:

- many words could not be translated because there was no translation in the dictionary for them
- most English synsets consist of one word
- absence of some relations used for generating

We created an experiment to get some statistical information for commonly used words. We created a group of 300 words randomly selected from the 5000 most used words in English [8]. Then we found English synsets containing these words (1709 synsets). The next table shows data from this sample of commonly used words.

	Totals	**Nouns**	**Adjectives**	**Verbs**	**Adverbs**
Synsets in a sample	1709	769	397	429	114
Total EN synsets with a Slovak synset	946 (55.4%)	491 (63.9%)	171 (43.1%)	237 (55.2%)	47 (42.2%)
Method A	559 (32,7%)	255	126	145	33
Method B	404 (23.6%)	201	97	67	39
Method C	505 (29.6%)	337	-	168	-
Method D	112 (6.6%)	67	-	45	-

Table 3. Statistics of sample synsets generation

There are some important numbers in this table in comparison to the results in table 2. Generating common synsets is much more successful. 32% of generated synsets are by method A which is the biggest increase out of all methods. Method B is almost at the same value which could be caused by the balanced location of univocal words in WordNet synsets.

5 Building the Slovak WordNet

5.1 Automatic synset building

The approach described above has been used to bootstrap a basic Slovak-English-German-Polish-Lithuanian dictionary[1]. The Slovak-English synset pairs have been generated as described before, the other languages have been pre-filled from other sources and then manually proofread. The Slovak part of the structure then server as a base for a small Slovak language WordNet.

We selected the ten thousand most frequent words from the Slovak National Corpus (balanced subcorpus prim-4.0-vyv). We then generated synsets for each of the noun, verb, adjective and adverb categories of these words. A web-based application is used to further edit the generated synsets and their relation to the English synsets. The application allows for general M:N mapping between English and other synsets – the English WordNet serves as a pivot language in the dictionary, even if the external appearance will be that of a *Slovak→other language* one. An additional link can be specified between synsets in other languages and Slovak synsets inside a set of synsets linked to the same English synset. This is used in job titles or animal nomenclature, where the (usually) gender-neutral English noun has two Slovak synsets assigned, one masculine[2] and one feminine. German, Polish and Lithuanian nouns (which mostly keep the same distinction as Slovak) are then linked with the corresponding Slovak synset.

5.2 Synset structure

Each synset has an optional gloss in its own language (parallels the English WordNet structure) – used only if further explanation or refining of the sense is desired.
There are several possible marks applied in the (non English) synset description:

- One or several constituent words in the synset can be marked as “major”, giving it a distinct visual realization in the final dictionary version.
- The whole synset can be marked as “imprecise”. This is used in cases where there is no direct semantic equivalent to the English synset, but the synset had to be filled in, most likely because it was a hypernym of other existing synset(s). This is mostly present in concepts that are realized in other languages as phrases or descriptions (e.g. the English noun *uxoriousness* has no Slovak language equivalent as a noun describing the trait – the meaning combines two rather different concepts, the verbal construction *byť pod papučou* and a dative noun phrase *oddanosť manželke*)
- Individual words in the synset can be marked as “unsure”. This is purely a temporary measure for the editor to record that he or she was unsure about the equivalence or meaning and the synset has to be re-checked later.

[1] Sponsored by the Slovak Online (Lifelong Learning Programme 504873-2009-LLP-SK-KA2-KA2MP) project.

[2] Strictly speaking, a Slovak masculine noun should be assigned into two different synsets, a general one encompassing both genders (or gender agnostic) and a strictly masculine hyponym. However, we considered this distinction too detailed for the purpose of the database.

Additionally, a synset in the database can be marked as "checked" (by an independent reviewer).

5.3 Verbs

Links between other parts of speech are straightforward; there are only a few isolated cases where the situation is more complicated (such as the inclusion of numerals as nouns, or English adverbs whose Slovak equivalents are classified as particles). On the other hand, verbs are more complicated. Features that deserver special care are negation, aspect and reflexivity.

Verb negation in Slovak is accomplished (with very few exceptions) by prefixing the verb with *ne-*, which is then seen as a separate, derived verb. We included the most frequent negative lexemes in the database if there was a corresponding English synset (e.g. *disagree↔nesúhlasiť);* for all other verbs, we have only the affirmative form.

Verb aspect in Slovak is mostly inherent in the lexical level – verbs can be either perfective, imperfective, or ambivalent (which is in fact just the conflation of both aspects into one lexeme), although ways of deriving the opposite aspect exist, such as prefixes turning an imperfective verb into the perfective and morphology root changes to turn a perfective verb into an imperfective one. In the database, we treat perfective/imperfective verb pairs as separate lexemes and assign them to separate synsets that are linked to the same English synset (unless there is a different English synset for the opposite aspect). The presence of both perfective and imperfective verbs inside one synset is prohibited and is automatically enforced by comparing the synsets entered against a list of perfective and imperfective verbs respectively. We do not include Slovak verbs that are only formally derived from the opposite aspect and are not used reasonably frequently in the language. In particular, frequentative/habitual verbs can be derived almost mechanically, but only the frequently used ones are included in the database.

Verb reflexivity is realized with special reflexive pronouns *sa*, *si* that are considered part of the lexeme, although they are written separately from the verb proper and their position in the sentence varies and can be quite remote from the verb itself. If there is a Slovak reflexive/non-reflexive verb pair and the meaning of both of the verbs corresponds to one English synset, both the reflexive and non-reflexive verbs are assigned to two different synsets linked to the same English synset, often separately for perfective and imperfective aspects (therefore producing in some cases four different Slovak synsets linked to the same English one).

Part of speech	**Nouns**	**Adjectives**	**Verbs**	**Adverbs**	**Totals**
Unique strings	12941	3321	1150	982	18394
Strings with one sense	10239	2305	953	702	14199
Word-sense pairs	18740	5551	1400	1505	27196
Synsets	9317	2329	830	549	13025
Synsets of one word	3916	773	426	141	5256

Table 4. Slovak WordNet statistics (at the time of writing)

6 Conclusion

It is clear that it is not possible to generate synsets for the whole WordNet database. Quality and quantity of Slovak synsets correspond to the usability of words in real life. The main problem is that there is no translation in dictionaries for some/many words in the WordNet database.

We used four different methods for generating Slovak synsets and each of them have their advantages and disadvantages, and some Slovak synsets were produced from English ones by more than one method and it is not possible to select the best one automatically. It is also important to balance the quality and quantity of results. For example: Method D uses and idea too complex to find Slovak synsets and the number of results is very low. Also, by joining two or more techniques (described in this article) together we can achieve better results. The output from all methods covered 34.4% of all English synsets.

Generated data were used to bootstrap a Slovak WordNet database. Some generated synsets had an incorrect sense or contained words with the wrong part of speech, but the synsets were manually checked and corrected if needed. A web-based application was created to simplify the whole process of building Slovak WordNet. This application was also used to create a basic Slovak-English-German-Polish-Lithuanian dictionary.

References

[1] Stamou S., Kemal O., Pala K., Christoudoulakis D., Cristea D., Tufiş D., Koeva S.,Totkov G., Dutoit D., and Grigoriadou M. BALKANET: A multilingual semantic network for the Balkan languages. In Proceedings of the International Wordnet Conference, pp. 12 –14, Mysore. 2002.

[2] EuroWordNet. http://www.illc.uva.nl/EuroWordNet/

[3] The Global WordNet Association. http://www.globalwordnet.org/

[4] WordNet – a lexical database for the English language. http://www.wordnet.princenton.edu/

[5] Lapoš P.: Verification of possibility to build EuroWordNet synsets based on on-line dictionaries. Diploma work. KPI FEI TU Košice. 2005.

[6] Sudynová M.: Generating tool for dictionary records. Diploma work. KPI FEI TU Košice. 2006.

[7] Slovník.sk: http://www.slovnik.sk/

[8] Word Frequency List: http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Project_Gutenberg_1-10000

[9] Dzurjuv O.: Computational linguistics – Generating sets of synonyms between languages. Semestral project. KPI FEI TU Košice. 2010.

[10] Dzurjuv O.: Computational linguistics – Generating sets of synonyms between languages. Diploma project. KPI FEI TU Košice. 2010.