From Multilingual Dictionary to Lithuanian WordNet

Radovan Garabík and Indrė Pileckytė

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

Abstract. In this paper we describe the motivation for building a small Lithuanian language WordNet out of a bilingual dictionary. The WordNet contains semantic relations for nouns, adjectives, adverbs and verbs, with mapping of synsets to the Princeton WordNet and the Slovak language WordNet. The process of combining various dictionaries to create an initial version and subsequent manual proofreading is described; the first preliminary version of the WordNet has been released.

1 Introduction

For the English language, the Princeton WordNet [7] became de facto a foundation for semantic analysis and annotation, and it inspired WordNets in many other languages. Such projects frequently use the same structure of semantic relations (often augmented by relations specific for the given language).

Considering their interconnections, WordNet projects can be divided into two groups: monolingual and multilingual ones. Monolingual WordNets capture semantic relationship for the given language without any out-of-the language links, while multilingual ones have some way of relating their synsets to other WordNets (most likely the English one).

This relation can form the main design principle of a WordNet, sometimes even to the point of the WordNet being created as a translation of the English one [6]; or the relations are included additionally [9]. There are several different projects that specifically aim to provide multilingual synonym databases, like BalkaNet [10], EuroWordNet [11], or WordNet Grid [8].

The described Lithuanian WordNet database started its life as a (small) multilingual dictionary[1] for students of Slovak as a foreign language.[2] The dictionary used Slovak as the pilot language, and the English part of it has been based on Princeton WordNet v. 3.0. Our database has been therefore created with the semantic relations in mind, and later we specifically extended the contents with the aim of creating bases for Slovak and Lithuanian WordNets.

2 Automatic Synset Generation

The database has been bootstrapped by an automatic synset generation. The method used for the Slovak synsets is described in [3] – in a nutshell, the method is based on translating synsets, hypernyms and hyponyms according to an existing bilingual dictionary and then taking an intersection of various combinations of the translations. The initial database

[1] The term 'dictionary' is perhaps a little ambitious, 'glossary' would be more appropriate.

[2] The dictionary also includes other languages, in particular German and Polish, but since they are not germane to the Lithuanian WordNet, we will not describe them here.

has been filled with Slovak synsets generated by a union of all the four methods (A, B, C, D) described therein. This database has been then manually proofread and extended, with the synsets being mapped to their equivalent English synsets, with the aim to cover (as a minimum) hypernyms for each Slovak synset – thus creating a complete semantic chains up to the top-level categories.

Since we lacked a computer readable English-Lithuanian dictionary, the Lithuanian part of the database has been generated differently – first we obtained a rough Slovak-Lithuanian dictionary based on Slovak-Esperanto and Esperanto-Lithuanian dictionaries provided by the `lernu.net`[3] portal. Using Esperanto as a pivot language had several advantages:

- Word suffixes in Esperanto denote unambiguously part of speech, therefore we obtained highly reliable separation of synsets into nouns, adjectives, adverbs and verbs.
- There is a very low amount of homonymy (although it does exist) in Esperanto [4], which limits the risk of carrying improper semantic chain into a given synset.
- Bilingual Esperanto-Slovak and Esperanto-Lithuanian dictionaries were available and we obtained a copyright agreement allowing us to use them for this purpose.

The dictionary entry consisted of one Esperanto word and its one or several translations. The size of Esperanto-Lithuanian dictionary was 11 529 entries or 16 268 words, Esperanto-Slovak 7 116 entries or 8 130 words. By combining the dictionaries, we obtained a simple Slovak-Lithuanian dictionary of 3 977 entries (one entry corresponds to one Slovak word and its possible Lithuanian translations), or 10 048 Lithuanian words – we can see that there was a substantial ambiguity in the translations.

The dictionary has been then manually proofread and corrected, with the emphasis on keeping 'precision' – i. e. the proofreaders were instructed to predominantly delete incorrect translations, in order to keep down the time needed to complete the task.

This proofread dictionary has been then used to automatically assign Lithuanian synsets to the Slovak ones via a simple substitution of Slovak literals with Lithuanian equivalents.

3 Database Structure

One entry in the database corresponds to one synset. In addition to the synset itself it contains optional definition (not used much), a link to one (or more) English synsets and an optional links to one or several Slovak language synsets. Generally, the relations in the database are $L{:}\,M{:}\,N$, where L is the number of English language synsets, M the number of Slovak language synsets and N that of Lithuanian language synsets – i. e. any number of synsets from any of the languages can be connected to any other number of synsets in the other languages, although in practise the relation is usually split into $L{:}\,M$ where $L = 1 \vee M = 1$; $M{:}\,N$ where $M = 1 \vee N = 1$; $L{:}\,M$ where $L = 1 \vee N = 1$; that is, we try to refrain from introducing complicated and hard to read connections and try to use simple, at most one-to-many relations between two languages. However, most of the entries are simple one-to-one.

[3] `http://lernu.net/`

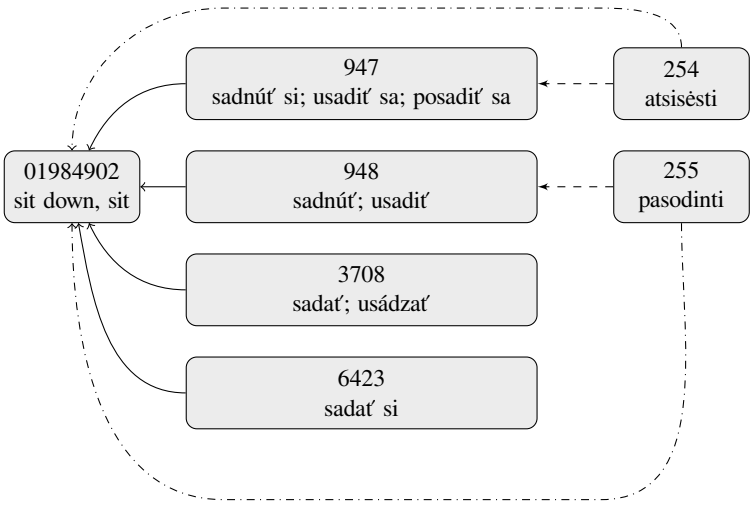

Fig. 1. Example of the interconnection between synsets. The four Slovak synsets correspond to reflexive (intransitive) perfective, transitive perfective, transitive imperfective and reflexive (intransitive) imperfective. Two Lithuanian lexemes for transitive and intransitive are linked to respective Slovak synsets.

3.1 Synset Microformat

The synsets in our database (both Lithuanian and Slovak) are written down using formalized formatting rules in order to ease further automated use and to include additional information (see Figure 2 for the syntax diagram). To put it informally, each synset consists of one or several literals separated by a semicolon; this allows us to include embedded clauses separated by a colon and a relative pronoun (such use is discouraged, but it is necessary to cover those English synsets that do not have direct Slovak or Lithuanian equivalents). Thus the literals can be multiword, simple two-word constructions (adjective+noun) are quite common.

A literal can have an optional annotation character '+' in front of it, this denotes that the literal is semantically 'most important' in the synset, i. e. this is *the* word that is usually used to express the meaning. Another optional annotation is formed by an optional gloss in parentheses, explaining or clarifying the literal in case its inclusion in the synset not obvious to the user, usually in the case of surprising homonymy or a rarely used meaning.

There are also two synset-wide annotations – a minus character and a question mark. Minus character in front of the synset means that the linked Princeton WordNet synset cannot be expressed clearly in the target language (i. e. the semantic meaning is too wide or too narrow, or it covers specific English-culture term that does not have a direct equivalent, or – rarely – there is an outright semantic lacuna in one language). This appears almost exclusively when trying to cover hypernyms of an already existing synset.

A question mark means that the annotator is not sure about the synset – either the synset itself, or its relation to other languages. In theory, this means that we should try to resolve the problems later and the annotation helps to keep the track of such problems.

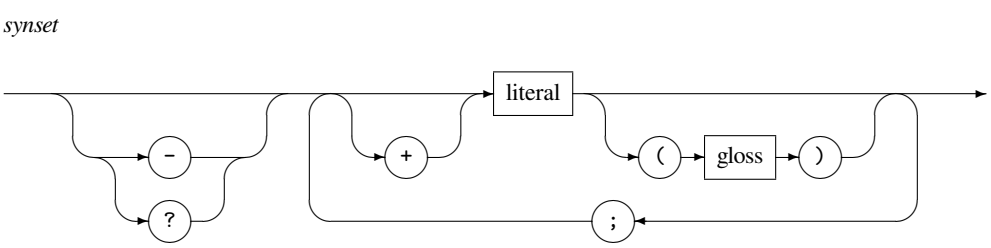

Fig. 2. Syntax diagram of synset definition. Definitions of *literal*, *gloss* and optional whitespace are not included for the sake of brevity

4 Nouns, Adjectives and Adverbs

The mapping for nouns, adjectives and adverbs is often straightforward. Sometimes one English synset is linked to two Slovak or Lithuanian synsets – this often appears when referring to gender distinct nouns that have only the gender neutral form in English (e. g. *10020890 doctor, doc, physician, MD, Dr., medico* is connected to two Lithuanian synsets, *2204 daktaras; gydytojas* (masc.) and *4914 daktarė; gydytoja* (fem.). Since the masculine/feminine gender distinctions in Lithuanian and Slovak are rather compatible, the mapping from Lithuanian synsets to Slovak ones is one-to-one.

Other kind of multiple correspondence is very rare, and although there is a number of homonyms and polysemous words in Lithuanian, we were able to disambiguate them quite clearly based on the English WordNet meanings – sometimes the distinction was even too detailed (e. g. Lithuanian word *veiksmas* appears in 6 different synsets, ranging from *14006945 action, activity, activeness (the state of being active)* to *07009640 act (a subdivision of a play or opera or ballet)*.

5 Verbs

5.1 Aspect

Both Lithuanian and Slovak verbs distinguish two aspects, perfective and imperfective, not as a morphological category, but as separate lexemes (though there are often derivation patterns converting between the two).

In Slovak, the base form is either perfective (e.g. *dať*) and the imperfective is derived semi-regularly with the *-va* morpheme (*dávať*), or the basic form is imperfective and there is a sizeable set of verbal prefixes turning it into perfective, often with considerable semantic differences (*robiť* $\rightarrow$ {*u,vy,za,pre,do,na,od*}*robiť*). There is also a class of ambivalent aspect verbs, but these can be though of as close homonyms. In the Slovak WordNet, we consistently keep both forms (if they exist) as separate synsets, linked to the same English synset.

In Lithuanian language, most of the verbs without any prefix are imperfective, whereas prefixed verbs denote a finished action. There are some exceptions, directional movement verbs are imperfective in the present tense – e.g. *atvykti* (inf., 'to come'), *atvyksta* (present tense, '[he] is coming'), however, in the past simple tense they denote a finished event

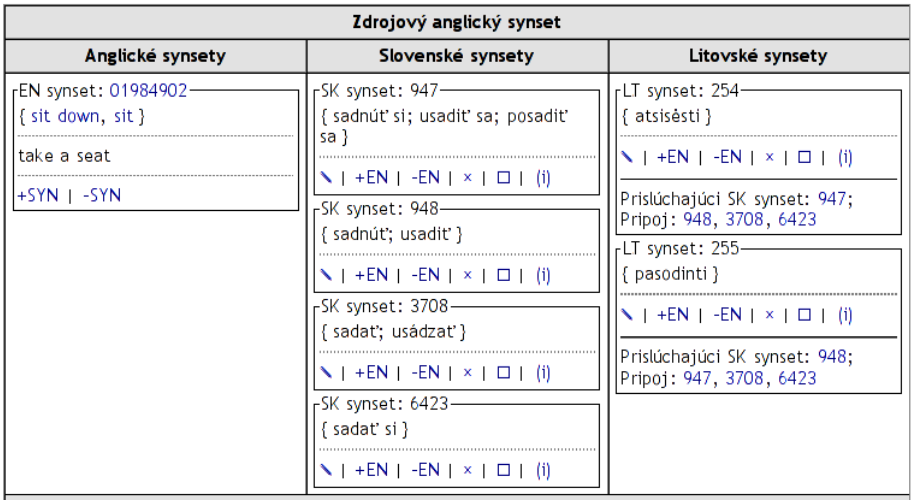

Fig. 3. Screenshot of the WordNet interface, with a view of corresponding English, Slovak and Lithuanian synsets

(*atvyko*, '[he] came'). Then there is a class of verbs neutral with respect to the aspect – e.g. *mirti 'to die'*.

In general, we preferred imperfective aspect to the perfective for simplicity, but we try to cover both aspects of the Slovak synset, if the same semantic meaning is preserved in Lithuanian.

Another exception are verbs indicating momentous (very short or abrupt) actions with suffixes *-el(ė)ti* and *-er(ė)ti*. In general, these forms were avoided in the WordNet, but they are included in cases where they tend to have a specific meaning – e.g. *gūžtelėti (pečiais)* 'to shrug (shoulders)'.

5.2 Reflexive verbs

Both Slovak and Lithuanian languages contain reflexive verbs, with approximately similar semantic behaviour. In Slovak, reflexivity is expressed by a separate reflexive pronoun/particle *sa* or *si*, which is nonetheless considered a part of the lexeme and we treat reflexive verbs as single units (literals including a space and the reflexive pronoun).

In Lithuanian, reflexive verbs have a reflexive affix *-si* or *-s*, which is attached to the end of the stem as an affix for prefixless verbs, but it is put as an infix after the prefix morpheme – e. g. *sukti* → *sukti**s***, but *nuprausti* → *nu**si**prausti*.

Syntactic reflexivity can express various semantic meanings, ranging from true reflexivity (action reflected towards oneself) through reciprocal, to pronominal reflexivity (where the reflexive status is obligatory but has no inherent meaning). There is often a ough conflation between reflexive and intransitive categories, and the non-reflexive and transitive ones.

In the Slovak WordNet we try to cover both reflexive and non-reflexive variants of the verb (if they both exist) in two separate synsets. In case where the reflexivity overlaps with transitivity, both synsets are mapped to the same English language synset (unless there are

separate transitive and intransitive English synsets). Lithuanian synsets are then mapped to the Slovak ones (not necessarily only related verbs, see Figure 1 for an example) if they cover the same meaning.

6 Manual Proofreading

The proofreading of both Slovak and Lithuanian parts was done almost simultaneously – the Slovak synsets have been proofread in two step process, first proofreading by one annotator and then a second proofreading by an independent one. Each step in itself consisted of two actions – verifying the completeness and correctness of literals in each synset, and verifying the synset position in the ontological hierarchy (i. e. its connection to the Princeton WordNet synset, its hypernyms and – if existing – hyponyms).

As the Slovak synsets acquired the "verified" status, corresponding Lithuanian ones have been proofread and edited as well, with paying attention to its interconnection to both the English and Slovak synsets.

The main Lithuanian language resources used for the proofreading were Modern Lithuanian Dictionary[4], Dictionary of International Words [1], Terminology Database of Lithuanian Republic[5], the website of the State Language Commission[6] and an encyclopaedic dictionary of computer science [2].

Only the terms approved by the Lithuanian Language Commission or present in one of the recommended (by the Commission) language resources were added to the Lithuanian WordNet. Therefore, colloquial expressions, neologisms and frowned-upon words were avoided at this phase of the proofreading (this however does not mean we are against their inclusion in the future).

7 Current Status

The Lithuanian WordNet started its life as a multilingual glossary, but it has grown up to be a small WordNet, with semantic hierarchy provided by Princeton WordNet. At the time of writing, the database composition is 7 874 noun synsets, 2 099 adjective synsets, 682 adverbial synsets and 533 verbal Lithuanian synsets. All of them are connected to the Slovak and English equivalents and the nouns, adjectives and adverbs are (once) manually proofread. Current work includes proofreading the verbs and extending existing word coverage. Once the database coverage and accuracy reaches satisfactory levels, its conversion into VisDic/DEBVisDic [5] could be considered, however the database still contains too many errors and omissions. Nevertheless, a preliminary version has been released[7] under GNU Affero General Public License, v. 3[8]; Creative Commons Attribution-ShareAlike 3.0 Unported License[9]; and Open Database License (ODbL) v1.0[10].

[4] Dabartinis lietuvių kalbos žodynas, `http://dz.lki.lt`
[5] Lietuvos Respublikos terminų bankas, `http://terminai.vlkk.lt/pls/tb/tb.search`
[6] Valstybinė lietuvių kalbos komisija, `http://vlkk.lt/`
[7] `http://korpus.sk/ltskwn_lt.html`
[8] `http://www.gnu.org/licenses/`
[9] `http://www.creativecommons.org/`
[10] `http://opendatacommons.org/licenses/`

Acknowledgments

The original multilingual dictionary was funded by the Slovak Online project[11]. Automatic synset generation and web based editing interface was provided by Faculty of Electrical Engineering and Informatics, Technical University of Košice. We thank Ján Genči and Ondrej Dzurjuv for their help.

References

[1] Bogušienė, V. and Bendorienė, A. (2008). *Tarptautnių žodžių žodynas.* Alma littera, Vilnius, Lithuania.

[2] Dagienė, V., Grigas, G., and Jevsikova, T. Anglų-lietuvių kalbų kompiuterijos žodynėlis. Retrieved from `http://www.likit.lt/en-lt/angl.html` on 17th October 2013.

[3] Dzurjuv, O., Genči, J., and Garabík, R. (2011). Generating sets of synonyms between languages. In *Natural Language Processing, Multilinguality. Proceedings of the 6th International Conference SLOVKO 2011*, Modra, Slovakia.

[4] Hana, J. (1998). Two Level Morphology of Esperanto. Master's thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

[5] Horák, A., Pala, K., Rambousek, A., and Povolný, M. (2006). DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. In *Proceedings of the Third International WordNet Conference – GWC 2006*, pages 325–328, Brno, Czech Republic.

[6] Lindén, K. and Carlson, L. (2010). FinnWordNet – WordNet på finska via översättning. *LexicoNordica – Nordic Journal of Lexicography*, (17):119–140.

[7] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

[8] Pease, A., Fellbaum, C., and Vossen, P. (2008). Building the Global WordNet Grid. In *Proceedings of the 18th International Congress of Linguists (CIL18)*, Seoul, Republic of Korea.

[9] Rudnicka, E., Maziarz, M., Piasecki, M., and Szpakowicz, S. (2012). A Strategy of Mapping Polish WordNet onto Princeton WordNet. In Kay, M. and Boitet, C., editors, *COLING (Posters)*, pages 1039–1048. Indian Institute of Technology Bombay.

[10] Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufiş, D., Koeva, S., Totkov, G., Dutoit, D., and Grigoriadou, M. (2002). BALKANET: A Multilingual Semantic Network for the Balkan Languages. In *Proceedings of the 1st Global WordNet Association conference*.

[11] Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, pages 5–7.

[11] `http://slovake.eu/`