Constructing a poor man's wordnet in a resource-rich world

Darja Fišer, Benoît Sagot

Lang Resources & Evaluation
DOI 10.1007/s10579-015-9295-6

ORIGINAL PAPER

Constructing a poor man's wordnet in a resource-rich world

Darja Fišer · Benoît Sagot

Abstract In this paper we present a language-independent, fully modular and automatic approach to bootstrap a wordnet for a new language by recycling different types of already existing language resources, such as machine-readable dictionaries, parallel corpora, and Wikipedia. The approach, which we apply here to Slovene, takes into account monosemous and polysemous words, general and specialised vocabulary as well as simple and multi-word lexemes. The extracted words are then assigned one or several synset ids, based on a classifier that relies on several features including distributional similarity. Finally, we identify and remove highly dubious (*literal, synset*) pairs, based on simple distributional information extracted from a large corpus in an unsupervised way. Automatic, manual and task-based evaluations show that the resulting resource, the latest version of the Slovene wordnet, is already a valuable source of lexico-semantic information.

Keywords Wordnet development · Multilingual lexicon extraction · Word-sense disambiguation · Distributional similarity

1 Introduction

Motivation. In the past decade, the role of lexical knowledge has increased in many areas of natural language processing as it has been shown that exploiting lexical

D. Fišer
Department of Translation Faculty of Arts, University of Ljubljana Aškerčeva 2, 1000 Ljubljana, Slovenia
e-mail: darja.fiser@ff.uni-lj.si

B. Sagot (✉)
Alpage, INRIA Paris-Rocquencourt & Université Paris-Diderot, Bâtiment Olympe de Gouges, Rue Albert Einstein, 75013 Paris, France
e-mail: benoit.sagot@inria.fr

Published online: 11 February 2015

resources benefits the performance of various tasks. For example, Gabrilovich and Markovitch (2006) have demonstrated that using encyclopaedic knowledge improves automatic document classification. Similarly, Nastase (2008) have employed it to help text summarisation. An improvement in a question-answering system that takes advantage of dependencies between words in a lexico-semantic network has been achieved by Harabagiu et al. (2000), the advantages of which have also been shown in word-sense disambiguation (Cuadros and Rigau 2006) and machine translation tasks (Carpuat and Wu 2007).

Several frameworks for organising and representing lexical knowledge have been proposed, such as ACQUILEX[1] (Copestake et al. 1993), Roget's Thesaurus[2] (Kirkpatrick 1987), MindNet[3] (Richardson et al. 1998), ConceptNet[4] (Liu 2003) or Cyc[5] (Matuszek et al. 2006). But one of the best known and most widely used lexico-semantic resources for natural language understanding and interpretation as well as in Semantic Web applications are Princeton WordNet (Fellbaum 1998) and its sister wordnets for languages other than English, developed within projects such as EuroWordnet (Vossen 1999), BalkaNet (Tufiş 2000), AsianWordNet (Sornlertlamvanich 2010) and the most recent project, called the Open Multilingual Wordnet[6] that has normalised and merged all the wordnets that allow redistribution, and currently contain wordnets for 27 languages (Bond and Foster 2013).

While it is true that wordnets were originally constructed manually in order to maximise linguistic soundness and accuracy of the developed database, such an endeavour is too time-consuming and expensive to be feasible for many languages still lacking a wordnet. This is why semi- or fully automatic approaches have become popular, which exploit various types of existing resources to facilitate the development of a new wordnet. However, a common problem with automatically induced wordnets is the necessary trade-off between limited coverage and the desired level of accuracy, both of which are required if the developed resource is to be useful in a practical application. An important advantage of our approach compared to related work is that our approach is much more lightweight and straightforward: it does not require complex processing available only for some languages (e.g., computing parses for syntactical distributional similarity), language-dependent rules (e.g., patterns for extracting hypernymy relations) or costly lexical resources (e.g., large-scale thesauri).

Contribution. In this paper we present a two-step automated approach for building wordnets. The approach is language-independent as shown with its successful application to the development of the French wordnet WOLF (Sagot and Fišer 2008). It is suited for research scenarios that do not allow for long-running resource development by a team of professional lexicographers. Instead, it recycles different types of resources that are already available for the target language, such as

[1] http://www.cl.cam.ac.uk/research/nl/acquilex/ [06.07.2014].

[2] http://www.bartleby.com/62/ [06.07.2014].

[3] http://research.microsoft.com/nlp/Projects/MindNet.aspx [06.07.2014].

[4] http://web.media.mit.edu/hugo/conceptnet/ [06.07.2014].

[5] http://www.cyc.com/ [06.07.2014].

[6] http://compling.hss.ntu.edu.sg/omw/ [06.07.2014].

machine-readable dictionaries, parallel corpora and Wikipedia in a way that maximises the amount of extracted lexical information from each type of the resource used. The created wordnets are directly aligned to Princeton WordNet as well as among each other. They can therefore be used in multilingual tasks, such as machine translation, cross-lingual information retrieval and question answering, the demand for which is growing with the increased multilinguality of the web (Nie 2010). As opposed to most related work relying solely on Wikipedia, our approach is not limited only to nominal concepts but can handle all parts-of-speech. Our approach is also comprehensive in that we handle monosemous and polysemous words, general and specialised vocabulary as well as simple and multi-word lexemes.

Preliminary and partial versions of the full methodology described here were already used for building the first freely available wordnet for Slovene (Fišer and Sagot 2008). However, since the first version, our techniques have been substantially refined and extended. In this paper, we recall how the first versions of sloWNet were built and, more importantly, how the refinements allowed us to more than double its size while increasing its overall accuracy. The result is the first freely available semantic lexicon for Slovene wordnet called sloWNet.

Overview. This paper is structured as follows: in Sect. 2 we give an overview of related work. In Sect. 3 we describe the process of extracting lexico-semantic information from structured, semi-structured and unstructured resources. In Sect. 4 we explain the two-step process used for merging this information and constructing sloWNet: we first built restricted resources containing only the literals with the highest certainty level and then used these in a large-scale enrichment step as training data for a maximum entropy classifier used for computing whether a less certain literal extracted from the existing resources is an appropriate candidate for a given synset. Section 5 is dedicated to the evaluation of the created resource, and Sect. 6 discusses the results and points towards future directions of research.

2 Related work

Automatic techniques for wordnet development can be divided in two approaches: the merge approach and the extend approach (Vossen 1999). In the merge approach, an independent wordnet for a certain language is first created based on monolingual resources and then mapped to other wordnets (Rudnicka et al. 2012). In the extend approach, which we used in this work, a fixed set of synsets from Princeton WordNet (PWN) are translated into the target language, preserving the structure of the original wordnet. The extend approach relies on the assumption that concepts and semantic relations between them are language independent, at least to a large extent. Apart from faster and cheaper construction of the lexical resource, the biggest advantage of this approach is that the resulting wordnet is automatically aligned to all other wordnets built on the same principle and therefore available for use in multi-lingual applications, such as machine translation and cross-language information retrieval.

The downside of the expand approach is that the target wordnets inherit any inconsistencies in PWN but are also biased by PWN and may, in an extreme case, become arbitrary (Orav and Vider 2004), (Wong 2004). For example, PWN contains a synset {*performer*, *performing artist*} which is defined as *an entertainer who performs a dramatic or musical work for an audience*. However, there is no word or phrase in Slovene that denotes the concept describing actors and singers collectively. Such cases have been dealt with in several wordnet development projects by providing the closest possible match for the synset and aligning the two wordnets with a *near_synonym* relation. In this way, the overall structure of straightforward cases remained intact and the exceptions encoded. Despite these difficulties, the approach is still attractive due to its much greater simplicity, which outweighs the language difference issues. What is more, the fact that the resulting resource is aligned with the PWN, and therefore also with all other resources and tools that follow the PWN principles, is highly valuable. This is why the expand model has been adopted in a number of recent projects, such as the BalkaNet (Tufiş 2000), MultiWordNet (Pianta et al. 2004) and BableNet (Navigli and Ponzetto 2010). With this we do not suggest that the expand approach is universally superior to the merge approach but that in our research settings, and we believe this applies to many other researchers as well, it was the optimal one. As a result, issues regarding the inventory or the relative position of synsets in the resulting semantic network do exist, but analysing them and adapting wordnet structure accordingly lies beyond the scope of this study.

Under the setting of the expand approach, approaches vary according to the type of resources that are available for the construction of a wordnet in a particular language. Early approaches link English entries from machine-readable bilingual dictionaries to PWN synsets under the assumption that their counterparts in the target language correspond to the same synset (Knight and Luk 1994; Yokoi 1995). A well-known problem with this approach is that bilingual dictionaries are generally not concept-based but follow traditional lexicographic principles, which is why the biggest obstacle is the disambiguation of dictionary entries. Also, bilingual machine-readable dictionaries often have limited coverage or are not digitally available for the relevant language pair at all.

This problem is overcome by a different set of approaches in which bi- or multilingual lexicons are extracted from parallel corpora (Resnik and Yarowsky 1997; Fung 1995). The main underlying assumption in these approaches is that senses of ambiguous words in one language are often translated into distinct words in another language. Furthermore, if two or more words are translated into the same word in another language, then they often share some element of meaning. This results in sense distinctions of a polysemous source word or yields synonym sets. As a result, parallel corpora have been utilised to induce synsets for a new language (Dyvik 2002; Ide et al. 2002; Diab 2004).

The third set of approaches that have become popular in the past few years draw upon Wikipedia, a large-scale collaborative encyclopaedic resource available for many languages. New wordnets have been induced by associating Wikipedia pages with the most frequent WordNet sense (Suchanek et al. 2008) by using structural information to assign Wikipedia categories to WordNet (Ponzetto and Navigli 2009)

or by extracting keywords from Wikipedia articles (Reiter et al. 2008). Vector-space models have been developed to map Wikipedia pages to WordNet (Ruiz-Casado et al. 2005; Declerck et al. 2006). The most advanced approaches use Wikipedia and related projects, such as Wiktionary, to bootstrap wordnets for multiple languages (de Melo and Weikum 2009; Navigli and Ponzetto 2010, 2012).

Our attempt in the construction of sloWNet has been designed to benefit from the combination of the available resources, which were of three different types: general and domain-specific bilingual dictionaries, parallel corpora and Wiki resources (Wikipedia and Wiktionaries). A basic version of the approach has already proved to be successful in our previous work where we created an initial Slovene wordnet drawing from these resources (Erjavec and Fišer 2006; Fišer and Sagot 2008). The focus in this paper, however, is to take the approach a step further and not limit the extraction of translations to monosemous words only but to extend it to polysemous words as well in order to take full advantage of each resource and at the same time mitigate the limitations each one of them brings by weighting candidates for synsets according to a selection of features.

3 Automatic extraction of lexico-semantic information

In this section we describe the extraction of translation pairs from three types of resources: structured (general and domain-specific dictionaries and lexica), semi-structured (Wikipedia articles) and unstructured (parallel corpora). In the extraction process our task is to extract as many translation variants for each word or multi-word expression as possible in order to capture as many of its senses as possible. The goal of the procedure is to obtain wordnet candidates from the extracted translation pairs in the form of pairs consisting of a literal and its synset, i.e. translation of a source word with an assigned synset id from wordnet.

The acquisition of Slovene wordnet candidates is based on PWN concepts (synsets) and proceeds as follows: we take a PWN literal which appears in several synsets (say, n_s synsets) and has many possible translations (say, n_t translations) according to the information extracted from the various resources. Our aim is therefore to select the best candidates among the $n_s.n_t$ possible ones by disambiguating each of these translations either in context thanks to parallel corpora, or out-of-context for all translation pairs extracted from dictionaries and Wikipedia. If the PWN literal is monosemous,[7] we simply assign the same concept (its synset id) to its translation. If the PWN literal is polysemous, we choose the best synset id for its translation by defining a synset proximity metric that is based on the initial restricted version of the Slovene wordnet.

[7] In this paper we use the term monosemous for such literals that only appear in one synset in the Princeton WordNet. While this is unproblematic in most cases, there is a possibility that some words only appear to be monosemous according to the lexical resource which is missing some senses because the resource is incomplete.

3.1 Extracting lexico-semantic information from structured resources

Bilingual dictionaries are very rich sources of lexico-semantic information that have already been compiled, analysed and structured, which is why they are an obvious first choice for harvesting the lexico-semantic information needed to populate a wordnet for a new language. Most bilingual dictionaries contain very little noise, have good coverage of general vocabulary across all parts of speech, contain translations for several senses of polysemous words and sometimes even definitions. However, bilingual dictionaries provide non-contextualised information and even when sense distinctions are explicitly encoded in the dictionary structure they are usually coarser-grained than PWN, which is why dictionary entries cannot be mapped directly to PWN concepts. Our approach for assigning synset ids to the translation pairs we extracted from dictionaries is described in Sect. 4. Here we present the dictionaries we used, the extraction process and the results we obtained.

Wiktionaries[8] are freely available collaboratively constructed bilingual dictionaries which were originally designed as lexical companions to Wikipedia. They contain definitions of words as well as some additional information, including their translations into other languages which are sometimes structured into senses. We used English and Slovene Wiktionaries and extracted translation pairs for all parts-of-speech from these two resources on the basis of translation sections within the articles. However, Wiktionaries do not (yet) have good coverage of Slovene, which is why the number of lexicon entries we were able to extract is relatively low: 7,029 translation pairs from the Slovene Wiktionary and 6,052 from the English Wiktionary. For each entry, we also tried to extract a gloss based either on the first sentence of the Wiktionary article, or, if available, from the short glosses associated with each sense.

In order to extract the general vocabulary that was largely missing in Wiktionary, we used a digitised **traditional English–Slovene** (Grad et al. 1999) **and a Slovene–English dictionary** (Grad and Leeming 1999). The dictionaries do not contain definitions but we were able to harvest 207,972 translation pairs from the English–Slovene dictionary and 72,954 from the Slovene–English one.

For domain-specific vocabulary we used **Wikispecies**,[9] which is a taxonomy of living species that includes Latin standard names as well as vernacular terms for the common species. This allowed us to extract 2,360 English–Slovene pairs. In a similar way, we obtained 31,702 translation English–Slovene pairs from the domain-specific thesaurus **Eurovoc**,[10] an on-line dictionary of informatics **islovar**[11] and a **military glossary** (Korošec et al. 2002).

The result of our extraction process is a large bilingual lexicon containing 282,789 unique English–Slovene translation pairs with the name of the resource(s) they originate from. The figures for this extracted bilingual lexicon are summarised in the upper part of Table 1.

[8] http://www.wiktionary.org/ [06.07.2014]

[9] http://species.wikimedia.org/ [06.07.2014].

[10] http://eurovoc.europa.eu/ [06.07.2014].

[11] http://www.islovar.org/ [06.07.2014].

Table 1 Quantitative information about the bilingual English–Slovene lexicons extracted from various available structured and semi-structured sources

Input resource	En–Sl unique pairs
English wiktionary	6,052
Slovene wiktionary	7,029
Wikispecies	2,360
Slovene–English dictionary	72,954
English–Slovene dictionary	207,972
Eurovoc and specialised vocabularies	31,702
Wikipedia	32,161

3.2 Extracting lexico-semantic information from semi-structured resources

Less structured than dictionaries but still with a much more predefined structure than free text is the on-line multilingual collaborative encyclopaedia **Wikipedia**.[12] We used English and Slovene Wikipedia for extracting a bilingual lexicon by following inter-language links that relate two articles on the same topic in the two corresponding Wikipedias. We enhanced the extraction process with a simple pattern-based analysis of article bodies based on regular expressions. More precisely, we first looked across each article body for all non-sentence-initial occurrences of its title, ignoring capitalisation, and compared the relative frequency of all capitalisation variants. This allowed us to resolve ambiguities arising from the capitalisation of article titles (e.g., *Grass_author*, *Grass_plant*). Our pattern-based analysis allowed us also to extract relevant information from the structure of the first sentence of each article, thus identifying synonyms for the key terms (e.g., *Cannabis*, *also known as marijuana*), their definitions (e.g., *Hockey is a family of sports in which two teams play against each other by trying to manoeuvre a ball or a puck into the opponent's goal using a hockey stick.*) and usage examples (e.g., *The true grasses include cereals, bamboo and the grasses of lawns (turf) and grassland.*). It would also be possible to further analyse article bodies in order to extract other semantically related terms (e.g., *hockey: sports [hypernym], hockey: ball, puck, goal, hockey stick [meronym]*) which we plan to do in our future work.

As a result, we obtained 32,669 Slovene–English entries that are predominantly nouns or nominal multi-word expressions (common or proper), yielding 32,161 unique translation pairs (see Table 1).

3.3 Extracting lexico-semantic information from unstructured resources

The final type of the resources used in our experiment are the unstructured resources, that is free text. We used the SEE–ERA.NET corpus (Tufiş et al. 2009), a 1.5-million-word subcorpus of JRC-Acquis (Steinberger et al. 2006) in eight languages. Apart from Slovene, we used English, Romanian, Czech and Bulgarian.

[12] http://www.wikipedia.org/ [06.07.2014].

Table 2 Quantitative information about the various multilingual lexicons extracted from the SEE-ERA.NET multilingual corpus

Lexicon languages	Unique entries
Slovene–Czech–Bulgarian	59,369
Slovene–Czech–English	52,192
Slovene–Czech–Romanian	57,674
Slovene–Czech–Bulgarian–English	55,768
Slovene–Czech–Bulgarian–Romanian	55,275
Slovene–Czech–Romanian–English	51,013
Slovene–Czech–Bulgarian–English–Romanian	49,892

We used different tools to PoS-tag and lemmatise the corpus before word-aligning it with Uplug (Tiedemann 2003). Because word-alignment was performed on single words only, we were not able to generate any translation equivalents for multi-word expressions. The output of the word alignment process is a file with word links between word occurrences and information on word link certainty.

This allowed us to build bilingual lexicons that include all translation variants of words as well as frequency, part-of-speech and token id information for each entry. Note that we base this decision on the fact that the quality of word-alignment is far from perfect in general but is improved when the languages in question are closely related, as is the case for example for Slovene and Czech. The bilingual lexicons we extracted range from 43,024 entries for the cs–en lexicon to 50,289 for the cs–bg one. These bilingual lexicons are then combined into five multilingual lexicons which all involve at least Slovene and two other languages. They contain between 52,193 (Slovene–Czech–English) and 55,768 (Slovene–Czech–Bulgarian–English) entries (see details in Table 2). Obviously, not all these candidates are correct; errors may appear for several reasons, such as tagging, lemmatisation, or alignment problems. However, many of these errors are eliminated in the next stage of the process.

3.4 Assigning synset ids to translation pairs

As explained above, creating or expanding a wordnet by preserving PWN structure and synset inventory can be viewed as generating (*literal, synset*) pairs. This is achieved by assigning synset ids to the extracted translation pairs. The resources we used can be divided in two groups: lexicon-based and alignment-based resources, the difference between them being that lexicon-based entries are not associated with a particular occurrence in a particular context while the alignment-based entries are. This is why the process of assigning synset ids differs for these two groups. The lexicon-based bilingual entries we extracted are extremely valuable because they are far more numerous and accurate than word-alignment based information.

Let us consider for example the following English–Slovene entry we extracted from Wiktionary: $(organ_{\text{en}}, organ_{\text{sl}})$. It does not contain any information that would

make it possible for us to determine which of the 6 PWN synsets containing $organ_{en}$ as a literal would be appropriate to be translated with $organ_{sl}$ in sloWNet. In this case, only the 'body part' sense of English $organ_{en}$ can be translated as $organ_{sl}$ in Slovene, but not the 'musical instrument' sense (which translates as $orgle_{en}$). In Wiktionary articles, translations of a given word are sometimes organised by senses that are associated with short glosses. These have been compared to PWN glosses in order to map Wiktionary senses to PWN synsets (Bernhard and Gurevych 2009; Casses 2010). The first sentence of a Wikipedia article can be used in a similar way (Ruiz-Casado et al. 2005). However, this is not the case for all Wiktionary entries or other resources in this category. We therefore decided to assign a synset id only to those translation pairs that contain a monosemous English word and postpone the disambiguation of polysemous lexicon-based entries to a later stage (see Sect. 4).

On the other hand, alignment-based entries contain contextual information that enables semantic disambiguation, as they are composed of translation equivalents which have been word-aligned at least once in the multilingual corpus. For each entry, we gathered the set of all possible synset ids associated with each word in each language involved (apart from Slovene) using the corresponding BalkaNet wordnets (Tufiş 2000). Since all BalkaNet wordnets use the same synset inventory and synset ids as PWN, we were then able to compute the intersection of ids for all languages. The result contains all synset that are shared among all non-Slovene literals in this particular multilingual lexical entry. We then assigned these synset ids to their Slovene equivalent.

To illustrate this process, Table 3 shows how a few entries from the en–cs–ro–bg–sl lexicon are disambiguated and associated with a synset id, thus generating (literal, synset) candidates. The first two 5-lingual entries provide different translation variants for the English noun 'party,' which are *strana* (cs), *partid* (ro), партия (bg) and *stranka* (sl) for 'political party,' and *večirek* (cs), *petrecere* (Ro), забава (bg) and *zabava* (sl) for 'social occasion'. A comparison of all the synsets these words appear in in the respective wordnets shows that the translation variants from the two lexicon entries do not have any synset ids in common, which suggests that they are translations of different senses of the polysemous English noun 'party'. A different intersecting synset id in each lexicon entry is therefore assigned to their Slovene translations, which results in generating two different Slovene candidates, namely (*stranka*, 07758173) for the 'political party' sense and (*zabava*, 07753857) for the 'social occasion' sense. On the other hand, the last two entries in the lexicon are for the English word 'army' and are the same in all languages except in Slovene. Since translation variants in both lexical entries share the same intersecting synset id, it is assigned to both Slovene variants. This generates two synonym candidates (i.e. the synset id in both candidates is the same), namely (*armada*, 07686671) and (*vojska*, 07686671). Using multiple language in this way on polysemous lexical entries eliminates most alignment errors. Indeed, it is rather unlikely that the same polysemy occurs in many different languages or that alignment errors lead to a non-empty intersection. Therefore, the intersection of all possible senses in each language is likely to output only the correct synset id(s). Obviously, this is even more so when using more different languages than when using only one language apart

Table 3 Example of lexical disambiguation based on multilingual word-alignment from a parallel corpus

English party	Czech strana	Romanian partid	Bulgarian партия партия	Slovene stranka
06992505	08042997	**07758173**	06600579	*→07758173*
07753857	04052451		**07758173**	
07758173	08120765			
07765339	07610762			
09728152	06366295			
	08120943			
	07897707			
	05875089			
	09604212			
	07758173			
	13109893			
party	večírek	petrecere	забава забава	zabava
06992505	07756360	06992505	01190333	*→07753857*
07753857	**07753857**	**07753857**		
07758173	07753722			
07765339				
09728152				
army	armáda	armată	армия армия	armada
07686671	00555727	**07686671**	**07686671**	*→07686671*
07694312	**07686671**	07701234	07701861	
	07701861	07701861		
		07694312		
army	armáda	armată	армия армия	vojska
07686671	00555727	**07686671**	**07686671**	*→07686671*
07694312	**07686671**	07701234	07701861	
	07701861	07701861		
		07694312		

Note that synset ids are PWN 2.0 synset ids. This is because this part of the work was carried out on BalkaNet wordnets, which are aligned to PWN 2.0

from English and Slovene, which is the minimum required for the intersection to actually be possible. On the other hand, the more languages are used, the more reliable but the less numerous the generated candidates will be because intersecting more translation sets in more languages can only lead to a smaller intersection.

Applied to the above-mentioned multilingual lexicons, this technique yielded five different sets of candidates filling different synsets with at least one Slovene literal. They include between 1,364 (sl-ro-cs-bg-en) and 4,232 (sl-cs-en) entries. Because the pre-processing stages, such as tagging, lemmatisation and word-alignment were not perfect, it is expected that the synsets created in this way will inherit some of the errors which will hopefully be filtered out by the classifier (see Sect. 4.3).

4 Automatic induction of synsets for wordnet extension

The development of the initial sloWNet was achieved in a three-step process. First, we created baseline versions of wordnets (Fišer and Sagot 2008) by using only (*literal, synset*) pairs obtained from the parallel corpus which could be disambiguated based on other languages, and monosemous words extracted from the dictionaries, lexica and Wikipedia which required no disambiguation. Such a restricted wordnet was relatively reliable but did not use full potential of the available resources. This is why, after making a few improvements on this initial sloWNet, we performed a large-scale extension process, aiming at taking full advantage of the resources and improving the coverage of the resource without lowering its accuracy (Sagot and Fišer 2011). In this section we briefly describe the two first steps and then give a detailed account of the novel extension step.

4.1 Developing baseline wordnets

The first step in the development of sloWNet was achieved in 2008, when the first version was created (Fišer and Sagot 2008).[13] For this first step only monosemous PWN literals were translated using bilingual resources, thus avoiding disambiguation issues. However, all PWN literals were used for adding target language literals in the synsets found by the alignment-based approach. If the same (*literal, synset*) pair was created from more than one resource (e.g., from a multilingual lexicon that was extracted from the word-aligned corpus and from a bilingual lexicon that was extracted from Wikipedia), the information on the source of the generated synset was retained. This enabled us to perform a simple heuristic filtering according to the reliability of each resource, on the number of different resources that assign a given literal to the same synset and on frequency information (for resources from the alignment approach).

Automatic insertion of Slovene literals to synsets inevitably leads to gaps in the hierarchy. Because we are aware of the importance of the conceptual density and hierarchy preservation principles for applications (Tufiş 2000), we inherited the structure and relations of the missing synsets from PWN, thus leaving many empty synsets. Therefore, in case an application runs into an empty synset, it can still use the relation information to access a more general concept. Other language-independent information (e.g., PoS, domain, semantic relations) was inherited from the PWN as

[13] See however (Erjavec and Fišer 2006) for preliminary experiments on building a Slovene wordnet from the Serbian wordnet (Krstev et al. 2004).

Table 4 Quantitative data about the number of non-empty synsets within the different sloWNet versions, and a comparison with PWN 2.0

	PWN 2.0	sloWNet 2.0	sloWNet 2.2	sloWNet 3.0
All	115,424	29,108	17,817	42,919
BCS1	1,218	714	1,203	1,208
BCS2	3,471	1,361	2,192	3,111
BCS3	3,827	1,611	1,232	2,698
Non-BCS	106,908	25,422	13,190	35,902
N	79,689	22,927	16,234	30,911
V	13,508	1,547	1,097	5,337
Adj	18,563	4,376	429	6,218
Adv	3,664	258	57	453

Figures are broken down by BCS category (see text) and by PoS. Note that sloWNet 2.2 and 3.0 use the synset inventory of PWN 3.0. Therefore, BCS information is approximate, as it was computed automatically

well. We also adopted the three Base Concept Sets (BCS) which were introduced in the BalkaNet project (Tufiş and Cristea 2002) and comprise 8,516 synsets that have been commonly agreed to be implemented by all consortium members in order to obtain a guaranteed overlap of lexicalised concepts between the BalkaNet languages, such as *building*, *vehicle*, *animal*, etc. The Base Concepts are the concepts that play the most important role in the various wordnets of different languages. This role is measured in terms of two main criteria: (1) a high position in the semantic hierarchy and (2) having many relations to other concepts (Weisscher 2013). The Base Concepts play a crucial role in wordnet building that is typically top-down: first, a core wordnet is developed around the Base Concepts that contains about 5,000–10,000 synsets and is highly compatible in coverage and semantic interpretation with wordnets in other languages, and then the core wordnet is extended beyond 20,000 synsets in a top-down fashion, given the semantic basis of the core wordnet.

The figures for the first version of sloWNet (version 2.0) are given in the second column in Table 4, together with corresponding figures about PWN 2.0, and more recent versions of sloWNet (version 2.2 and version 3.0 resulting from the work described in this paper).

4.2 Enhancing baseline sloWNet

After the restricted and automatically produced version of sloWNet (version 2.0) was built, it underwent some improvement steps. First of all, because legal aspects regarding the use of the traditional English–Slovene and Slovene–English dictionaries were unclear, we re-ran all experiments without using this resource. Second, due to poor parsing of Wikipedia articles, many synsets contained duplicate literals that were identical except in stress markings (e.g., *kolo* and *koló*) which were therefore normalised and merged. Since sloWNet was used for manual semantic annotation of a corpus (Fišer and Erjavec 2009), it was also extensively manually edited in order to delete erroneous senses of words that were annotated and add the

missing ones. This is why the published Slovene baseline wordnet (version 2.2), used during the large-scale extension experiments described in the next section, is significantly different and smaller than the automatically produced version 2.0. Finally, sloWNet 2.2 uses the synset inventory from PWN 3.0, whereas sloWNet 2.0 uses the synset inventory from PWN 2.0.[14] Note however that, in the end, we did use the traditional dictionaries during that large-scale extension step.

4.3 Large-scale wordnet extension

Restricting the use of a bilingual lexicon to monosemous English literals is a safe but limited approach that does not exploit the available resources to their full potential. However, using lexicon-based candidates generated from polysemous English literals is only possible if we can establish the likelihood with which a word should be added to a particular synset, i.e. can compute the semantic distance between a given Slovene literal and synset id. We designed such a technique based on the already-existing Slovene wordnet (version 2.2) and we present it in this section.

4.3.1 Using a probabilistic classifier

Our technique relies on a probabilistic classifier that uses various features associated with each (*literal, synset*) candidate. The underlying idea is as follows: we have a baseline wordnet at our disposal, and a large set of lexicon-based candidates to evaluate. We extract all (*literal, synset*) pairs that are already in the baseline wordnet and consider these candidates as valid ones while all the other candidates are considered invalid, thus creating a "copper standard", i.e. a reasonable although noisy training set for a probabilistic model. The training set is noisy for two reasons: first, the baseline wordnet itself contains noise because not all synsets were manually validated; second, and more importantly, many of our new candidates are valid even though they are not in the baseline wordnet. In fact, such candidates are exactly those that we are looking for to extend our wordnet. In order to use the copper standard as the training set for a classifier, we need to extract suitable features for the candidates which was performed with the Maximum-Entropy package megam (Daumé 2004) based on the features described below. The result of our classifiers on training data is a certainty value between 0 and 1. We empirically set the threshold at 0.1 (see Sect. 6.1 for motivations for this value) and added all the candidates that pass the threshold to the wordnet.

4.3.2 Feature selection

This section contains a description of the features we used to train our candidate evaluation models. The most important feature models the semantic proximity between a literal and a synset. Let us illustrate it on the previous example

[14] The conversion from one synset inventory to another was achieved based on an automatic PWN 2.0 to 3.0 mapping (Erjavec, p.c.).

($organ_{en}$, $organ_{sl}$). In PWN (3.0), 6 synsets contain the literal $organ_{en}$, which is why we also generate 6 different (*literal, synset*) candidates from the bilingual entry ($organ_{en}$, $organ_{sl}$). We now need to know which of these 6 candidates are valid, i.e. to which of the 6 corresponding synsets the Slovene literal $organ_{sl}$ should be added in sloWNet. We therefore compute the semantic similarity of the Slovene literal $organ_{sl}$ w.r.t. each of these 6 synsets.

We first represent each sloWNet synset by a bag of words obtained by extracting all literals from this synset and all the synsets up to two nodes apart in sloWNet, i.e. related via a path of length at most two involving any type(s) of lexical relation(s).[15] For example, the synset {$organ_{en}$, *pipe organ*} in PWN is represented by the bag of words {*glasbilo*, *Anton Bruckner*, *glasbenik*, *Johann Sebastian Bach*, *pisalni*, *klavirska*, *harmonika*,...} ('musical instrument,' 'Anton Bruckner,' 'musician,' 'Johann Sebastian Bach,' 'writing$_{adj}$,' 'piano$_{adj}$,' 'accordion,' 'device,'...).

Next, we use a distributional semantic model for evaluating the semantic similarity of *orgle* w.r.t. this bag of words. We use the freely-available SemanticVectors package (Widdows and Ferraro 2008),[16] which relies on the Lucene indexing system.[17] This package is able to build a word-document frequency matrix from a large set of documents, reduce the dimensionality of this matrix by a random projection technique, and finally extract one semantic vector per word from this reduced matrix. The package then allows for assessing the distributional semantic similarity of two bags of words using Latent Semantic Analysis. The documents we used for building such distributional semantic models are 334,000 lemmatised paragraphs from the FidaPLUS corpus (Arhar and Gorjanc 2007) (180,000 distinct lemmas).[18] Applied to our example, the semantic similarity between $organ_{sl}$ and the synset {$organ_{en}$, *pipe organ*} is only 0.021, while the similarity between $organ_{sl}$ and one of its valid synsets, {$organ_{en}$}, defined as *a fully differentiated structural and functional unit in an animal that is specialised for some particular function*, is 0.668. Indeed, in Slovene, $organ_{sl}$ has the 'body part' meaning but not the 'musical instrument' (*orgle*, in Slovene).

Apart from that semantic similarity measure, the other features we used are the following. Let us consider a (literal, synset) candidate (l_t, s) (i.e. l_t is a literal in the target language, here Slovene) that has been generated because our bilingual resources provided entries of the form $(l_{e,1}, l_t)\ldots(l_{e,n}, l_t)$, where all PWN literals $l_{e,i}$'s are among s's literals. The number of such PWN literals is one of the features. Each possible source (e.g., the English wiktionary) corresponds to one feature, which receives value 1 if and only if at least one of the $(l_{e,i}, l_t)$ bilingual lexical entries was extracted from this source. Moreover, we extract the lowest polysemy index among all the occurences of $l_{e,i}$. For example, if the least polysemous $l_{e,i}$ is in

[15] This threshold of 2 was empirically found to be the best balance between the number of related words (a threshold of 1 or 0 would have provided us too few, a threshold of 3 or more too many) and the relevance of the related words (a threshold of 3 or more gathers many literals which are not relevant as descriptors of the input synset).

[16] http://semanticvectors.googlecode.com [06.07.2014].

[17] http://lucene.apache.org [06.07.2014].

[18] The Slovene lemmatisation was performed using the ToTaLe system (Erjavec et al. 2005).

two PWN synsets, this feature receives value 2. The idea is that if the candidate is generated from at least one monosemous PWN literal, then it is very likely to be correct, whereas if it is generated from only highly polysemous PWN literals, it is much more questionable. Finally, the number of tokens in l_t is used as a feature as well (literals with many tokens are usually not translations of PWN literals but rather glosses that arise from Wikipedia or Wiktionary, and are therefore incorrect).

4.3.3 Building classification models

The resulting models are shown in Table 5. They clearly show that semantic similarity is relevant and useful as it is the feature with the highest weight. As expected, the lowest polysemy index among English literals also contributes positively, as does the number of different English literals yielding the generation of the candidate, and the number of sources involved. On the other hand, as predicted as well, the number of tokens in the target language literal negatively contributes to the certainty score. Finally, the different sources are also associated with a weight.

4.3.4 Results of the classification

After having trained these models, we used them to score all 685,633 Slovene candidates generated from our bilingual resources as explained in Sect. 3.4. Using the above-mentioned threshold on the models' output, we retained 68,070 candidates. Among these candidates, 5,056 (7 %) correspond to (*literal, synset*) pairs already present in sloWNet 2.2, which means that 63,014 (93 %) new ones were added; as a consequence, 25,102 synsets that were previously empty in sloWNet now have at least one Slovene literal.

Quantitative information on the resulting wordnets (sloWNet 3.0) is provided in the last column of Table 4. In short, sloWNet 3.0 has as much as 141 % more non-empty synsets than before the extension. As far as (*literal, synset*) pairs contained in

Table 5 MaxEnt models for ranking new (literal, synset) candidates, trained on baseline wordnets

Feature	Weight in the model
Semantic similarity	6.24
No. of sources	0.55
No. of distinct English lits.	0.33
Lowest polysemy for Eng. lits.	2.69
No. of tokens of the TL lit.	−1.87
Source: Wikipedia	0.92
Source: English Wiktionary	0.27
Source: Slovene Wiktionary	−0.07
Source: SpeciesWiki	0.10
Source: En–Sl dictionary	0.15
Source: Sl–En dictionary	0.79

the resources are concerned, the increase is even higher: the extension of sloWNet has increased the number of such pairs from 24,081 to 82,721 (+244 %). The evaluation of the newly added (*literal, synset*) pairs is described in Sect. 6.1. Evaluations of the resulting extended resource with respect to other wordnets (a gold standard Slovene wordnet and two other automatically generated wordnet-like semantic repositories) is given in Sects. 6.3 and 6.4. A task-based evaluation in a machine translation setting is given in Sect. 6.5.

5 Cleaning noisy synsets

Despite the satisfying results obtained in the previous section, the technique we used, as all state-of-the art methods for the population of wordnets, is still far from perfect, resulting in noisy synsets. This is why we developed a language-independent, corpus-based approach to detect outliers in automatically generated synsets and filter them out in order to obtain a cleaner, more useful lexico-semantic resource for human use as well as for various NLP tasks (Sagot and Fišer 2012).

The cleaning approach falls within the scope of distributional methods for detecting semantic similarity between words (Lin et al. 2003), but instead of identifying most closely related words according to the contexts they appear in, we start from a (noisy) list of synonym candidates in the form of an automatically induced wordnet. In a way, our task is not very different from the lexical substitution framework (Mihalcea et al. 2010), with the exception that we are most interested in the bottom of the ranked list of potential synonyms. In addition, our notion of synonymy is much stricter because it is our aim to clean all the synsets in an automatically created wordnet, which is very fine-grained.

At the same time, the notion of polysemy that is of key importance for this work is translation-motivated. This means that regardless of the number of synsets a word appears in, the distinction between those senses that are lexicalised differently is the only relevant one in this work.

We focused on identifying and eliminating the most obvious errors in synsets that occurred due to errors in word-alignment of parallel corpora (e.g. misaligned elements of multi-word expressions) and inappropriate word-sense disambiguation of homonymous words (e.g. assigning a valid translation of one sense of a homonymous source word to all its senses). It is precisely these errors in wordnets that have the biggest impact in NLP applications and decrease the value of the resource the most.

Our approach for cleaning noisy synsets relies on a simple hypothesis: lexemes, defined here as (*literal, synset*) pairs, tend to co-occur in corpora with other lexemes that are semantically related, as made explicit by relations between synsets in a wordnet. This is possible because, provided that the wordnet is large enough, this technique can provide a sufficient number of semantically related lexemes for most lexemes with a high precision (the precision of sloWNet 3.0 has been evaluated as 86 %, see Sect. 6).

The method we used for cleaning our wordnets can be divided in two steps:

1. Co-occurrence-based evaluation of the similarity between each nominal occurrence in a large (monolingual) corpus and their possible synsets as provided by the input wordnet;
2. Global assessment of all nominal (*literal, synset*) pairs based on these similarity measures. Note that in the work described here we have restricted our search for outliers to nominal synsets only and we did not take into account multiword literals. This means that we currently consider as literals only tokens tagged as nouns, verbs, adjectives and adverbs.

5.1 Basic co-occurrence-based scoring of (literal, synset) pairs

In order to achieve step 1 in the enumeration above, we first associate each synset from the input wordnet with a set of related synsets, i.e. a subset of all synsets (nominal or not) that are related to the base synset by relation paths of length 0, 1 or 2, based on manually designed relation patterns shown in Fig. 1. Second, we associate each synset pair with the list of its related literals, i.e. all literals that belong to any of its related synsets.

Next, given an occurrence of a nominal literal in the corpus, we look at all literals that co-occur in the same paragraph. We chose paragraphs as contexts because sentences are too small to provide us with enough literals, and because paragraphs constitute the smallest linguistically motivated and typographically discernible text units that are larger than sentences. We then apply a variant of the wordnet-based Lesk algorithm for word sense disambiguation (Lesk 1986). Lesk's algorithm relies on the assumption that the relatedness of two words is proportional to the extent of overlaps of their dictionary definitions. This algorithm was later extended to also use literals from related synsets as well as wordnet usage examples in addition to the dictionary definition (Banerjee and Pedersen 2002). We adapted this idea to our task, by comparing each paragraph where a given literal occurs, represented by its content words (co-occurring literals) with each possible synset for this literal, represented by the literals found in their respective sets of related synsets, as defined above.

More formally, let l be a nominal literal in paragraph p. We refer to the set of all synsets containing a literal l' in the input wordnet as $S(l')$, and to the number of such synsets $|S(l')|$. Let $C(p)$ be the set of (PoS-tagged) literals in paragraph p, and $occ(l', p)$ the number of occurrences of a literal l' in p. We refer to the set of literals related to a synset s in the input wordnet as $R(s)$. Finally, let $length(p)$ be the number of tokens in p. Each (*literal, synset*) pair of the form (l, s), with $s \in S(l)$, receives for paragraph p a local score $local\_score(l, s, p)$ defined as follows:

$$local\_score(l, s, p) = \frac{1}{length(p)} \sum_{l_{\text{rel}} \in C(p) \cap R(s)} \frac{occ(l_{\text{rel}}, p)}{|S(l)|} \quad (1)$$

The corpus-wide score $global\_score(l, s)$ for the (literal, synset) pair (l, s) is then simply the sum of the local scores of each of its occurrences:

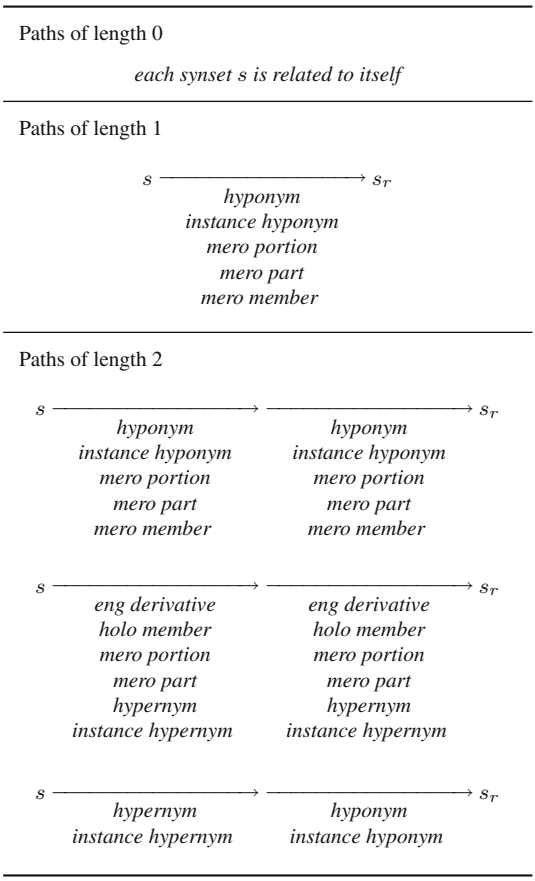

Fig. 1 Relation paths starting from a synset s and leading to its related synsets s_r

$$global\_score(l,s) = \sum_{p} occ(l,p) \cdot local\_score(l,s,p) \tag{2}$$

Let us illustrate this on an example. Consider the Slovene noun *ikona* ('icon (religion)' or 'icon (computer science)'). It appears in 4 synsets in sloWNet 3.0, among which:

- eng-30-07269916-n {*icon*}; the Slovene literal *ikona* is correct in this synset; excerpt of the related literals: *znak*, *točka*, *simbol*, *računalništvo*...('character,' 'pixel,' 'symbol,' 'computer science'...).
- eng-30-03931044-n {*icon*, *ikon*, *image*, *picture*}; the Slovene literal *ikona* is not correct in this synset; related literals: *fotografija*, *podoba*, *predstaviti*, *prikaz*...('photography,' 'image,' 'to represent,' 'representation'...);

In our corpus, the Slovene noun *ikona* occurs 3,488 times. The global score for the correct (*ikona*, eng-30-07269916-n) pair, based on the above-mentioned related literals, is only 1.02, whereas that for the incorrect (*ikona*, eng-30-03931044-n) pair

it is 5.99. This shows that global scores do not necessarily allow us to correctly detect the erroneous (*literal, synset*) pair. Therefore, we take into account additional information, as shown in the next section.

5.2 Extracting outlier candidates for (*literal, synset*) pairs

At this stage, we have for each (*literal, synset*) pair a global score that is the sum of the local scores of its occurrences in the corpus. We first normalise this global score by dividing it by the sum synset_global_score(s) of the global scores of all (*literal, synset*) pairs involving the same synset s. This is used to assess the contribution of a given literal among all literals in s. Let us call $L(s)$ the set of all literals that belong to the synset s in the input wordnet. We define synset_global_score(s) in a straightforward way:

$$\text{synset\_global\_score}(s) = \sum_{l \in L(s)} \text{global\_score}(l, s). \tag{3}$$

The contribution of l to the synset s is then:

$$\text{contribution}(l, s) = \frac{\text{global\_score}(l, s)}{\text{synset\_global\_score}(s)}. \tag{4}$$

This contribution is then normalised by the number of occurrences occ(l) of the literal in the corpus, thus leading to the final score for the (*literal, synset*) pair (l, s):

$$\text{score}(l, s) = \frac{\text{contribution}(l, s)}{\text{occ}(l)}. \tag{5}$$

If we go back to the example given in Sect. 5.1, the synset global score for eng-30-07269916-n is 1.02, and is 234 for eng-30-03931044-n. Their respective contributions are thus 1 and 0.026. Consequently, our last formula leads to a score of 0.287 for (*ikona*, eng-30-07269916-n), whereas the score for (*ikona*, eng-30-03931044-n) is 0.007. The final score now correctly identifies the correct vs. incorrect (literal, synset) pairs. Additional examples are provided together with their manual evaluation in Sect. 6.6, which is, along with Sect. 6.7, dedicated to evaluation and validation procedures of outlier candidates.

6 Evaluation of the results

In previous sections we have described the development of sloWNet: first, we developed a baseline version (sloWNet 2.0), second, we performed some manual improvements of the generated wordnet, and third, we extended the resulting sloWNet (version 2.2) with additional lexical information and identified outliers. Because only the final two steps are novel and because they have contributed to a considerable increase compared to previous versions, we begin with a manual evaluation of the extension step. We evaluate the accuracy of the candidates we obtained as well as the accuracy of the candidates we discarded. Next, we perform

two series of contrastive evaluations of the extended wordnet. With this we will gain insight into the precision and recall of the wordnets we created, before and after the extension. First, we compare it with a small-scale gold standard, a small, manually constructed wordnet for Slovene (SWN). Second, we compare it with other automatically generated wordnets, namely the multilingual Universal WordNet (UWN) (de Melo and Weikum 2009) and the latest version of BabelNet (version 2.0) (Navigli and Ponzetto 2010, 2012). Finally, we illustrate how this extended wordnet performs in task-based settings. In addition, we also provide insights into the quality of the outlier detection task, via manual assessments by an expert and crowdsourcing-based validation results.

In all our evaluation settings we assess if the synset assigned to a given literal is correct (i.e. if it is an appropriate lexicalisation of the concept in question). In order for the candidate to be considered valid, it has to be a perfect match for the assigned synset; if the literal denotes a more general or more specific concept (a hyper- or hyponym) than the concept represented by the synset in question, it is marked as incorrect.

6.1 Manual evaluation of the wordnet extension step

Before evaluating sloWNet as a whole, we wanted to measure the accuracy of our extension approach (see Sect. 4). We therefore randomly selected 400 (literal, synset) candidates and evaluated them manually. Since we have found that the errors performed by the automatic sense assignment step are not very fine-grained, we did not see the need to check inter-annotator agreement. Instead, manual evaluation was performed by a single annotator who used only two tags: "YES" if it was correct to add that literal to the synset, and "NO" if it was wrong, regardless of the reason for the error and its semantic relatedness to the synset. The accuracy of a set of candidates is as usual the proportion of candidates receiving the "YES" tag. Moreover, in order to assess the quality of our scoring technique, we compared the accuracy of the candidates per quartile w.r.t. their certainty scores. The results are shown in Table 6. We observe a strong correlation between the certainty score they received and the accuracy of the candidates, leading us to set the threshold value at 0.1. Other threshold values could have been used: higher values would have provided candidates with an even higher accuracy but the scale of the wordnet extension would have been lower; on the other hand, lower threshold values would have extended our wordnets even more but would have also introduced more noise. The 0.1 value, which corresponds approximately to the upper decile, seemed to provide a good balance. It leads to retaining 68,070 candidates (out of 685,633) that, however, have a precision of only 64 %.[19]

[19] In experiments conducted for applying this extension technique to the French wordnet WOLF, the same 0.1 threshold leads to retaining a higher proportion of candidates, namely 55,159 out of 177,980, which have a much higher precision (83 %). This is related to the archaic words present in the Slovene-English dictionaries we use for extending sloWNet and suggests that this dictionary is not the best resource for wordnet construction but was nevertheless used since it is the only extensive bilingual dictionary available, which is not uncommon in realistic research scenarios.

Table 6 Manual evaluation of 400 (literal, synset) candidates generated during the extension step and manual evaluation of the candidates that were added to sloWNet, out of which 36 (9 %) passed the threshold (score $\geq$ 0.1)

	#candidates	Accuracy (%)	Standard error (%)
All candidates evaluated manually	400	25	2
Candidates passing the threshold (score ≥ 0.1)	**36 (9 %)**	**64**	**8**
Accuracy of the discarded candidates (score < 0.1)	100	21	2
Accuracy in the upper (fourth) quartile	100	44	5
Accuracy in the third quartile	100	32	5
Accuracy in the second quartile	100	13	3
Accuracy in the lower (first) quartile	100	10	3

Accuracy figures are provided with the corresponding standard error, which provides an estimate of the margin of error

6.2 Manual evaluation of the extended wordnet

The most straightforward way to evaluate the accuracy of a wordnet is to randomly select a significant amount of (*literal, synset*) pairs and evaluate them manually. In order to obtain a meaningful per-PoS evaluation, we have decided to evaluate 100 randomly selected (*literal, synset*) pairs per PoS. This also allows for an estimate of the overall accuracy of the extended sloWNet (version 3.0), by weighting per-PoS accuracy scores by the relative number of (literal, synset) pairs for each PoS.

The results of this evaluation are provided in Table 7. It shows that the overall accuracy of sloWNet 3.0 is 82 %. With the proposed method we were able to generate most nominal synsets that at the same time have a very high accuracy (87 %). The only more accurate are adverbial synsets (96 %), which is understandable since they have the lowest degree of polysemy. The results for adjectives (85 %) are comparable to those of nouns, only verbal synsets perform much worse (59 %). On the one hand, this can be expected since verbs are much more polysemous (while average polysemy for nouns in PWN is 1.24, it is 2.17 for verbs[20]), but on the other hand their translations depend on target language syntax much more than translations of nouns or adjectives. This is why they are a much more difficult problem to address with the proposed approach.

6.3 Contrastive evaluation of the extended wordnet against a small-scale gold standard

A direct manual evaluation such as the one described in the previous section leads to precise overall accuracy results. However, such an evaluation does not provide insights into at least two questions:

(i) the recall of sloWNet 3.0, and
(ii) the accuracy of BCS synsets in sloWNet 3.0.

[20] http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html [06.07.2014]

Table 7 Results of the manual evaluation of the extended sloWNet (version 3.0)

PoS	No. of pairs in sloWNet 3.0	Correct pairs (%)	Incorrect pairs (%)
Noun	55,383	87	13
Adjective	12,438	85	15
Verb	14,053	59	41
Adverb	847	96	4
All synsets	82,721	82	18

In order to obtain information on these issues, we compared sloWNet 3.0 to the Slovene WordNet (SWN). SWN is a small-scale manually built gold standard, obtained by validating the results of the preliminary Slovene wordnet construction experiments based on the Serbian wordnet (Erjavec and Fišer 2006). Because it has been developed manually, SWN only contains synsets from the three Base Concept Sets (Tufiş and Cristea 2002).[21] Therefore, evaluating sloWNet 3.0 on those synsets which are not empty in SWN is a first step towards an answer to question (ii) above. Moreover, because SWN has been developed manually, it is reasonable to make the assumption that a non-empty SWN synset contains all Slovene literals that should be found in that synset. In other words, one can estimate the recall of sloWNet 3.0 by comparing its coverage w.r.t. non-empty synsets in SWN. However, such an evaluation is strongly biased, as it is restricted to the above-mentioned BCS synsets only. This evaluation therefore cannot be considered a definitive answer to question (i) above.

The accuracy-oriented evaluation was performed as follows. First, we consider any (*literal, synset*) pair that is found in both SWN and sloWNet 3.0 as correct. Second, in order to assess the accuracy of (*literal, synset*) pairs found in sloWNet 3.0 but not in SWN, we randomly selected 100 such pairs per category and evaluated them manually. For adjectival synsets, there are only 45 such pairs, so we evaluated them all. As there are no adverbial synsets in SWN, no figures can be obtained on such synsets.

The results of this evaluation are given in Table 8. The average accuracy result which we obtain, namely 70 %, is much lower than the overall accuracy of sloWNet 3.0 which we obtained in the previous section, namely 82 %. This is because evaluation against SWN is restricted to BCS synsets which denote general concepts, typically lexicalised with high-frequency vocabulary. Since general, frequent words are typically highly polysemous, they present the biggest challenge in automatic sense assignment, causing the lower accuracy score. In order to confirm this we randomly chose 100 pairs among the 67,393 sloWNet 3.0 pairs that are in synsets which are empty in SWN. This evaluation yielded an accuracy score of 92 %. This latter figure, together with the 70 % accuracy score on pairs from the non-empty SWN synsets, results in a new estimate of 86 % for the new overall accuracy of sloWNet 3.0. The discrepancy between the 82 % obtained in the previous section and the 86 % measured here is thus an artefact of the random selection process

[21] Note that SWN does not contain any adverbial synsets and only a few adjectival synsets.

Table 8 Results of the contrastive accuracy-oriented evaluation of the extended sloWNet (version 3.0) w.r.t. the non-empty synsets in the small-scale manually developed Slovene WordNet (SWN)

PoS	All sloWNet 3.0 pairs in non-empty SWN synsets					
	In sloWNet 3.0 and SWN		In sloWNet 3.0 only		Total	
	No. of pairs	Accuracy (%)	No. of pairs	Accuracy (%)	No. of pairs	Accuracy (%)
Noun	4,971	100	4,943	46	9,914	73
Adjective	71	100	45	51	116	81
Verb	2,239	100	3,059	37	5,298	64
Adverb	0	–	0	–	0	–
All					15,328	70

Table 9 details our results w.r.t. SWN from a recall-oriented perspective. It is difficult to interpret these results because they are computed only on BCS synsets, which contain literals that in general display a higher degree of polysemy (e.g. *plant*), as opposed to literals denoting specialised concepts (e.g. *hellebore*), and therefore cause a negative bias for recall. On the other hand, taking SWN non-empty synsets as necessarily complete would again be an incorrect approximation, causing a positive bias for recall. Table 8 shows a comparison of incomplete SWN synsets with their extended sloWNet 3.0 counterparts where we can see that many correct BCS (*literal, synset*) pairs are found in sloWNet 3.0 but not in SWN.

6.4 Contrastive evaluation of the extended wordnet against two other automatically generated wordnets

Another way to evaluate sloWNet, and more specifically to evaluate the approach we used for building it, is to compare it with comparable resources, namely other automatically generated wordnets. We have evaluated it against two highly multilingual wordnets, namely BabelNet (Navigli and Ponzetto 2010, 2012)[22] and the Universal WordNet (UWN) (de Melo and Weikum 2009).

Even though both UWN and BabelNet are built from the same basic resource as sloWNet, it must be recalled that their aim is to be massively multilingual networks, while we focused on translating the Princeton WordNet from English to an individual language, here Slovene. As a result, the Slovene subpart of UWN is much smaller than sloWNet as it contains only 9,924 (literal, synset) pairs,[23] to be compared with the 82,721 such pairs in sloWNet 3.0. This is not the case with BabelNet (version 2.0), which contains as many as 131,964 literals. However, as we

[22] Note that the first versions of BabelNet did not contain any Slovene literals. Only the recently published BabalNet 2.0 does.

[23] 115 Slovene UWN (*literal, synset*) pairs have a literal that contains at least one comma, which seems to be more a separator between possible literals than part of unique literals. Moreover, some literals include a stress marker (mentioned above and removed from sloWNet since version 2.0). Before evaluating sloWNet 3.0 against the UWN, we "improved" the UWN by correcting these issues. Therefore, our evaluation is in a way biased in favour of UWN.

Table 9 Results of the contrastive recall-oriented evaluation of the extended sloWNet (version 3.0) w.r.t. the small-scale manually developed Slovene WordNet (SWN). See text for a discussion on the relevance of such recall figures

PoS	SWN-only pairs	All sloWNet 3.0 pairs in non-empty SWN synsets			Recall (%)
	No. of pairs	No. of pairs	Accuracy (%)	No. of correct pairs	
Noun	2,669	9,914	73	~7,245	73
Adjective	7	116	81	94	93
Verb	2,311	5,298	64	~3,371	59
Adverb	0	0	–	0	–
All	4,987	15,328	70	~10,710	68

will see, the accuracy of the Slovene subpart of BabelNet is much lower than that of sloWNet 3.0.

Table 10 shows the number of (literal, synset) pairs for each of the 7 possible situations obtained by crossing the presence or absence of a pair in each of the three resources. Comparing sloWNet 3.0 with UWN and with BabelNet respectively, we get the following results:

- Among the 9,924 (literal, synset) pairs in the Slovene subpart of UWN, 5,590 (56 %) are in sloWNet 3.0 as well. On the other hand, 77,131 (*literal, synset*) pairs are in sloWNet 3.0 but not in UWN (i.e. 93 % of sloWNet 3.0).
- Among the 131,964 (*literal, synset*) pairs in the Slovene subpart of BabelNet 2.0, 14,707 (11 %) are in sloWNet 3.0 as well. On the other hand, 69,014 (*literal, synset*) pairs are in sloWNet 3.0 but not in BabelNet 2.0 (i.e. 82 % of sloWNet 3.0).

Overall, as many as 64,663 (*literal, synset*) pairs are only found in sloWNet 3.0. Only 901 pairs are both in BabelNet and in UWN but missing in sloWNet 3.0.

In order to better quantify and analyse the differences between these three resources in terms of accuracy, we carried out a manual evaluation of 50 randomly selected (*literal, synset*) pairs for each of the 7 possible situations mentioned above. The results are shown in Table 10. Based on the results, we can draw the following conclusions:

- The overall accuracy score obtained for sloWNet 3.0 in this evaluation is 88 %. This is to be compared with the 82 % obtained above, which was computed on a larger amount of manually evaluated (*literal, synset*) pairs. This shows that the "real" accuracy of sloWNet 3.0 is in the mid-80's, probably around 85 %. It also shows that in this contrastive evaluation, the differences are significant only if they are at least a few percent higher.
- The overall accuracy of sloWNet 3.0 and UWN are similar, despite the fact that sloWNet 3.0 is much larger than UWN. Or in other words, there are 18 times fewer Slovene UWN-only pairs than there are sloWNet-only pairs.
- The accuracy of the 64,663 sloWNet-only (literal, synset) pairs is around 86 %; this is to be compared with the accuracy of UWN-only and BabelNet-only pairs, which is much lower (72 and 70 % respectively).

Table 10 Comparative results between sloWNet 3.0, UWN and BabelNet 2.0

	In BabelNet 2.0		Not in BabelNet 2.0	
	In UWN	Not in UWN	In UWN	Not in UWN
(a) Detailed results				
In sloWNet 3.0				
No. of pairs	2,239	12,468	3,351	64,663
Accuracy	98 %	98 %	92 %	86 %
Not in sloWNet 3.0				
No. of pairs	901	116,356	3,433	-
Accuracy	100 %	70 %	72 %	

	In BabelNet 2.0	Not in BabelNet 2.0
(b) Contrastive results against BabelNet 2.0		
In sloWNet 3.0		
No. of pairs	14,707	68,014
Accuracy	98 %	86 %
Not in sloWNet 3.0		
No. of pairs	117,257	-
Accuracy	70 %	

	In UWN	Not in UWN
(c) Contrastive results against UWN		
In sloWNet 3.0		
No. of pairs	5,590	77,131
Accuracy	94 %	88 %
Not in sloWNet 3.0		
No. of pairs	4,334	–
Accuracy	78 %	

	sloWNnet 3.0	BabelNet 2.0	UWN
(d) Overall scores			
No. of pairs	82,721	131,964	9,924
Accuracy	88 %	73 %	87 %

- The three approaches used for building these resources are complementary in the sense that virtually all 2,239 (literal, synset) pairs that are common to all three resources are correct; what is more, most pairs that are in at least two of the three resources are correct; the lowest score in that regard concerns pairs that are in sloWNet 3.0 and in UWN but not in BabelNet 2.0 (92 % accuracy);
- BabelNet, which is very large, is also quite noisy, and therefore not fully reliable; this can be seen from the fact that only 70 % of BabelNet pairs that are not in sloWNet are correct, whereas 78 % of UWN pairs that are not in sloWNet are correct.

Table 11 Slovene (lowercased) literals found in sloWNet 3.0, UWN and BabelNet 2.0 in the synset 1503061-n (PWN literals: {*bird*}; PWN definition: *warm-blooded egg-laying vertebrates characterised by feathers and forelimbs modified as wings*)

Literal	sloWNet 3.0	UWN	BabelNet 2.0	Comment
aves			x	Wrong (class Aves, Latin)
ptiči		x	x	Wrong (plural form)
ptice		x	x	Wrong (plural form)
ptič	x	x	x	Correct
ptica	x	x	x	Correct
seznam ptičev			x	Wrong ('list of birds,' from Wikipedia)
ptičev			x	Wrong (genitive plural form)
seznam ptic			x	Wrong ('list of birds,' from Wikipedia)
avafauna			x	Wrong (Latin)

More specifically, and apart from real disambiguation issues, errors among sloWNet-only pairs are mostly related to strange and/or archaic words, whereas errors among UWN-only pairs and BabelNet-only pairs are often related to normalisation errors: English literals, titles of Wikipedia pages that are not literals (e.g., *Seznam Arheoloških Dob* 'List of archeological ages'), Slovene words that are correct semantically but are in the wrong part of speech, in feminine form, preceded by a numeral, followed by a dot or a disambiguation word from Wikipedia (e.g., *Mars (bog)* 'Mars (god)'), etc. Table 11 gives an example of a synset with all literals from sloWNet 3.0, UWN and BabelNet 2.0, including a short comment for each literal. In total, we find 9 literal candidates for this synset in all three resources. sloWNet contains two, both of which are correct. UWN contains the two correct ones plus two incorrect ones and the noisiest is BableNet which contains 9 literal candidates, including the two correct ones.

Given the results of this section and the previous one, we believe that sloWNet 3.0 can be considered as the most adequate Slovene wordnet to date.

6.5 Task-based evaluation of the extended wordnet

The evaluations presented in the previous sections provide direct insights into the quality of sloWNet 3.0. However, developing a wordnet is not necessarily a goal *per se*, which is why we decided to carry out a small-scale task-based evaluation as well. In this section, we present the results of an evaluation of the extended sloWNet which was used to improve machine translation at the lexical level (Fišer and Vintar 2010). Mistranslations often arise due to inadequate word-sense disambiguation of polysemous words and detection of multi-word expressions, and parallel wordnets can help with both problems.

In order to examine the importance of correct sense identification in an MT task, we created a small parallel corpus of 500 articles from the EU news portal that contained about 120,000 Slovene and 140,000 English tokens. We lemmatised, PoS-tagged and sentence-aligned the corpus and then semantically disambiguated all

WordNet:	081114 koza → goat, caprine animal
Presis:	Almost 360 million of pigs, sheep, a **smallpox** and cattle and more billion of poultry execute every year in European Union because of meat.
Human:	Every year nearly 360 million pigs, sheep, **goats** and cattle and several billion poultry are killed for their meat in the EU.

Fig. 2 An example of an improved lexical translation in MT with sloWNet

literals in the corpus with the freely-available graph-based UKB tool (Agirre and Soroa 2009), i.e. each of them received a unique synset id depending of their meaning in context. In sense assignment, UKB takes into account only direct (*literal, synset*) pairs, not their hypernyms or hyponyms, which could also be utilised in a future extension of the experiment. Next, we machine-translated both parts of the corpus with two MT systems; the rule-based Presis[24] and the statistical GoogleTranslate,[25] and compared the machine-translated solutions with human translations, which we treated as a gold standard, and translation equivalents obtained via synset ids from the two wordnets.

A comparison of MT-output, WN-equivalents and the human translations show that there were about 38,000 polysemous tokens (32 % of all the polysemous tokens in the corpus) in the Slovene part of the corpus. About 40 % of them were translated identically by both MT systems, wordnet-based WSD and the gold standard. But there were 1,558 tokens (4.1 % of all the polysemous tokens in the corpus) for which Slo → Eng Presis and 867 Google translations (2.3 % of all the tokens in the corpus) did not match the translations in the gold standard but were assigned the correct wordnet sense. This is illustrated in Fig. 2 where the word *koza* was incorrectly translated by Presis as *smallpox*, while *goat*, the correct translation, was suggested by sloWNet.

When translating in the opposite direction, there were 48,000 polysemous tokens (34 % of all the tokens in the corpus) and only about 32 % of them were translated identically by both MT systems, wordnet-based WSD and the gold standard. Interestingly, the discrepancy between semantic misrepresentation of polysemous tokens by the MT systems with respect to wordnet-based WSD was even larger: 3,730 tokens (2.7 % of all the polysemous tokens in the corpus) that were mistranslated according to the gold standard by Presis were correctly disambiguated with sloWNet, and 901 (1.9 % of all the polysemous tokens in the corpus) by Google.

In a random sample of 200 sentence pairs that were manually checked, there were also 166 multi-word expressions which were not identified as such by the machine-translation system and therefore incorrectly translated, but were found in wordnet, e.g., *biotska raznovrstnost* 'biotic diversity' instead of *biodiversity*; *vezani les* 'tied wood' instead of *plywood*.

This analysis shows that the extended sloWNet, when used in parallel with PWN, can be a very useful resource in MT systems, especially with polysemous words and multi-word expressions that are a major source of errors by MT systems, rule-based

[24] http://presis.amebis.si/prevajanje/ [06.07.2014].

[25] http://translate.google.com/ [06.07.2014].

and statistical alike. The reason why GoogleTranslate performed better than Presis overall is that Google's MT uses parallel texts found on the web, which was also the source of our parallel corpus and had probably already been detected by Google.

6.6 Manual evaluation of outlier candidates

As mentioned above, the overall error rate in the extended sloWNet has been evaluated as being about 15 %, i.e. around 12,000 incorrect (*literal, synset*) pairs. Given our set of outlier candidates, we have empirically chosen a threshold on the outlier score such that the number of candidate outliers has the same order of magnitude than the estimated number of erroneous (literal, synsets) pairs. This resulted in a threshold of 4×10^{-6}, thus generating 12,578 candidate outliers, i.e. approximately one third of all outlier candidates.

We manually evaluated a random sample of 100 candidate outliers. Among these, the proportion of (*literal, synset*) pairs which have correctly been detected as errors is 64 %. These figures can be compared with the estimated overall error rates in the input wordnets, namely 15 % as recalled above. Considering that we expanded sloWNet using thresholds that led to a reasonable balance between recall (more candidates is better) and precision (less erroneous candidates is better), we inevitably included erroneous candidates as well. The fact that our outlier detection method manages to suggest candidate outliers out of which 64 % are real errors is very good: first, it means that our outlier detection algorithm manages to spot many more errors than randomly selecting (*literal, synset*) pairs; second, this outlier detection algorithm relies heavily on its input wordnet, i.e. on the extended wordnet: in other words, the information available to the outlier detection algorithm includes the entire extended wordnet, something that the wordnet extension algorithm could obviously not rely on.

Examples of candidate outliers from sloWNet extracted from our manual evaluation data are shown in Table 12. Apart from the synset and literal we indicate the corresponding score as well as the outcome of manual evaluation in which the

Table 12 Example of manually evaluated candidate outliers

Literal	Synset id	English literals in the synset	Score ($\times 10^3$)	Eval
aktiva	05154517	*plus, asset*	0.002	YES
cilj	05868477	*end*	0.004	YES
dan	15113229	*period, period of time, time period*	0.001	NO
dan	15157225	*day*	0.004	NO
dan	06210791	*light*	0.003	YES
dan	06832572	*n, N*	0.004	YES
datelj	15159583	*date, day of the month*	0.000	YES
del	05867413	*division, part, section*	0.003	NO
del	13809207	*constituent, component (part), part, portion*	0.003	NO
delež	05256358	*part, parting*	0.004	YES

We show the first 10 pairs in the evaluation data set, which was randomly extracted from the full sets of candidate outliers

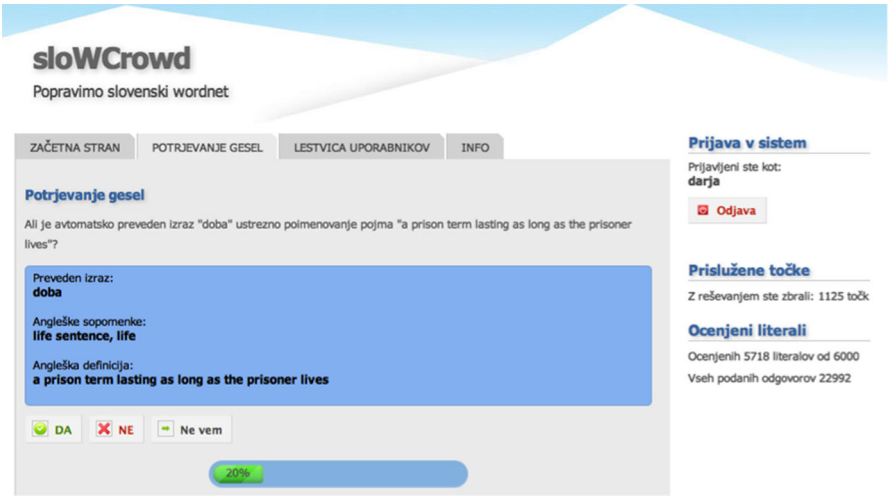

Fig. 3 An example of literal validation in sloWCrowd

"YES" label means that the (literal, synset) pair has been correctly detected as incorrect, while the "NO" label means that the (*literal, synset*) pair is indeed correct, and that its detection as a candidate outlier is erroneous.

6.7 Crowdsourcing-based validation of outlier candidates

The identified outlier candidates were earmarked for manual examination during which they would be rejected as errors and therefore deleted from the wordnet or validated as correct and kept in the resource. In order to facilitate manual work, we have developed a simple on-line tool called sloWCrowd (Tavčar et al. 2012) that works on the principle of crowdsourcing. The tool is open-source and based on popular technologies, such as PHP and MySQL. It consists of an administrator and a user interface. The administrator interface enables the creation of crowdsourcing projects, management of on-going projects and export of the results, while the user interface allows users to vote on the (in)correctness of the randomly displayed literals. The reliability of each user is checked against a gold standard so that the users with a very low accuracy can be automatically excluded from the final results. In order to achieve as high consensus about the answers as possible, the same question is repeated five times, each time to a different user, and the final decision is based on the majority vote. The user interface for validating outlier candidates is shown in Fig. 3 where the user is asked the following question: *Is the automatically translated expression X an appropriate lexicalisation of the concept Y?*

To date,[26] 275 users have provided 34,867 answers, including answers to gold standard requests, and have validated 7,276 outlier candidates. On average, each user provided 126.79 answers, whereas the maximum number of answers provided

[26] July 6, 2014.

by a single user is 4,200 and the minimum is 1. Users' accuracy ranges between 25 and 100 %, but is 79.72 % on average. According to the majority vote, 44 % of the outlier candidates have been voted as correct and 56 % as incorrect. This is in line with the results of manual evaluation of a sample of 400 outlier candidates, 64 % of which have been considered as genuine errors (see Sect. 6.6). All the outlier candidates that were rejected by the majority of the users were deleted from sloWNet. The crowdsourcing task will continue until we collect votes for all 12,578 candidate outliers we obtained from our automatic outlier detection procedure. Once all the votes are collected, the outlier candidates with the majority negative vote will be deleted from sloWNet and version 4.0 will be announced.

7 Conclusions and future work

In this paper, we have described the different resources and techniques we used for automatic construction, extension and cleaning of sloWNet, a wordnet for Slovene. We first outlined the construction of the baseline wordnet based on bilingual lexicons extracted from Wikipedia, Wiktionaries and other bilingual resources which we used for translating monosemous literals, and word-aligned parallel corpora for translating and disambiguating polysemous literals. Then we described a follow-up experiment in which we used the same bilingual lexicons much more exhaustively than the baseline wordnets. By using various features, including distributional similarity, we were able to reuse the same resources for translating and disambiguating polysemous literals as well, which had been dealt with only by word-aligned corpora up to this point. This enrichment step has increased the number of non-empty synsets in sloWNet from 17,817 to 42,919. The number of (*literal, synset*) pairs in sloWNet went up from 24,081 to 82,721 (+244 %).

The resulting wordnet was then carefully evaluated, both in terms of accuracy of the content and as a resource in a machine-translation setting. The accuracy of (*literal, synset*) pairs is estimated at approximately 85 %. These figures show that the enhanced resource has a much higher coverage than the baseline wordnet and that it outperforms the gold Slovene WordNet.

The latest version of sloWNet has been uploaded to *sloWTool*,[27] a freely available tool that incorporates browsing, editing and visualisation of wordnet content with hyperbolic graphs and images (Fišer and Novak 2011). It is freely available and based on MySQL and PHP technologies, which makes the tool light-weight, portable and efficient. Scripts for automatic database transformations from and into several standardised formats, such as DEBVisDic XML and LMF, are provided so that a wordnet for another language can be imported at any time. The on-line browser is simple to use for non-experts but also enables advanced searching and view settings for expert users that can enter complex search queries and decide which fields to display as well as toggle between a mono- and a multilingual option. Through sloWTool, sloWNet is now available to language students, translators and other linguists who can examine the Slovene lexical inventory and semantic

[27] http://nl.ijs.si/slowtool/ [06.07.2014].

Fig. 4 An example of a Slovene wordnet synset {*prst*, *zemlja*} with the corresponding English (in *red*) and French (in *black*) synonyms, an English definition and usage examples, and semantic relations. (Color figure online)

network which they can also compare to English and French since wordnets for these languages are cross-aligned via the Princeton WordNet synset IDs and are available on the sloWTool as well (Fig. 4).

In the future we plan to work in five complementary directions. First, more attention should be given to the multi-word expressions, such as phrasal verbs, compound nouns and idiomatic expressions, since they represent a substantial segment of our semantic repository and pose a major obstacle in NLP applications. Second, the extraction of lexico-semantic information from Wikipedia and Wiktionary can be improved even further by adding definitions and examples to the created wordnets as well as extend and validate the current network by mining semantically related words from article bodies. Third, we have already started adapting the alignment-based approach to work with non-parallel texts (Fišer et al. 2012), which is a very promising line of research as large comparable corpora are much easier to obtain from the rich web data. Fourth, there are many more features that could be used for lexical disambiguation still keeping our development process lightweight (i.e. without the need for advanced NLP tools that are rarely available for most languages, such as parsers or WSD systems). Such features could include Lesk-like measures for comparing contexts of definitions or glosses; similarity between cognates, etc. And last but not least, since our approach has already proven efficient and useful for two languages as different as French and Slovene (Sagot and Fišer 2008; Fišer and Sagot 2008; Sagot and Fišer 2011, 2012a, b), for which the amount and nature of the available sources is very different as well, we would like to create wordnets for other under-resourced languages, such as Croatian.

We believe the work presented in this paper has two main consequences. First, it shows that it is possible to build large-scale reliable wordnets with fully automatic approaches (although manual work was involved in intermediate steps of the construction process, it has affected only a small number of senses included in the latest version of sloWNet). Second, this work has resulted in a freely available lexical semantic resource for a language that was lacking such a resource, which is large and accurate enough to be used in real NLP applications. The developed sloWNet is distributed under the Creative Commons BY-SA 3.0 licence at http://nl.ijs.si/sloWNet.

Acknowledgments The work described in this paper was funded in part by the French–Slovene PHC PROTEUS project 22718UC "Building Slovene–French linguistic resources: parallel corpus and wordnet" (2010–2011), by the French national grant ANR-09-CORD-008 "EDyLex" (2010–2013) and by the Slovene national postdoctoral grant Z6-3668.

References

Agirre, E., & Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the association for computational linguistics (EACL'09)*, Athens, Greece, pp. 33–41.

Arhar, Š., & Gorjanc, V. (2007). Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa (The FidaPLUS corpus: A new generation of the Slovene reference corpus). *Jezik in slovstvo*, *52*(2), 95–110.

Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Computational linguistics and intelligent text processing*, (pp. 136–145). Berlin: Springer.

Bernhard, D., & Gurevych, I. (2009). Combining lexical semantic resources with question and answer archives for translation-based answer finding. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP (ACL '09)*, Suntec, Singapore, pp. 728–736.

Bond, F., & Foster, R. (2013). Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st annual meeting of the association for computational linguistics*, Sofia, Bulgaria, pp. 1352–1362.

Carpuat, M., & Wu, D. (2007). Improving statistical machine translation using word sense disambiguation. In *The 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007)*, pp. 61–72.

Casses, B. (2010). Final paper of research experience for undergraduates for artificial intelligence, natural language processing and information retrieval, The University of Colorado at Colorado Springs. http://www.cs.uccs.edu/~jkalita/work/reu/REUFinalPapers2010/Casses.pdf.

Copestake, A., Sanfilippo, A., Briscoe, T., & de Paiva, V. (1993). The ACQUILEX LKB: An introduction. In T. Briscoe, A. Copestake, & V. de Paiva (Eds.), *Inheritance, defaults and the lexicon* (pp. 148–163). New York, NY: Cambridge University Press.

Cuadros, M., & Rigau, G. (2006). Quality assessment of large scale knowledge resources. In *Proceedings of the 2006 conference on empirical methods in natural language processing (EMNLP '06)*, Sydney, Australia, pp. 534–541.

Daumé III, H. (2004). Notes on CG and LM-BFGS optimization of logistic regression. Paper available at http://pub.hal3.name#daume04cg-bfgs, implementation available at http://hal3.name/megam/

Declerck, T., Pérez, A.G., Vela, O., Gantner, Z., & Manzano-Macho, D. (2006). Multilingual lexical semantic resources for ontology translation. In *Proceedings of the international conference on language resources and evaluation (LREC 2006)*, Genova, Italy.

de Melo, G., & Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on information and knowledge management (CIKM '09)*. ACM, New York, NY, United States, pp. 513–522.

Diab, M. (2004). The feasibility of bootstrapping an Arabic WordNet leveraging parallel corpora and an English wordnet. In *Proceedings of the Arabic language technologies and resources*.

Dyvik, H. (2002). Translations as semantic mirrors: From parallel corpus to wordnet. In (2002). In *Post-proceedings of the ICAME 2002 conference (revised version)*, Gothenburg, Sweden.

Erjavec, T., & Fišer, D. (2006). Building Slovene WordNet. In *Proceedings of the international conference on language resources and evaluation (LREC 2006)*, Genova, Italy.

Erjavec, T., Ignat, C., Pouliquen, B., & Steinberger, R. (2005). Massive multi lingual corpus compilation: Acquis communautaire and totale. In *2nd language & technology conference*, April 21–23, 2005, Poznań, Poland. Vetulani, Z. (ur.). Human language technologies as a challenge for computer science and linguistics: in memory of Maurice Gross and Antonio Zampolli: proceedings. Poznań: Wydawnictwo Poznańskie Sp. z o.o., 2005, pp. 32–36.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fišer, D., Ljubešić, N., & Kubelka, O. (2012). Addressing polysemy in bilingual lexicon extraction from comparable corpora. In N.C.C. Chair, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), *Proceedings of the eight International conference on language resources and evaluation (LREC 2012)*, Istanbul, Turkey.

Fišer, D., & Novak, J. (2011). Visualizing sloWNet. In *Proceedings of the conference on electronic lexicography in the 21st century: New applications for new users (eLEX2011)*, Bled, Slovenia.

Fišer, D., & Erjavec, T. (2009). Semantic concordances for Slovene. *Cognitive Studies - Études cognitives*, *9*, 89–100.

Fišer, D., & Sagot, B. (2008). Combining multiple resources to build reliable wordnets. In *Proceedings of the 11th international conference on text, speech and dialogue (TSD 2008)*, Brno, Czech Republic.

Fišer, D., & Vintar, Š. (2010). Uporaba wordneta za boljše razdvoumljanje pri strojnem prevajanju. In *Proceedings of the 13th international multiconference information society—IS 2010.*

Fung, P. (1995). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd annual meeting on association for computational linguistics (ACL '95)*, Cambridge, Massachusetts, United States, pp. 236–243.

Gabrilovich, E., & Markovitch, S. (2006). Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st national conference on artificial intelligence (AAAI'06)*. AAAI Press, pp. 1301–1306.

Grad, A., & Leeming, H. (1999). *Slovene–English dictionary*. Zagreb: DZS.

Grad, A., Škerlj, R., & Vitorovič, N. (1999). *English–Slovene dictionary*. Zagreb: DZS.

Harabagiu, S., Moldovan, D., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Girju, R., Rus, V., & Morarescu, P. (2000). Falcon: Boosting knowledge for answer engines. In *Proceedings of TREC-9*, pp. 479–488.

Ide, N., Erjavec, T., & Tufiş, D. (2002). Sense discrimination with parallel corpora. In *Proceedings of the ACL'02 workshop on word sense disambiguation: Recent successes and future directions (WSD '02)*, Philadelphia, Pennsylvania, United States, pp. 61–66.

Kirkpatrick, B. (1987). *Roget's thesaurus of English words and phrases*. Penguin: Penguin reference books.

Knight, K., & Luk, S. K. (1994). Building a large-scale knowledge base for machine translation. In *Proceedings of the twelfth national conference on artificial intelligence (AAAI '94)*, Seattle, Washington, United States, pp. 773–778.

Korošec, T., Fekonja, M., Jehart, A., Pečelin, F., & Ulčar, M. (2002). *Vojaški slovar*. Ljubljana: Ministrstvo za obrambo.

Krstev, C., Pavlović-Lažetić, G., & Obradović, I. (2004). Using textual and lexical resources in developing serbian wordnet. *Romanian Journal of Information Science and Technology*, *7*(1–2), 147–161.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on systems documentation (SIGDOC'86)*, Toronto, Canada, pp. 24–26.

Lin, D., Zhao, S., Qin, L., & Zhou, M. (2003). Identifying synonyms among distributionally similar words. In *Proceedings of the 18th international joint conference on artificial intelligence (IJCAI 2003)*, Acapulco, Mexico, pp. 1492–1493.

Liu, H. (2003). Unpacking meaning from words: A context-centered approach to computational lexicon design. In P. Blackburn, C. Ghidini, R. M. Turner, & F. Giunchiglia (Eds.), *Modeling and using context: Fourth international and interdisciplinary conference, context 2003*. Springer, Stanford, California, United States, pp. 218–232.

Matuszek, C., Cabral, J., Witbrock, M., & Deoliveira, J. (2006). An introduction to the syntax and content of Cyc. In *Proceedings of the 2006 AAAI spring symposium on formalizing and compiling background knowledge and its applications to knowledge representation and question answering*, pp. 44–49.

Mihalcea, R., Sinha, R., & McCarthy, D. (2010). Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th international workshop on semantic evaluation (SemEval 2010)*. Los Angeles, California, United States, pp. 9–14.

Nastase, V. (2008). Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP '08)*. Honolulu, Hawaii, pp. 763–772.

Navigli, R., & Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, Uppsala, Sweden, pp. 216–225.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217–250.

Nie, J. Y. (2010). *Cross-language information retrieval synthesis lectures on human language technologies*. San Rafael, CA: Morgan & Claypool Publishers.

Orav, H., & Vider, K. (2004). Concerning the difference between a conception and its application in the case of the estonian wordnet. In *Proceedings of the 2nd international conference of the Gobal WordNet Association (GWC-2004)*, Brno, Czech Republic, pp. 285–290.

Pianta, E., Bentivogli, L., & Girardi, C. (2004). Fighting arbitrariness in wordnet-like lexical databases—A natural language motivated remedy. In *Proceedings of the 1st international conference of the Global WordNet Association (GWC-2002)*. Mysore, India.

Ponzetto, S. P., & Navigli, R. (2009). Large-scale taxonomy mapping for restructuring and integrating wikipedia. In *Proceedings of the 21st international jont conference on artifical intelligence (IJCAI'09)*, Pasadena, California, United States, pp. 2083–2088.

Reiter, N., Hartung, M., & Frank, A. (2008). A resource-poor approach for linking ontology classes to Wikipedia articles. In J. Bos & R. Delmonte (Eds.), *Semantics in text processing. STEP 2008 conference proceedings, research in computational semantics*. College Publications, pp. 381–387.

Resnik, P., & Yarowsky, D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the ACL SIGLEX workshop on tagging text with lexical semantics: Why, what, and how?*, Washington, DC, United States, pp. 79–86.

Richardson, S. D., Dolan, W. B., & Vanderwende, L. (1998). Mindnet: Acquiring and structuring semantic information from text. In *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics* , Montreal, Canada, pp. 1098–1102.

Rudnicka, E., Maziarz, M., Piasecki, M., & Szpakowicz, S. (2012). A strategy of mapping Polish Wordnet onto Princeton Wordnet. In *Proceedings of COLING 2012: Posters*. Mumbai, India, pp. 1039–1048.

Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005). Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In *Proceedings of advances in web intelligence*.

Sagot, B., & Fišer, D. (2008). Building a free French wordnet from multilingual resources. In *Proceedings of ontolex 2008*, Marrakech, Morocco.

Sagot, B., & Fišer, D. (2011). Extending wordnets by learning from multiple resources. In *LTC'11: 5th language and technology conference*. Poznań, Pologne. http://hal.inria.fr/hal-00655785

Sagot, B., & Fišer, D. (2012a). Automatic extension of WOLF. In *Proceedings of the 6th international Global Wordnet Conference (GWC2012)*, Matsue, Japan.

Sagot, B., & Fišer, D. (2012b). Cleaning noisy synsets. In *Proceedings of the international conference on language resources and evaluation (LREC 2012)*, Istanbul, Turkey.

Sornlertlamvanich, V. (2010). Asian wordnet: Development and service in collaborative approach. In *Proceedings of the 5th international conference of the Global WordNet Association (GWC-2010)*, Mumbai, India.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from Wikipedia and wordnet. *Journal of Web Semantics*, *6*(3), 203–217.

Tavčar, A., Fišer, D., & Erjavec, T. (2012). slowcrowd: Orodje za popravljanje wordneta z izkoriščanjem moči množic. In *Proceedings of the 8th language technologies conference, within the proceedings of the 15th international multiconference information society (IS 2012)*, Vol. C, Ljubljana, Slovenia, pp. 197–202.

Tiedemann, J. (2003). *Recycling translations—Extraction of lexical data from parallel corpora and their application in natural language processing*. Ph.D. thesis, Uppsala Universitet, Uppsala, Sweden (Studia Linguistica Upsaliensia 1).

Tufiş, D. (2000). BalkaNet—Design and development of a multilingual balkan wordnet. *Romanian Journal of Information Science and Technology Special Issue*, *7*, 107–124

Tufiş, D., & Cristea, D. (2002). Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet. In *Proceedings of LREC 2002 workshop on wordnet structures and standardisation*, Las Palmas, Spain, pp. 35–41.

Tufiş, D., Koeva, S., Erjavec, T., Gavrilidou, M., & Krstev, C. (2009). Building language resources and translation models for machine translation focused on south Slavic and Balkan languages. In Machačová, J., & Rohsmann, K. (Eds.), *Scientific results of the SEE-ERA.NET pilot joint call*, pp. 37–48.

Vossen, P. (Ed.). (1999). *EuroWordNet : A multilingual database with lexical semantic networks for European languages*. Dordrecht: Kluwer.

Weisscher, A. (2013). GWA base concepts. http://globalwordnet.org/gwa-base-concepts/

Widdows, D., & Ferraro, K. (2008). Semantic vectors: A scalable open source package and online technology management application. In *Proceedings of the international conference on language resources and evaluation (LREC 2008)*, Marrakech, Morocco.

Wong, S. H. S. (2004). Fighting arbitrariness in wordnet-like lexical databases—A natural language motivated remedy. In *Proceedings of the 2nd international conference of the Global WordNet Association (GWC-2004)*. Brno, Czech Republic, pp. 234–241.

Yokoi, T. (1995). The EDR electronic dictionary. *Communications of the ACM*, *38*(11), 42–44.