GROUNDING SLOWNET ON SLOVENE CORPUS DATA

Darja FIŠER
Department of Translation, Faculty of Arts, University of Ljubljana

Maciej PIASECKI
Bartosz BRODA
Department of Artificial Intelligence, Institute of Informatics, Wroclaw University of Technology

Fišer, D., Piasecki, M., Broda, B. (2013): Grounding sloWNet on Slovene corpus data. Slovenščina 2.0, 1 (2): 82–112.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_05.pdf.

Wordnets can be translated from another language or can be built from corpus evidence. The transfer approach is easier and quicker, which is why it has been most widely used. However, it has a big disadvantage that the created resource does not necessarily reflect the language in question. This is why in this paper we test a language-motivated approach that uses linguistically annotated corpus data and basic statistical methods to extract lists of semantically similar words that are then incorporated into the wordnet for Slovene. The approach was originally developed for Polish but because the algorithm itself is language-independent and can use minimally annotated corpus resources in any language, it is also attractive for other languages that are still lacking an extensive wordnet or a similar semantic lexicon. An important advantage of the approach is that it relies on real linguistic evidence harvested from a corpus, yielding a linguistically sound organization of the vocabulary. As all the previous approaches used for the construction of Slovene wordnet were transfer-based and relied on the English Princeton WordNet, the encouraging results obtained in the presented experiment will be a welcome complement to the existing semantic network.

Keywords: lexical semantics, wordnet, semantic similarity, semantic relations

1 INTRODUCTION

sloWNet, a wordnet for Slovene, has been developed in a number of steps, taking advantage of several types of available bi- and multilingual language resources, such as bilingual dictionaries, parallel corpora and Wikipedia (Fišer, Sagot 2008). All these approaches have in common that they follow the so called transfer approach (Vossen 1999), which means that they take over the structure of Princeton WordNet (Fellbaum 1998), the oldest and most extensive existing wordnet that was developed for English, and find Slovene equivalents for the same set of concepts.

However, despite all its attractive advantages, such as ease of construction and cross-lingual alignment, the transfer approach also suffers from some serious drawbacks, such as conceptual and lexical dependencies on the source language, which may in an extreme case result in an arbitrary and unrepresentative resource for the target language (see Vider 2004, Wong 2004). The discrepancies between the source and the target language are most visible when they do not share semantically identical equivalents, which is caused by two phenomena (Bantivogli et al. 2004): (1) *lexical gaps* (a concept which is lexicalized in the source language can only be translated descriptively into the target language) and (2) *denotation differences* (the most suitable translation equivalent of a lexical unit is more specific or more generic than its source counterpart).

In a study that analyzed the results of the transfer model for the wordnet subtree for the semantic field of *Relatives* for Slovene, the following disadvantages of the approach have been identified (Fišer 2005):

- **Culture-specific Concepts**: because some synsets in Princeton WordNet are ideologically or religiously motivated, they are not suitable for inclusion in the Slovene resource (e.g. *Virgin Mary* as the hyponym of *mother*);

- **Denotation Differences**: because there are 14 different English expressions for *ancestor* that are organized into 7 different synsets in Princeton WordNet, as many as 4 levels in the transferred Slovene tree contain identical synsets with the only existing translation equivalent *prednik*, the granularity of which is not linguistically justified;
- **Lexical Gaps**: concepts, such as *antediluvian* and *empty nester* are not lexicalized in Slovene and can only be translated descriptively, which has little practical value in the created resource and therefore does not justify the inclusion in a linguistically sound network; and
- **Semantic Relations**: there are some inconsistencies in Princeton WordNet, such as the lumping of neutral and marked synonyms for the concept *grandfather* on the one hand, while splitting them into several synsets for the concept *father* on the other. Since the transfer model preserves the original structure, such inconsistencies are inherited in the Slovene tree as well. An indication that not all the relations among concepts are language-independent is the case of *father-in-law*, which is a hyponym of *father* in Princeton WordNet. The Slovene equivalent *tast* would fit much better under *in-laws* because unlike *father*, *father-in-law* is not a *blood relative*. The next big issue with the hyper-/hyponymy relation is the unsystematic treatment of female nouns in Princeton WordNet. For example, while *forefather* and *foremother* are co-hyponyms, *ancestress* is a hyponym of the *ancestor*.

Due to these shortcomings the work presented in this paper does not tackle the problem of wordnet creation using the transfer model but instead approaches the task from a completely different angle and extracts all the relevant lexico-semantic information from the largest Slovene reference corpus Gigafida (Logar Berginc, Šuster 2009). As a result, we obtain language-motivated lists of semantically related words and a linguistically sound organization of the vocabulary. We achieve this by adapting the wordnet expansion algorithms, originally developed for Polish, to Slovene in order to test whether they work for another language as well. With the analysis of the first results we also wish to outline further refinements and enhancements of the approach for future work on fully automated methods of wordnet expansion for Slovene.

This paper is structured as follows: in the next section we present related work. Then, we focus on the resources and tools that were used in the experiment. In Section 4 we give an overview of the experimental setup, evaluate and discuss the results. We then conclude the paper with some final remarks and ideas for future work.

2 RELATED WORK

The task of extending a wordnet with additional literals or synsets is typically performed in two phases: first, descriptions of lexico-semantic relations are extracted from a text corpus, and second, the acquired knowledge is used to identify the most appropriate places for each new literal in the existing semantic network. Lexico-semantic relations are represented by sets of word pairs that can be extracted by a range of methods, where most of them follow two main paradigms: the one based on *Lexico-syntactic Patterns* (Hearst 1992) and those that follow *Distributional Semantics* (Harris 1968), briefly described below. The corpus can include structured text, e.g. in the Wikipedia style, and its structure can be utilized during relation extraction, but here we focus on methods assuming unstructured text on input.

The pattern-based approaches rely on a list of lexico-syntactic patterns in which two lexical units frequently occur in an identifiable lexical semantic relation, e.g. the pattern *NP1 is a kind of NP2* extracts a pair of NPs or their heads as a hypernym-hyponym pair. Manually constructed patterns were first applied to text corpora by Hearst (1992) for the extraction of hypernyms. Apart from manual construction, patterns can be statistically learned from the corpus, e.g. (Pantel, Pennacchiotti 2006). Patterns are language-dependent to some extent, as they require some form of morpho-syntactic processing, e.g.

(Piasecki et al. 2009), especially for Slavic languages. What is more, attempts to extract relations other than hypernyms were less successful.

On the other hand, Distributional Semantics (Harris 1968) stipulates that the similarity of distributions of some words across different lexico-syntactic contexts is evidence of a close semantic relation among those contexts. The stronger the similarity, the closer the meanings of the lexical units are. A context can be limited to a block of text of *k* words surrounding the given word *w* or a sentence including *w*. A context can be described simply by other words occurring in it as features or by words linked to *w* by particular lexico-syntactic relations – in this case a feature is an occurrence of the pair: word and relation linking it to *w,* e.g. "modified by *red*" or "a subject of *ride*". The value of a *Measure of Semantic Relatedness* (MSR) is calculated for the words *x* and *y* by comparing the frequencies of their occurrences with different features. In the case of features based on syntactic relations and contexts limited to sentences, MSR produces values that are more correlated with wordnet-like lexico-semantic relations, i.e. higher values are produced by a MSR for pairs of synonyms, hypernyms, meronyms, etc. MSRs based on simple word co-occurrences as features have a tendency to express more associative semantic relations.

Unlike pattern-based approaches, which are limited only to the words that co-occur in a particular pattern, Distributional Semantics techniques can be used for almost any word pair, i.e. both words must occur with a minimal frequency in order to be able to obtain their good descriptions and comparison. It is hard to find a theoretically motivated minimal frequency because a corpus can include errors or accidental word associations due to, for example, the use of metaphors. Our experience with building MSRs for Polish showed that for most words this threshold is somewhere between 100 and 200 occurrences, (see Piasecki et al. 2009). Because high recall is an important desideratum in the work presented in this paper, we have opted for MSR as the main source of information.

Many ways of MSR computing have been proposed (see Ruge 1992; Lin, Pantel 2002; Weeds, Weir 2005). They all share the starting point, which is the construction of a coincidence matrix of co-occurrences of words (rows) and their features describing contexts of their use (columns) in a large corpus.

The main differences between them are the following:

(1) how contexts are defined,

(2) how raw frequencies are normalized, and

(3) how the final MSR value is calculated.

In our previous work have experimented with several different settings for MSRs reported in literature in our previous work (see Piasecki, Broda 2007; Broda, Piasecki 2008), and are using the best-performing settings in this work: Point-wise Mutual Information (PMI) as the association measure and cosine as the similarity measure (see Section 3.3). Also, since Slovene is a morphologically rich language and the language tools available for Slovene are limited, we apply our Distributional Semantics methods to texts that have been previously converted to lemmas.

SPRIČEVALO (CERTIFICATE)	MRR	RELATION
potrdilo (certificate)	0.271532	hypernym
dokazilo (certificate)	0.231892	hypernym
diploma (diploma)	0.221888	hyponym
izpit (exam)	0.209232	related
listina (document)	0.207115	hypernym

Table 1: An example of words most associated to the noun *spričevalo* (*certificate*) by MSR.

However, the obtained list of highly semantically related words for a given word *w* is not enough to identify its appropriate place or places in the wordnet structure. As the examples in Table 1 show, *w* can be polysemous and pertain to several locations in wordnet (e.g. *spričevalo – school certificate* vs. *spričevalo – legal document*). Next, *w* can be associated by MSR with its synonyms and direct hypernyms (e.g. *spričevalo* > *potrdilo/dokazilo – certificate*), but also with indirect hyper/hyponyms, co-hyponyms, meronyms (e.g. *spričevalo* > *diploma - diploma*, *spričevalo* > *izpit – exam*) and words that are semantically related but are not linked by any wordnet relation.

The next step, then, is to attach the generated lists of semantically related words to the most appropriate positions in the existing semantic network. The best-known taxonomy induction methods utilize only the existing hypernymy structure in incremental wordnet expansion steps. Several machine-learning methods have been used to induce taxonomies from hypernym-hyponym pairs, such as decision trees (Witschel 2005) or k-nearest neighbors (Widdows 2003) for a limited set of domains of concrete and frequent nouns. In their seminal paper, Snow et al. (2006) propose a probabilistic wordnet-expansion method based on a probabilistic model of the taxonomy which reports promising results that, however, were not reproduced successfully in a reimplementation of their algorithm (see Piasecki et al. 2012a).

The approach used in this paper goes beyond the related work in three respects. First, in our previous work (Piasecki et al. 2012a), the wordnet hypernymy structure is perceived as a very important wordnet relation, but not the only one that describes lexical units and synsets. Thus, we aim at utilizing all different types of links in the expansion of Slovene wordnet as well. Second, the algorithm is based on the assumption that the relation extraction method produces some noise in the results, so we cannot identify the exact place (synset) for a new lemma as such but an area (a wordnet subgraph). And last, contrary to a rich body of the related work, we do not assume any shape of the lexical semantic network, but we try to build it in a way that faithfully reflects the language data. The contribution of this paper is thus the application of the algorithm to another language which has not been attempted before. Based on automatic and manual evaluation of the results we will then propose future refinements of the approach, especially tailored to Slovene language properties and the tools and resources available.

3 RESOURCES AND TOOLS USED

3.1 Gigafida

The Gigafida corpus is a 1.15 billion word reference corpus of Slovene and is as such currently the largest and most extensive text collection of Slovene (Arhar Holdt et al. 2012). It was developed within the national project Communication in Slovene (2007-2013) and contains texts of various types and genres such as literary texts, newspaper articles and Internet contents that were published between 1995 and 2011. The corpus was split into paragraphs and sentences, tokenized, part-of-speech tagged and lemmatized, so that is readily available for use via a concordancer as well as for NLP applications.

3.2 sloWNet

sloWNet is a concept-based semantic lexicon in which nouns, verbs, adjectives and adverbs are grouped into sets of synonyms (synsets) that are then organized into a hierarchical network with lexical and semantic relations, such as hyper- and hyponymy, antonymy, meronymy etc. The synsets represent concepts which are defined with a short gloss and usage examples while most synsets also have a domain label and a mapping to the SUMO/MILO ontology (Pease 2011), the largest existing formal public ontology.

sloWNet is based on a Princeton WordNet that was originally developed for the English language (Fellbaum 1998). Slovene equivalents for synsets were obtained automatically by leveraging existing bi- and multilingual resources, such as a bilingual dictionary, a multilingual parallel corpus and Wikipedia (see Fišer, Sagot 2008). Recently, a large-scale extension of sloWNet has been achieved by training a maximum entropy classifier in order to determine appropriate senses of translation candidates extracted from heterogeneous bilingual resources (see Sagot, Fišer 2012a). In addition, automatic detection of candidate outliers has been performed within the framework of distributional semantics by comparing the immediate neighborhood of literals in sloWNet and their contexts in a reference corpus (see Sagot, Fišer 2012b) with the goal of eliminating noise from the automatically generated resource.

The most recent version of sloWNet has 82,721 literals, which are organized into 42,919 synsets. Apart from single words, sloWNet also contains many multiword expressions and proper names. Nouns are still by far the most frequent, representing more than 70% of all synsets. While 66% of all the literals in sloWNet are monosemous, their average polysemy level is 2.07.

The methodology of sloWNet construction has three important implications that we try to address in this work:

(1) The resource is based on a semantic network originally produced for a foreign language, so it might be biased towards the organization and distinction of senses typical of English and therefore inadequately reflects the semantic inventory of Slovene.

(2) Slovene equivalents for synsets were harvested from several already available language resources of limited coverage, which is why we were able to obtain equivalents only for some synsets while the rest are still empty, leaving gaps in the network.

(3) Due to the automatic generation of synsets, word-sense disambiguation was not perfect, resulting in noisy synsets that have a negative impact on applications using sloWNet, and should therefore be eliminated.

3.3 SuperMatrix

SuperMatrix is a system for semantic text analysis, especially aimed at supporting automatic acquisition of lexical semantic relations from large corpora (Broda, Piasecki 2008). The main functionality of the *SuperMatrix* is related to the automated construction of corpus-based Measures of Semantic Relatedness (MSRs) and wordnet-based testing of the constructed MSRs. An MSR is a function that takes a pair of words and returns a value, which describes how closely semantically related the two words are. MSR construction follows a typical blueprint:

1. corpus preprocessing,
2. co-occurrence matrix construction,
3. matrix filtering and transformation, and
4. row similarity computation.

The depth of corpus preprocessing depends on the available language tools. However, for morphologically rich languages, lemmatization is a minimal requirement for obtaining a useful MSR, in order to avoid describing different word forms of the same lemma as semantically distinct from each other. Lemmatization can introduce some errors to MSR quality when a word form is mapped to a wrong lemma and, in the case of homographs, when a lemma with a different meaning is selected. Such errors are especially harmful when they are systematic and change the statistical blueprint of the context vector. However, practice showed that the percentage of MSR errors caused by lemmatization is very small among one-word lemmas (Piasecki et al. 2009). Multiword lexemes are a much bigger problem. They are much less frequent and require a more advanced method of identification than the recognition of a sequence of lemmas, (see Kurc et al. 2012).

In addition, a corpus parsed by a shallow parser or a dependency parser is a good basis for the construction of a highly accurate MSR, i.e., an MSR which assigns higher values for pairs of lemmas linked by one of the lexico-semantic relations, e.g. synonymy, hyper-/hyponymy, holo-/meronymy and other relations described in wordnets. Syntactic relations define linked word pairs and disambiguate context words to some extent. However, the development of a parser for Slovene has only started at the time of the experiment, which is why we decided to use only a part-of-speech tagger that also provides lemmatization.

Corpus data contain a lot of phenomena that have a negative impact on the results, e.g. very low frequencies, accidental co-occurrences due to errors produced by language tools (e.g. incorrectly assigned lemmas), which is why they must be filtered out before they can be used for similarity calculations. Moreover, many frequent words, such as *new*, *good*, *high*, *man*, *be*, occur in many contexts, are recorded as features for many words, and, as a result, can increase MSR values for weakly related or unrelated words. This is especially true with infrequent words that are described by a limited number of features. As a consequence, raw frequencies produce skewed results, which is why several weighting algorithms have been implemented in *SuperMatrix*. Our previous experiments show that the Point-wise Mutual Information (PMI) measure (Lin and Pantel 2002) gives the best results. *SuperMatrix* can also reduce dimensions of a matrix using, for example, Singular Value Decomposition. Finally, a vector similarity measure is applied to the matrix in order to obtain a ranked list of similar lemmas. *SuperMatrix* offers most well-known similarity measures but it has been shown that the simple cosine measure produces the best results in most cases.

The system also supports an automated evaluation of the selected MSR using synonymy tests that are automatically generated from wordnet, called *Wordnet-Based Synonymy Test* (WBST). The test is described in detail in (Piasecki et al. 2009) but the procedure is quite straightforward. Each test item consists of a question word that has been selected from the wordnet data, its synonym (the correct answer) taken from the same synset (or its direct hypernym in the case of singleton synsets including only the question word) and *k* distractors (words taken form other synsets). The task is to select the most related word to the question word among the presented candidates using only the MSR value. For example, for the word *svet (council)* the algorithm has to choose between *gomolj* (tuber), *izvirnost* (originality), *odbor* (committee) – the correct answer – and *odobravanje* (approval).

3.4 Wordnet Weaver

WordnetWeaver is a tool that extends the wordnet editing system called *WordnetLoom* (Piasecki et al. 2012b) with an automated wordnet expansion facility. It utilizes the results of the *Activation Area Attachment Algorithm* (AAAA) that generates suggested attachment places for new lemmas, i.e. lemmas that are not yet present in a wordnet. A suggested *attachment* is a synset to which a new lexical unit for the given new lemma can be added as a synonym – the ideal case, a hyponym – a typical case for expanding the existing wordnet structure, or linked via a lexical or semantic relation, such as hypernymy, meronymy or indirect hyponymy. The algorithm takes into account all the lexical and semantic relations found in Princeton WordNet: synonymy, hypernymy, hyponymy, holonymy, meronymy, co-hyponymy, co-meronymy and antonymy. Moreover, as all automated methods for the extraction of the lexico-semantic relations produce some errors, attachment points in *WordnetWeaver* are presented in the context of *attachment areas*, i.e. connected subgraphs of the wordnet hypernymy graph such that each synset of the selected subgraph expresses a strong enough semantic relation – in terms of the semantic fit calculated with the help of AAAA to the new lemma. A suggested attachment is always a synset with the highest value of the semantic fit in the given attachment area. Attachment areas for a new lemma are a subset of all *activation areas* identified on the basis of the semantic fit.

WordnetWeaver then presents attachment areas (i.e. top-scored suggestions) in a visual, graph-based editor and enables their verification, correction as well as manual editing of the wordnet structure. Contrary to other automated wordnet construction methods mentioned in Section 2, the aim of AAAA is to generate suggestions for lexicographers who then make the final wordnet expansion decisions, not to expand the wordnet fully automatically. Thus, AAAA is intentionally set up for slight sense over-generation in order to increase the coverage. The refinement of AAAA that would allow fully automated wordnet expansion is still an open research question.

As there is no perfect way of extracting relations, AAAA tries to utilize and combine all the extraction methods available. We assumed that its result can be represented for all types of methods as a set of triples: <*l1, l2, w*>, where *l1* s a new lemma and *l2* a lemma already in wordnet while *w* is the weight assigned to the pair of words joined in a given semantic relation.

L1	L2	W	RELATION
desnica (right wing)	levica (left wing)	0.456875	antonym
desnica (right wing)	opozicija (opposition	0.334268	related
desnica (right wing)	koalicija (coalition)	0.301908	related
desnica (right wing)	politik (politician)	0.297513	related
desnica (right wing)	stranka (party)	0.293900	hypernym

Table 2: Examples of triplets <*l1*, *l2, w*> from a MSR-based knowledge source for the AAAA algorithm (the weight W is the similarity value).

As can be seen from the example of triplets including the word *desnica – right wing* in Table 2, that are used as an input to AAAA, they are all reasonable, and describe the new word (here *desnica*) by its hypernyms, antonyms and other related words. However, in terms of recall, the system failed to pinpoint the other sense of this polysemous noun, namely the *desnica – right hand*. Most likely, this is due to skewed corpus evidence, which is skewed towards the political sense of the word.

The *l1* and *l2* are linked with a lexico-semantic relation according to a corpus-based relation extraction method, and *w* is the weight assigned to the pair by the given method. We refer to such a set of triples as a knowledge source (KS).

Each KS is produced by a different extraction method and can have different coverage (in terms of new lemmas described), interpretation of weight values and accuracy. Snow et al. (2006) used two KSs and had a probabilistic interpretation in both weights. However, this is not true in case, e.g. pattern-based methods work on single occurrences of word pairs and no reliable probability values can be calculated for them. MSR values are also not probabilities. Thus AAAA takes only a minimal assumption that weights are values expressing *semantic fit* of two lemmas. Moreover, the vast majority of pairs is extracted by patterns on the basis of singular or at most a few occurrences. Weights based on probability cannot be calculated for such pairs due to the lack of statistical evidence. AAAA therefore also introduces *global* weights for the whole KSs that can be used in parallel or instead of *local weights* included in triples. Global weights can be estimated on the basis of the accuracy of a KS obtained from manual inspection of a sample.

Taking triples from the desired KSs, the AAAA algorithm is composed of three steps:

1. The *semantic fit* between a new lemma on the input to AAAA and each lemma in the wordnet is calculated by collecting triples and weights (local and/or global) from all KSs. The semantic fit for a lemma pair can be calculated in many different weights but a simple sum of weights mostly gives good results.
2. The *semantic fit* between the input lemma *l1* and each synset *X* in a wordnet is calculated on the basis of the semantic fit between *l1* and the existing synset members, as well as the neighborhood of *X*.
3. And then, connected subgraphs *(activation areas)* of the lexico-semantic network are identified, (for details see Piasecki et al. 2012a, Broda et al. 2011).

Step 2 originates from the observation that errors in KSs cause the support for linking a pair of lemmas to be often directed to wrong places, e.g. *a bus* can be linked to *a vehicle*, while in fact this is too general and a correct link is *a car*. However, we assume that, in the case of a good KS, those suggested wrong places are accessible from the correct place via short paths in the graph of the wordnet relations, e.g. *a vehicle* is in a distance of 4 hypernymy links in WordNet 3.0 from *a car*. Thus, for a given KS triple <*l1, l2, w*>, due to its possible error, we have to consider a whole area of the wordnet relation graph around *l2* as potential places for *l1*. Moreover, *l2* can be ambiguous and can correspond to several synsets. So, each of these synsets have to be treated as defining a potential area for a sense of *l1*.

Having in mind the above observations, in step 2, during the calculation of the semantic fit to a given synset *X*, we try to compensate for the errors of KSs by taking into account not only the semantic fit of lemmas included in *X* but we also consider a part of the semantic fit of *l1* to synsets linked by short paths to *X*. We assume that in the case of good KSs it is more likely that a top level hypernym of *l1* is mismatched with a more specific hypernym than with a different lemma that is very weakly related to it. Thus a contextual, indirect semantic fit is collected only from the synsets linked by relatively short paths. The amount of the indirect semantic fit replicated depends on the length of the path and relation types of the links in the path, e.g. many relation extraction methods barely distinguish among close hyper/hyponyms but are better at differentiating synonyms and antonyms. The amount of the semantic fit replicated corresponds to the likelihood of errors of the given type and is described by the functions of transmittance and impedance that are parameters of AAAA and can be tuned on the basis of training data. AAAA has been described in detail, e.g., in (Piasecki et al. 2009) and (Piasecki et al. 2012a).

In step 3 we first identify the activation areas, and then select the attachment areas. We assume that due to the nature of KS errors, a high semantic fit is distributed around the appropriate places and that the most appropriate places – suggested attachments – are characterised by the highest values of the fit as being a kind of centres.

AAAA has so far been successfully applied to the development of the Polish wordnet (plWordNet) extensively (Piasecki et al. 2009). Also, an automated evaluation of the AAAA performance on Princeton WordNet (Fellbaum 1998) has been performed (Broda et al. 2011). The latest version of the algorithm – *Lexical Activation Area Attachment Algorithm* (LAAA) – is presented in (Piasecki et al. 2012a). In LAAA, the wordnet graph is searched for the indirect support on the level of synset members without the mediation of synsets.

4 EXPERIMENTAL SETUP

The application of the AAAA algorithm to a new language is limited only by the available language resources and corpus processing tools. The minimum requirements are: a large enough corpus and a means for constructing an MSR from it. For morphologically rich languages, Part-of-Speech tagging and lemmatization is also very useful.

In this initial experiment on Slovene wordnet expansion with *WordnetWeaver,* we have limited our work to the most frequent single-word nouns, i.e. nouns that occurred at least 1,000 times in the Gigafida corpus. There were 36,026 such nouns, 8,981 of which are already in sloWNet. This was a pragmatic decision in order to examine the first results as quickly as possible and make any necessary changes for future large-scale experiments. But the selected setting is not a limiting factor of the algorithm as such as most of the methods developed for Polish were aimed at low-frequency data (see Piasecki et al. 2009). On the other hand, the results for very frequent words should be better due to the statistical nature of applied methods.

The corpus had been PoS-tagged and lemmatized by a statistical PoS-tagger and lemmatizer called Obeliks (Grčar et al. 2012). It was then converted to a simple plain-text format. In addition, sloWNet had to be converted to the plWordNet XML format for use in *WordnetWeaver*. Apart from that, no other changes were required, which is a great advantage of the tools that were initially developed for Polish because this means that they can be used with other resources and for other languages with relatively little effort.

4.1 Extracting semantically related words

The measure of semantic relatedness is the most fundamental knowledge source for AAAA as it has good coverage (i.e. it provides similarity values for every pair of lemmas that are frequent enough in the corpus), and facilitates the discovery of lexico-semantic relations between words. In comparison to a KS that contains pairs of semantically related lemmas extracted with manually constructed patterns, which has a much higher precision than MSR, the coverage of the pattern-based KS is much lower as only a limited number of pairs can be found in the corpus.

As work on dependency parsers for Slovene is still on-going and we wanted to avoid additional manual work required for pattern-based approaches in this preliminary work, the MSR was constructed with a simple window-based approach. That is, target lemmas are described by all the other content lemmas (nouns, adjectives, verbs, adverbs) co-occurring in a small text window (3 lemmas before and after the target lemma), stopping at paragraph boundaries. The small size of the window was motivated by our previous experiments for Polish (Piasecki, Broda 2007) during which we observed that an MSR based on smaller windows provides a closer estimate of MSRs based on partial dependency parsing.

Since there is no *a piori* best method for MSR development and several are implemented in SuperMatrix, we selected the best-performing one with WBSTs based on the existing part of sloWNet. We generated questions with three detractors and a correct answer. On the 20,308 generated questions we achieved the best results for PMI weighting extended with the discounting factor and cosine similarity function (Lin, Pantel 2002). MSR chose the correct answer in 72.37% of all the questions in WBST.

4.2 Attaching the words to sloWNet

The most straightforward adaptation of AAAA to sloWNet requires importing sloWNet to the *WordnetWeaver* scheme and a preparation of knowledge sources. We have prepared two KSs based on MSR. The first one is based on the similarity lists for lemmas. That is, for each lemma l_x we compute 20 most similar lemmas l_y using the above-described MSR. This KS then takes the form of pairs <l_x, l_y, *msr(x,y)*>, where *msr(x,y)* is a value of MSR between the two lemmas.

Table 3 contains 20 highest-ranking triplets for the word *termin*. As the generated most similar lemmas show, the word is polysemous and can refer to *technical term* (65%), *deadline* (15%) or *time period* (10% suggestions). The meaning of the word cannot be guessed from the suggestion of similar lemmas in 2 (10%) cases, which are at the same time the only two triplets that are completely useless in terms of attaching the headword into wordnet. The rest are vaguely (5%) or closely related (65%), are the headword's hypernyms (15%) or refer to the domain the headword belongs to (5%). By far the most useful of these triplets for the lexicographer are those of the hypernym kind whereas the closely related ones can serve as reminders where to attach the word in the network.

The other KS uses *bi-directional* similarity lists. It is a subset of the above knowledge source with additional filtering. For l_x the pair <l_x, l_y, *msr(x,y)*> is included only if there is also a pair <l_y, l_x, *msr(y,x)*> among the 20 most similar items for l_y.

The second KS is clearly correlated with the first one. However, the bi-directional similarity list express a significantly higher precision in representing wordnet relations, and by combining these two KS we emphasise strongly lemma pairs that provide more reliable information.

L1	L2	W	RELATION
termin (term)	izraz (expression)	0.173217	hypernym
termin (term)	pojem (concept)	0.173076	hypernym
termin (deadline)	urnik (schedule)	0.171332	closely related
termin (term)	kontekst (context)	0.153846	closely related
termin (term)	definicija (definition)	0.152953	closely related
termin (term)	teorija (theory)	0.150855	closely related
termin (time)	spored (listing)	0.150815	closely related
termin (term)	concept (concept)	0.149574	hypernym
termin (deadline)	datum (date)	0.147092	closely related
termin (?)	vsebina (content)	0.143041	unrelated
termin (term)	terminologija (terminology)	0.142419	domain
termin (term)	pomen (meaning)	0.141739	closely related
termin (term)	smisel (sense)	0.136470	closely related
termin (?)	praksa (practice)	0.132857	unrelated
termin (term)	interpretacija (interpretation)	0.132739	closely related
termin (time)	razpored (plan)	0.132211	closely related
termin (term)	tema (topic)	0.130169	closely related
termin (deadline)	teden (week)	0.129162	closely related
termin (term)	vidik (aspect)	0.127669	vaguely related
termin (term)	razumevanje (understanding)	0.127527	closely related

Table 3: Triplets <*l1*, *l2*, *w*> produced by an MSR.

The suggested attachment areas for the word *desnica* (right wing) are spread around the following sloWNet synsets presented below in the order of their semantic fit (the best first):

1. {*duša, glava, nekdo, oseba, posameznik, smrtnik, človek*} (person)
2. **{*ideologija, nazor, politična teorija, politični nazor*} (political view)**
3. {*besednjak, slovar*} (vocabulary)
4. {*aliansa, koalicija, liga, pakt, zavezništvo, zveza*} (coalition)
5. {*oblast, pravilo, predpis, vladanje*} (government)

In the political sense, the word *desnica* should best be attached as a hyponym of synset *ideologija, nazor, politična teorija, politični nazor* (political view). Closely related, but not in the hyper-hyponymy chain, is also the synset *oblast, pravilo, predpis, vladanje* (government) as well as the synset *aliansa, koalicija, liga, pakt, zavezništvo, zveza* (coalition). The top attachment suggestion *duša, glava, nekdo, oseba, posameznik, smrtnik, človek* (person) probably originates from the word's body part sense and is not suitable for the political sense of the word.

4.3 Evaluation of the results

WordnetWeaver and AAAA were designed to help a linguist in expanding an existing wordnet structure with new lemmas. Thus, the evaluation of the algorithm's performance should focus on this practical aspect. In order to gain a comprehensive insight into the performance of the adopted approach, we evaluate the results both automatically and manually.

4.3.1 AUTOMATIC EVALUATION

For automatic evaluation of the results, we follow the evaluation methodology proposed by (Broda et al. 2011). The idea of the evaluation is simple: first, we remove some literals from the existing sloWNet structure; then we run AAAA for those literals and see how close to the original place in sloWNet (in terms of the length of hyper-/hyponymy paths) the removed literals were re-attached by the AAAA. It means that the algorithm works perfectly if a literal is reattached directly to the same synset. AAAA suggestion for re-attaching the given lemma as, e.g., an indirect hyponym of the original synset is considered to be less correct. Ideally, we would like to remove all occurrences of one lemma in sloWNet at a time, next reattach it and the process repeats for each test lemma. Reattaching single lemmas would alter sloWNet structure as little as possible. However this way of performing evaluation is computationally very expensive as several complex operations are repeated for each test in the testing environment. Thus, we remove a package of 50 lemmas at a time. For evaluation purposes, we randomly selected a sample of the 1,000 nouns meeting the frequency threshold that was also set to 1,000 (see Section 3).

Several evaluation strategies are possible, each giving a different perspective on the algorithm performance (Broda et al. 2011). From the lexicographers' point of view, the algorithm performs well if there is at least one correct suggestion that is relatively close to the proper place in a wordnet structure, i.e., the correct place and the suggestion can be seen on the same screen of WordnetLoom in a distance up to 6 links.

Applying the *closest path* evaluation strategy for each test lemma we check only one suggestion, the one that is closest to the correct place of the original position of the lemma in the wordnet. This strategy is intended to measure how useful suggestions are for a linguist assuming that having at least one suggestion in close distance is helpful.

In the *best supported* strategy only one the top-scored suggestion (with the highest semantic fit) provided for a given test lemma by the algorithm is analyzed. The *best* strategy shows how much we can trust the highest-scored suggestions.

Finally, in the last evaluation strategy we simply check *all* suggestions generated by the algorithm per a test lemma.

Table 4 presents the results of the described evaluation methodology for all three strategies. The *acceptable distance* to the original place was set to 6 on the basis of the experience of lexicographers with using visual graph-based wordnet editing in WordnetLoom (Piasecki et al. 2009). The distance is measured on the hypo-/hypernymy and mero-/holonymy graphs with the exception that we can only traverse one edge of mero-/holonymy at the end of the path (as these relation can take us to completely unrelated parts of the wordnet very quickly).

Dist.	Closest [%]	Best [%]	All [%]
0	15.0	5.9	3.7
1	19.7	13.9	4.6
2	19.0	13.9	6.0
3	11.7	8.2	4.9
4	8.1	9.0	5.3
5	5.5	6.4	6.8
6	0.2	0.7	0.8
Σ	79.2	57.9	32.2

Table 4: Results of the automatic evaluation procedure for sloWNet expansion.

The achieved results are significantly lower than for Polish (Broda et al. 2011), i.e. 91.1% according to the *closest* strategy, 78.8% for the *best* strategy (called there strongest) and 75.7% for the *all* strategy, in the case of frequent Polish lemmas with the frequency ≥ 1000. The coverage was 99% for words (at least one suggestion generated) and 66% for the known senses of the test lemmas. We expected such differences, as we have employed much simpler and less precise, window-based MSR in the case of Slovene data, and we did not used additional, pattern-based KSs. On the other hand, the results are encouraging as for almost 80% of the words the algorithm suggested at least one correct place for attachment. Also, the correct attachment places are mostly close to the original place in the wordnet structure (i.e., the results are shifted towards closer distances than 6). AAAA provided a suggestion for 94% of words from the random sample and found 29.6% of word senses for each word on average.

4.3.2 MANUAL EVALUATION

For a more qualitative insight into the results, we also performed a manual evaluation on 100 random lemmas included in the automatic evaluation. In manual evaluation, 5 highest-ranking attachment suggestions were checked for each lemma, amounting to 500 candidate-attachment pairs.

The evaluated lemmas were first categorized into monosemous or polysemous. Based on the attachment suggestions for polysemous lemmas, we checked whether our algorithm was able to detect only one of its senses or more. We took into account only the 5 top-ranked suggestions for each lemma because checking longer lists would be too time consuming in a realistic lexicographic scenario. Next, we tried to label each attachment suggestion with one of the 10 lexico-semantic relations included in wordnets and produced by our algorithm: *synonymy, hypernymy, hyponymy, holonymy, meronymy, co-hyponymy, co-meronymy, antonymy, close, vague*, or *no relation*. The *no relation* label is intended for clear errors of the algorithm. The *close* label is used for cases where the candidate-attachment pair is clearly semantically related but the relation type is not found in the current version of sloWNet (e.g. *Occupation-Place* such as *pošta-poštar* [post-postman], *Activity-Occupation* such as *učenje-učitelj* [teaching-teacher]). The *vague* label, on the other hand, is used for cases where the candidate-attachment pair is in a more loose associative relation that will probably not be encoded in wordnet (e.g. same semantic field such as *politika-debata* [politics-debate]).

Overall, the results of manual evaluation are very encouraging as no cases were found where all the attachment suggestions for a lemma would be completely unrelated. What is more, only 1 out of 100 lemma received no better attachment suggestion than a vague association, and an additional 1 got at best a closely related one. On the other hand, as many as 38 lemmas had no erroneous attachment suggestions, which means that the lexicographers who are responsible for selecting the best attachment candidates will be presented with very little noise that would slow down their work.

Category	Freq.	%
synonym	22	4.40%
hypernym	74	14.80%
hyponym	9	1.80%
holonym	9	1.80%
meronym	12	2.40%
antonym	1	0.20%
co-hyponym	40	8.00%
co-meronym	2	0.40%
closely related	171	34.20%
vaguely related	50	10.00%
unrelated	110	22.00%
total	500	100.00%

Table 5: Frequency counts of association candidates per relation type.

As Table 5 shows, almost 34% of the suggested association candidates can easily be labeled with one of the standard lexico-semantic relation types from wordnet. By far the most frequent one is the hypernymy relation that was selected in almost 15% of the cases. There were quite a lot of co-hyponymy (8%) and synonymy (4%) attachments as well while the rest of the relations were much more rare. A further 34% of the suggestions were very closely related to the lemmas, 10% were loosely associated to them while 22% of the association candidates were not related at all to the lemmas they were assigned to.

When analyzing the semantic nature of the randomly selected lemmas in the evaluation sample, we observe that 62% of them are monosemous and 38% polysemous. This is very similar to the polysemy level of nouns in the latest version of sloWNet, where 66% of the literals are monosemous. A single sense prevailed for 58% of the otherwise polysemous lemmas in the evaluation sample, while association candidates refer to different senses in 42% of the cases. This is a well-known phenomenon of distributional semantics where a Zipfian distribution of senses in the corpus causes skewed context vectors of polysemous words, which are thus heavily biased towards the most frequent sense in the corpus.

Cat.	**Mono.**	**Poly.**			**Σ**
		1 sense detected	>1 sense detected	Σ poly	
synonym	62	22	16	38	100
hypernym	11	5	3	8	19
hyponym	40	13	7	20	60
holonym	3	1	2	3	6
meronym	4	2	3	5	9
antonym	4	5	2	7	11
co-hypo	1	0	0	0	1
co-mero	19	6	3	9	28
closely related	1	1	0	1	2
vaguely related	51	16	2	18	69
error	22	4	5	9	31

Table 6: Frequency counts of lemmas with at least 1 association suggestion per category. The first column contains semantic categories (Cat.), and the rest are their frequency counts for monosemous (Mon.) and polysemous (Poly.) words as well as the sum total (Σ).

Table 6 shows frequency counts of semantic categories that appeared at least once among the association suggestions per lemma. Because we counted all the relation types that were suggested for each lemma, and a single lemma could have suggestions belonging to a single category or up to five different categories, the total count is more than 100. Hypernymy and co-hyponymy are still the most frequent in this setting, suggested for 60% and 28% of the lemmas, respectively. Both are more frequently suggested for monosemous nouns, while polysemous ones have more suggestions for synonyms, hyponyms, holonyms, meronyms and co-meronyms. Polysemous nouns contain a slightly higher number of erroneous attachment candidates and a much higher number of vaguely and closely related suggestions than the monosemous ones. Interestingly, the polysemous nouns for which only one sense was detected by the algorithm, contain the least noise and vague association candidates.

5 CONCLUSIONS

In this paper we presented the first results of applying *WordnetWeaver* to Slovene data in order to extend Slovene wordnet. The approach, which had never been ported to a new language before, uses statistical methods to extract lists of semantically similar words from a large reference corpus of Slovene, and then identifies the part of the wordnet hierarchy these words should be attached to. Automatic and manual evaluations of the results show that the algorithm was successfully ported to a new language and is already useful in its most basic setting. However, the state-of-the-art results for Polish suggest that further improvements of measures of semantic relatedness are still possible, for example by using a constraint-based approach, a dependency parser, and testing more measures with more parameters. Similarly, the attachment algorithm could further be improved by optimizing parameters of the algorithms, for example by using meta-heuristics like in (Kłyk et al. 2012), and providing additional knowledge sources, such as pattern-based lists of semantically related word pairs.

In the future, we wish to investigate methods that would enable us to extend the current functionality of the attachment algorithm to expand sloWNet fully automatically, requiring no human intervention for reaching the final decision where to add a new word in wordnet. A somewhat different but very interesting area of research would be to adapt the attachment algorithm to be able to use corpus data in order to analyze the semantic network in sloWNet that is based on Princeton WordNet and find suspicious areas in the network that does not correspond to the linguistic evidence harvested from the corpus and should therefore be improved.

ACKNOWLEDGMENTS

We would like to thank Tomaž Erjavec for performing conversions of sloWNet into the plWordNet XML format and of the Gigafida corpus into plain text. We would also like to thank the reviewers for their insightful comments on our paper.

REFERENCES

Arhar Holdt, Š., Kosem, I., and Logar Berginc, N. (2012): Izdelava korpusa Gigafida in njegovega spletnega vmesnika. *Proceedings of 8th Language Technologies Conference IS-LTC-12*: 16–21. Ljubljana.

Bentivogli, L., and Pianta, E. (2004): Looking for Gaps. *Proceedings of EURALEX-2000*: 663–669. Stuttgart.

Broda, B., and Piasecki, M. (2008): SuperMatrix: a General Tool for Lexical Semantic Knowledge Acquisition. *Speech and Language Technology Conference, Volume 11 of Lecture Notes in Computer Science*: 239–254. Berlin.

Broda, B., Kurc, R., Piasecki, M., and Ramocki, R. (2011): Evaluation Method for Automated Wordnet Expansion. *Security and Intelligent*

Information Systems: 293–306. Berlin.

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

Fišer, D. (2005): Pristopi k izdelavi leksikalnih podatkovnih zbirk. *Jezik in slovstvo*, 50 (6): 17–32.

Grčar, M., Krek, S., and Dobrovoljc, K. (2012): Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. *Proceedings of 8th Language Technologies Conference IS-LTC-12*: 89–94. Ljubljana.

Harris, Z. S. (1968): *Mathematical Structures of Language*. New York: Interscience Publishers.

Hearst, M. A. (1992): Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceeedings of COLING-92*: 539–545. Nantes.

Kłyk, Ł., Myszkowski, P. B., Broda, B., Piasecki, M., and Urbansky, D. (2012): Metaheuristics for Tuning Model Parameters in Two Natural Language Processing Applications. *Proceedings of AIMSA-12*: 32–37. Varna.

Kurc, R., Piasecki, M., and Broda, B. (2012): Constraint Based Description of Polish Multiword Expressions. *Proceedings LREC-12*: 2408–2413. Istanbul.

Lin, D., and Pantel, P. (2002): Concept Discovery from Text. *Proceedings of COLING-02*: 577–583. Taipei.

Pantel, P., and Pennacchiotti, M. (2006): Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. *Proceedings of COLING-ACL-06*: 113–120. Sydney.

Pease, A., (2011): *Ontology: A Practical Guide*. Angwin, CA: Articulate Software Press.

Piasecki, M., Kurc, R., Ramocki, R., and Broda, B (2012a): Lexical Activation Area Attachment Algorithm for Wordnet Expansion. *Proceedings of AIMSA-12*: 23–31. Varna.

Piasecki, M., and Broda, B. (2007): Semantic Similarity Measure of Polish Nouns Based on Linguistic Features. *Business Information Systems: 10th International Conference, Volume 4439 of Lecture Notes in Computer Science*: 381–390. Berlin.

Piasecki, M., Marcińczuk, M., Ramocki, R., and Maziarz, M. (2012): WordnetLoom: a Wordnet Development System Integrating Form-based and Graph-based Perspectives. *International Journal on Data Mining, Modelling and Management*, 5 (3): 210–232.

Piasecki, M., Szpakowicz, S., and Broda, B. (2009): *A Wordnet from the Ground Up*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Ruge, G. (1992): Experiments on Linguistically-based Term Associations. *Information Processing and Management*, 28 (3): 317–332.

Snow, R., Jurafsky, D. and Ng, A. Y. (2006): Semantic Taxonomy Induction from Heterogenous Evidence. *Proceedings of ACL-06*: 801–808. Sydney.

Vider, K. (2004): Concerning the Difference between a Conception and its Application in the Case of the Estonian WordNet. *Proceedings GWC-04*: 285–290. Brno.

Weeds, J., and Weir, D. (2005): Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity. *Computational Linguistics*, 31 (4): 439–475.

Widdows, D. (2003): Unsupervised Methods for Developing Taxonomies by Combining Syntactic and Statistical Information. *Proceedings of NACL-03*: 197–204. Edmonton.

Witschel, H. F. (2005): Using Decision Trees and Text Mining Techniques for Extending Taxonomies. *Proceedings of Learning and Extending Lexical Ontologies by using Machine Learning Methods, Workshop at ICML-05*: 61–68. Leipzig.

Wong, S. H. S. (2004): Fighting Arbitrariness in WordNet-like Lexical Databases: A natural language motivated remedy. *Proceedings of GWC-04*: 234–241. Brno.

UTEMELJEVANJE sloWNeta NA KORPUSNIH PODATKIH

Wordnet lahko izdelamo na podlagi že obstoječega tujejezičnega wordneta ali pa kot osnovo za gradnjo vzamemo korpusne podatke. Prvi pristop je preprostejši in enostavnejši, zaradi česar ga razvijalci tudi najpogosteje uporabljajo. Vendar ima ta pristop veliko pomanjkljivost, predvsem to, da tako izdelan vir ne odseva nujno jezika, za katerega je bil izdelan. Zato v pričujočem prispevku predstavljamo pristop, ki izhaja iz jezikovnih podatkov, pridobljenih iz jezikoslovno označenega referenčnega korpusa, iz katerega smo s pomočjo preprostih statističnih metod izluščili sezname semantično podobnih besed, ki smo jih nato vključili v wordnet za slovenščino. Pristop je bil prvotno razvit za poljščino, vendar je privlačen tudi za druge jezike, saj zanj potrebujemo minimalna jezikovnotehnološka orodja in vire, zato ga je enostavno uporabiti tudi za jezike, za katere obsežen wordnet ali podoben semantični leksikon še ne obstaja. Druga pomembna prednost uporabljenega pristopa pa je, da temelji na izpričani jezikovni rabi, pridobljeni iz korpusa, ki se nato kaže v jezikovno utemeljeni organizaciji besedišča v izdelani semantični mreži. Glede na to, da so vsi naši dosedanji pristopi za izdelovo slovenskega wordneta celotno strukturo prevzeli iz Princetonovega WordNeta, ki je bil izdelan za angleščino, bodo spodbudni rezultati, dobljeni s pričujočo metodo, koristno dopolnjevali obstoječo semantično mrežo.

Ključne besede: leksikalna semantika, wordnet, semantična podobnost, semantične relacije