Corpus-based Semantic Relatedness for the Construction of Polish WordNet

Bartosz Broda[1], **Magdalena Derwojedowa**[2], **Maciej Piasecki**[1], **Stanisław Szpakowicz**[3,4]

[1] Institute of Applied Informatics, Wrocław University of Technology
[2] Institute of the Polish Language, Warsaw University
[3] School of Information Technology and Engineering, University of Ottawa
[4] Institute of Computer Science, Polish Academy of Sciences
bartosz.broda@pwr.wroc.pl, derwojed@uw.edu.pl, maciej.piasecki@pwr.wroc.pl, szpak@site.uottawa.ca

Abstract

The construction of a wordnet, a labour-intensive enterprise, can be significantly assisted by automatic grouping of lexical material and discovery of lexical semantic relations. The objective is to ensure high quality of automatically acquired results before they are presented for lexicographers' approval. We discuss a software tool that suggests synset members using a measure of semantic relatedness with a given verb or adjective; this extends previous work on nominal synsets in Polish WordNet. Syntactically-motivated constraints are deployed on a large morphologically annotated corpus of Polish. Evaluation has been performed via the WordNet-Based Similarity Test and additionally supported by human raters. A lexicographer also manually assessed a suitable sample of suggestions. The results compare favourably with other known methods of acquiring semantic relations.

1. Introduction

The construction of a wordnet is expensive, mainly due to the high linguistic workload. The recent developments in automatic acquisition of lexico-semantic relations suggest cost reductions. Our project to construct a wordnet for Polish (*plWordNet*) (Derwojedowa et al., 2008) explores this path as a supplement to a well-organized and well-supported effort of a team of lexicographers. We envisage the software-assisted creation of 15000 to 20000 *lexical units* (LUs). The *core* of about 7000 LUs, constructed manually, will help "bootstrap" the remainder semi-automatically.

A software tool will generate synset member suggestions based on a *measure of semantic relatedness* (MSR) with a given LU v: LUs semantically related to v should receive significantly higher values than unrelated LUs. Another tool will assist the lexicographers in identifying instances of lexico-semantic relations among the LUs deemed sufficiently related. An MSR is particularly useful if it places LUs most closely related to v near the top of the suggestion list.

The process relies on a morphosyntactically annotated corpus of Polish: the IPI PAN corpus (IPIC) (Przepiórkowski, 2004) of about 254 million tokens. We work at the morphosyntactic level, effective enough given rich Polish inflection, because no efficient, accurate parser for Polish is available in the public domain yet.

We have earlier constructed an MSR for Polish nouns (Piasecki et al., 2007a) and discussed its evaluation (Piasecki et al., 2007b). Here, we present a natural continuation of that work: MSRs for verbs and adjectives, the remaining open categories in *plWordNet*.

Work financed by the Polish Ministry of Education and Science, Project No. 3 T11C 018 29.

2. Rank Weight Functions in similarity extraction

Our method of MSR construction (Piasecki et al., 2007a) follows a scheme of distributional similarity extraction that generalises many existing approaches:

- a co-occurrence matrix of LUs and *contexts* is constructed,
- next, the matrix row values are transformed in order to emphasise co-occurrence regularities,
- finally, the computed similarity of row vectors is used as an estimate of the similarity of the corresponding LUs.

A rich description of a context would usually be based on parsing results, e.g., (Ruge, 1992; Lin, 1998; Weeds and Weir, 2005). A context is represented as an instance of a lexico-syntactic relation, e.g., `an object of` *`eat`*. Relation instances could be identified effectively given the output of a sufficiently accurate syntactic parser. This cannot be done for Polish, because no such parsers have been placed in the public domain. We found a satisfactory replacement, however. Context descriptions rely on *lexico-morphosyntactic constraints*, that is to say, associations among word forms which depend on their morphosyntactic characteristics, e.g., a possible agreement between an adjective and a noun on gender, number and case. In an inflectional language like Polish, the morphosyntactic description of word forms (rather than word order) delivers most of the structural information. Morphosyntactic associations are also simpler to recognise, since this requires only a tagger and a constraint representation formalism. The main disadvantage is that identification is in many cases imprecise and the overall error depends on the tagger error rate.

Most known methods of transformation and similarity computation are based directly on the frequencies of LUs, with

context co-occurrences collected from a corpus, e.g. (Freitag et al., 2005; Weeds and Weir, 2005). We have observed that noise and bias in an unbalanced corpus, as is the case with IPIC, can significantly influence the extracted MSR. In order to compensate for that, we follow a typical blueprint for MSR computation, namely:

1. *global selection* of features on the basis of global statistical evaluation and perhaps a heuristic assessment;
2. transformation of the matrix cells (or rows), either globally or only by referring to the compared rows of LUs;
3. local selection of features for the comparison of two units;
4. similarity calculation for a pair of row vectors.

At the local selection level, however, we have introduced (Piasecki et al., 2007a) a *Rank Weight Function* (RWF) as a means of abstracting from the corpus frequencies in weights assigned to *features* of the given LU; a feature is a particular instance of a lexico-morphosyntactic relation.
In the case of RWF, we assume that what contributes most information to a LU's description is not the feature's exact frequency. Instead, we take its *relevance* relative to the other features that describe the given LU. Feature relevance for a LU can be estimated on the basis of the mutual association of those features, observed in the language data, e.g., measured by the statistical significance of the co-occurrence. The ordered set of the relevant features describes the meaning of the given LU, and the meanings of two LUs can be compared by the comparison of the corresponding sequences of features ordered by their relevance. We believe that differences in the feature's initial values, directly related to the corpus frequencies, are mostly an artefact of the corpus bias. One should not depend on them too strictly during similarity calculation.
Thus, the core of a RWF is a mapping from feature frequencies to features ranks, based on their order of relevance.

1. Let $\mathbf{M}$ be a co-occurrence matrix, w_i – a LU, c_j – a feature, $\mathbf{M}[w_i, c_j])$ – the co-occurrence frequency of w_i together with c_j.
2. For the given w_i, we recalculate the weighted values of the corresponding cells, using a *weight function* f_w: $\forall_c \mathbf{M}[w_i, c] = f_w(\mathbf{M}[w_i, c])$.
3. Features in a row vector $\mathbf{M}[n_i, \bullet]$ are sorted in the ascending order on the weighted values.
4. The k highest-ranking features are selected; e.g., k = 1000 works well.
5. For each selected feature c_j a new value is calculated: $\mathbf{M}[w_i, c_j] = k - rank(c_j)$
where $rank(c_j)$ calculates the position of c_j (starting from zero) in the ranking based on f_w.

The construction of RWF has been inspired by the *neighbour set comparison technique* introduced in (Lin, 1997) and modified in (Weeds and Weir, 2005). It was applied to the *comparison* of the results of two MSRs. In our approach we use rank vectors in *calculating* MSR and ranks are the values of the features.
In step 2 one can use any function that produces values comparable to those of the weight function. The idea, however, is to apply a function that somehow measures the relevance of the feature c_j to the LU w_i. Natural candidates for the weight functions are measures based on probability distribution or on Information Theory. In (Piasecki et al., 2007b) several functions have been tested; the best result was achieved with the *t_score* measure, e.g., (Manning and Schütze, 2001), applied as the *weight function* f_w:

$$f_w(w,c) = t\_score(w,c) = \frac{\mathbf{M}[w,c] - \frac{TF_w TF_c}{W}}{\sqrt{\frac{TF_w TF_c}{W}}} \quad (1)$$

$TF_w = \sum \mathbf{M}[w, \bullet]$, $TF_c = \sum \mathbf{M}[\bullet, c]$ are the total frequencies of LUs and features, respectively, and *W* is the number of words processed. An event is an occurrence of a word in text – features are occurrences filtered by some constraint. We assumed the strict threshold of significance of LU – feature co-occurrence for 0.5% ($t\_score \geq 2.567$). In this step we filter out all such c for which $t\_score(w,c) < 2.567$.
In step 4 only a predefined number of the k best feature is selected for the description of the given LU w_i. This is how we want to eliminate less relevant, more accidental features. The exact value of k is set up experimentally. It depends of the type of f_w applied and it seems to depend on the level of sparseness of the matrix – sparser matrices seem to require lower values of k.
Finally, in step 5, the best feature receives k as its new value, the remaining features – the subsequent values in descending order. As the number of features selected is often lower than k (because of matrix sparseness and the *t_score* threshold), we also tested a different scheme, in which the best feature was assigned the value equal to the number of the features selected. The results, however, were slightly worse.

3. Lexico-morphosyntactic constraints

Lexico-morphosyntactic constraints are expressed in JOSKIPI, a specialised language implemented as part of a rule-based tagger (Piasecki, 2006). The constraints refer to the morphosyntatic properties of tokens, but prior disambiguation is assumed. For most of the constraint, especially those referring to some sort of morphosyntactic agreement between word forms, we achieve in practice the expressive power similar to that of a chunker. Constraints that test only the existence of a sequence of specified tokens, e.g., a close occurrence of an adverb before a adjectival LU, are intrinsically less precise, but many of them are still very informative, as shown below. At present, all constraints return Boolean values and associate two words: one represents the LU under consideration, the other is a semantically descriptive modifier or predicate.[1] A constraint can process

[1] The mechanism allows the use of non-Boolean features, however, and the analysis of more than one word in the context.

the whole sentence. A set of Boolean values produced by constraints describes a context.
There is ample literature on the methods of representing and acquiring lexical semantics of verbs. Most methods focus on subcategorisation, verb frames and semantic restrictions on verb arguments. We stay closer to the surface: the distribution of morphologically informed patterns dictates a real-valued MSR that helps identify pairs of closely related verbs.
The proposed MSR for verbs combines several constraints on token occurrences (we have italicised the lexical elements of the constraints):

- a particular *noun* as a potential subject of the given verb (NSb in Table 1) ,
- a *noun* in a particular case as a potential verb argument (NArg), see Fig. 1,
- a present or past participle of the given verb as a modifier of some *noun* (VPart),[2]
- an *adverb* in close proximity to the given verb (VAdv), see Fig. 2.

In the constraint `NArg` presented in Fig. 1, written in JOSKIPI (Piasecki, 2006), in the first `and` expression we check whether there is a form that is the head of an utterance, e.g., a finite form, at position 0 – mnemonics come from the IPIC tagset (Przepiórkowski, 2004). Next we look to the right (`rlook`) for a noun form in the accusative case or a next verb. `$SR` is a variable that stores the position of iteration. Finally, we check whether the right iteration stopped on a noun in accusative – in that situation we have found an argument. In the next `and` expression, we look for an appropriate noun to the left, but this time we have to check if a preceding verb form is separated by a punctuation mark or conjunction.
In the constraint `NArg` presented in Fig. 2, we test the presence of an adverb first in the left context of a verb, next in the right context. In both circumstances we look for an adverb that it is not separated from the verb form in the position 0 by any other verb form. As there is no agreement between a verb and an adverb, we can depend only on their proximity.
Constraints produce Boolean values, but the procedure is not strict: a small number of errors is acceptable, and the errors seem to be compensated by the large amount of data processed.[3]
MSRs for adjectives were constructed as a by-product of larger projects in (Hatzivassiloglou and McKeown, 1993; Freitag et al., 2005). Extraction of distributional features was also discussed in (Lapata, 2001; Boleda et al., 2004; Boleda et al., 2005), but applied in the semantic classification of adjectives. We have identified three types of constraints as the potential semantic descriptors of adjectives:

- an occurrence of a particular *noun* as modified by the given adjective (ANmod) – we look for a noun which agrees on case, gender and number,[4]
- an *adverb* in close proximity to the given adjective (AAdv),
- the co-occurrence with an *adjective* that agrees on case, number and gender as a potential co-constituent of the same noun phrase (AA).

The last feature was advocated in (Hatzivassiloglou and McKeown, 1993) as expressing negative semantic information: only unrelated adjectives can sit in the same noun phrase. Our corpus data, however, suggest that it is too strong a bias. In addition, our AA constraint also accepts coordination of adjectives, and then related adjectives can co-occur in a noun phrase. In the end, we used the AA feature in a positive way, just like the other features. Features of all three types, weighted and filtered by an RWF, were used in the discovery of contexts of occurrences of particular adjectives.
The AA constraint was applied in two different ways:

- as part of a joint large matrix together with the two other constraints: different parts (columns) of row vectors generated by different constraints, but the matrix processed as a whole – this usage is encoded ANmod+AAdv+AA in Table 1,
- two separate matrices were created: one joint for ANmod+AAdv and another for AA only.

In the second situation, the similarity values were calculated separately on the basis of both matrices separately processed and next linearly combined:

$$MSR_{Adj}(l_1, l_2) = \\ \alpha\ MSR_{ANmod+AAdv}(l_1, l_2) + \beta\ MSR_{AA}(l_1, l_2) \quad (2)$$

The values of the coefficients were selected experimentally; $\alpha = \beta = 0.5$ gave the best results.
A linear combination of separate matrices, that is to say, a linear combination of two MSRs, produced better results than the joint matrix ANmod+AAdv+AA.

4. Results and evaluation

For the needs of a general automated test of MRS accuracy, we have adapted the idea of a *WordNet-Based Similarity Test* (WBST) (Freitag et al., 2005) to an evaluation similar to what we had done with nouns (Piasecki et al., 2007b). WBST consists of pairs $\langle q, A\rangle$: *question-word* (q) – four *answer-words* (A) among which only one is a near-synonym of q. In our case, a WBST is generated on the basis of *plWordNet*. An instance of the test is built as follows: first, for a LU q (here an adjective or a verb) included in *plWordNet* its near synonym LU s is randomly selected and added to A; next, three other LUs not in q's synsets (detractors) are randomly drawn from *plWordNet* to complete

[2] A subtle agreement test and additional structural conditions distinguish such pairs from verb-complement pairs.

[3] Sometimes only systematic tagger errors influenced the results.

[4] Here we use almost the same constraint as that presented for nouns in (Piasecki et al., 2007b) but in the reverse direction.

```
or( and(
     in(pos[0],fin,ger,praet,impt,imps,inf,ppas,ppact,pcon,pant),
     rlook(1,end,$SR,or(
            in(pos[$SR],fin,ger,praet,impt,imps,inf,ppas,ppact,
                                                  pcon,pant,conj,interp)
            and( in(pos[$SR],subst,depr),
                 equal(cas[$SR],acc)
                 inter(base[$SR],"particular noun") )
            )),
     and( in(pos[$SR],subst,depr),
          equal(cas[$SR],acc) )
    ),
   and(
    in(pos[0],fin,praet,impt,imps,inf),
    llook(-1,begin,$SL,or(
           in(pos[$SL],fin,ger,praet,impt,imps,inf,ppas,ppact,
                                                 pcon,pant,conj,interp)
           and( in(pos[$SL],subst,depr),
                equal(cas[$SL],acc)
                inter(base[$SL],"particular noun") )
           )),
    and( in(pos[$SL],subst,depr),
         equal(cas[$SL],acc)
         and(
           llook($SL,begin,$VSL,in(pos[$VSL],fin,ger,praet,impt,imps,inf,
                                   ppas,ppact,pcon,pant,conj,interp)),
           in(pos[$VSL],conj,interp)
         ) )
   ),
   and(
       similar for left arguments of gerunds and participles
   )  )
```

Figure 1: Example of a constraint: a *noun* in a particular case as a potential verb argument (NArg).

```
or( and(
     in(pos[0],fin,praet,impt,imps,inf,ppas,ppact,pcon,pant)
     llook(-1,begin,$AL,or(
            in(pos[$AL],fin,ger,praet,impt,imps,inf,ppas,ppact,
                                                pcon,pant,conj,interp),
            and( equal(pos[$AL],adv),
                 inter(base[$AL],"particular adverb") )
     )),
     equal(pos[$AL],adv) ),
   and(
     a similar constraint for gerund forms and the left context
   ),
   symmetric constraints for non-gerund verb forms and the right context
)
```

Figure 2: Example of a constraint: an *adverb* in close proximity to the given verb (VAdv).

A in the pair. During evaluation, MSR generates values for the pairs $\langle q, a_i \rangle$, $a_i \in A$, expected to favour s.

Many synsets in *plWordNet*have at most three LUs, and there are many singletons. We had to modify the test slightly. In order to get a better coverage of LUs by WBST questions, and not to leave LUs in singleton synsets untested, the direct hypernyms of LUs from singleton synsets were considerded as their near-synonyms and used to form QA pairs in (Piasecki et al., 2007a). We named this modification the WBST with Hypernyms (WBST+H). The inclusion of hypernyms in QA pairs did not make the test easier, as was shown in (Piasecki et al., 2007a).

We generated 3532 question-answer pairs for adjectives and 3086 for verbs. The QA pairs encompass 2718 different adjectives (among them 569 occur over 1000 times in the corpus) and 2984 different verbs (702 occur more than 1000 times). Some of them occur in QA pairs more than once but with different near-synonyms.

Table 1 shows the results for different MSRs on the same tests. For WBST+H the baseline random selection is 25%. We divided the analysed adjectives and verbs into two groups by their frequency in IPIC: those occurring > 1000 and the others. We can thus compare our results with (Freitag et al., 2005) who presented their results only for LUs with > 1000 occurrencies.

Working with the same generated co-occurrence matrices for verbs and adjectives, we compared the application of RWF with four other measures:[5]: Lin's measure (Lin, 1998), CRMI (Weeds and Weir, 2005), RFF (Geffet and Dagan, 2004) and Freitag et al.'s measure (Freitag et al., 2005) (the authors call it "optimal"). From a large number of proposed solutions, we selected only the measures based on lexico-syntactic features. Lin's measure was included in the set due to its significant influence on the subsequent research. CRMI has been extensively compared with several other approaches showing significant improvement. RFF was chosen due to the idea of feature selection present in it. RFF is calculated in two phases: during the first phase features are evaluated and the best 100 are selected, reweighted and used in LU similarity calculation during the second phase. In all three approaches the similarity computation is based in some way on Mutual Information weighting, which is also often used by other methods. Finally, Freitag et al.'s approach is one of the few that deal with the similarity of adjectives and verbs, and the only one known to us with which we can directly compare.

Two measures, namely Lin's measure (the predecessor of CRMI and RFF) and RWF, significantly surpass the other two reimplemented measures in almost all cases of the four types of applications: adjectives – all and frequent (> 1000) – and verbs, see Table 1. There only is no statistically significant difference between CRMI and RWF in the case of all verbs, respectively: 71.99% and 73.45%. Following (Dietterich, 1997), we applied McNemar's test in order to check statistical significance of the difference between CRMI and RWF in the case of all verbs. McNemar's test is based on the contingency table of the number of examples misclassified (here an incorrect answer) by both, one of or any classifier. For all verbs we recorded: $n00 = 557$ (wrong answers from CRMI and RWF), $n01 = 262$ (CRMI $-$, RWF $+$), $n10 = 307$ (CRMI $+$, RWF $-$), $n11 = 1959$ (both correct). Because $\chi^2 = 3.40246$ and the threshold for the null hypothesis rejection is 3.841459 (the level of confidence: 95%), the better result of RWF is not significant enough. For all verbs, however, the result of RWF is significantly better than the results of Lin's measure ($\chi^2 = 48.510166$), as well as the result of CRMI in comparison with Lin's measure ($\chi^2 = 20.803951$) .

Lin's measure is closest to RWF in accuracy for frequent verbs: $\chi^2 = 0.023392$, so the difference is not statistically significant. There also is no statistically significant difference in the case of frequent adjectives: $\chi^2 = 2.6940298$.

When applied to *all* adjectives or verbs, however, RWF achieves much better accuracy than Lin's measure with the level of confidence higher than 95%: $chi^2 = 13.326478$ for all adjectives and $chi^2 = 48.510166$ for all verbs. Thus, RWF surpasses all four measures for lexical units that are less frequent, described by fewer and less reliable features. This means that RWF appropriately selects features for infrequent LUs, where some features occur randomly. Moreover, as there is no frequency-based filtering in RWF, it shows the ability to select automatically the most informative features for the given LU. We had observed the same for nouns (Piasecki et al., 2007b).

In the case of RWF, we also determined experimentally the threshold k for the number of features selected, achieving the best results with $k = 5000$ for frequent adjectives, $k = 1000$ for frequent verbs, $k = 1000$ for all adjectives, and $k = 500$ for all verbs.

It should be emphasised that verb matrices have lower percentage of non-zero cells than adjective matrices. It seems that for less dense matrices lower values of k give better results, as the lower ranks are occupied by more accidental features. This observation is also supported by the lower k values identified for experiments with all LU. Infrequent LUs have lower percentage of strongly associated features and more accidental features. Lower values of k result in the selection of only the more reliable features. An automatic mechanism of the k value adjustment on the basis of data analysis would be a valuable extension of the RWF method. It must be emphasised, however, that the range of results achieved for different k values is limited, e.g., in the case of frequent verbs and the joint matrix NSb+NArg+VPart+VAdv: for $k = 100$ 73.23%, $k = 500$ 76.24%, $k = 1000$ **77.12%**, and $k = 5000$ 76.88%. Results become stable around $k = 300$ and only a slight tuning is required by finding the optimal value of k. There was a similar result for nouns (Piasecki et al., 2007b).

The comparison with the (Freitag et al., 2005) is less direct, as we did not reimplemented their approach. They report results for a different wordnet and different corpus. Still, the comparison is very promising for RWF and for the well-known Lin's measure.

In the case of verb constraints, the highest results by a single type of a constraint is generated, surprisingly, by a simple closest adverb identification, see Fig. 2. NArg(dat) and NArg(inst) matrices are too sparse and the identification of

[5]We reimplemented Lin's measure, CRMI and RFF; we *cite* results from (Freitag et al., 2005).

	Frequent LUs				All LUs				*Freitag*
Features	Lin	CRMI	RFF	RWF	Lin	CRMI	RFF	RWF	
NArg(acc)	69.60	66.43	56.06	**72.45**	62.56	62.46	45.64	**66.55**	
NArg(dat)	**44.97**	19.72	37.53	26.05	**33.58**	17.96	28.65	22.24	
NArg(inst)	**64.13**	46.40	49.80	59.07	**52.03**	40.81	41.56	51.02	
NArg(loc)	**64.13**	54.47	50.75	62.79	50.18	44.02	39.55	**50.86**	
Nsb	62.95	58.35	49.49	**63.18**	51.54	52.38	40.58	**54.94**	*63.8*
VPart	**55.66**	42.04	48.54	46.00	**45.90**	34.94	39.48	41.20	
VAdv	72.68	53.60	55.50	**75.30**	62.07	45.67	43.37	**64.02**	
NArg(acc+dat+inst+loc)	74.82	68.65	56.45	**74.98**	65.51	69.47	46.29	**70.15**	
NSb+NArg+VPart+VAdv	76.88	70.23	55.34	**77.12**	68.17	71.99	48.17	**73.45**	
AAdv	60.05	13.40	62.62	**62.81**	48.65	12.94	49.82	**52.19**	
AA	**77.58**	50.47	64.12	76.14	**69.16**	46.30	54.12	68.37	
ANmod	**76.39**	71.01	64.06	75.27	71.68	70.60	58.57	**72.47**	
ANmod+AAdv	77.40	73.14	65.56	**77.71**	72.25	72.33	59.44	**74.71**	*76.4*
(ANmod+AAdv)⊕AA	81.65	75.95	67.44	**82.91**	75.70	75.47	61.29	**77.77**	
ANmod+AAdv+AA	79.65	76.64	66.12	**79.90**	75.50	76.21	60.52	**77.97**	

Table 1: Experiments with MSRs. Frequent LUs had > 1000 occurrences in the IPI PAN corpus. (Freitag et al., 2005) presented results for LUs with > 1000 occurrences in their corpus.

a subject generates too many errors (we do not apply any parser). In the case of a joined matrix, however, RWF selects features enough effectively to achieve a result that is significantly better than any single verb matrix.

In the case of adjectives, the differences of accuracy achieved for different types of constraint are much smaller. The joined matrix is also better than any single one. The claim of (Hatzivassiloglou and McKeown, 1993) that co-occurence of two adjectives in one noun phrase (clearly indicated in Polish by their morphological agreement) is a negative feature is contradicted by the result of AA alone and AA combined with other matrices. Moreover, the difference between the results of (ANmod+AAdv)⊕AA and ANmod+AAdv+AA suggests that the semantic information carried by the AA constraint is in some way orthogonal to other adjective constraint.

In order to compare the results of MSR with human performance, we randomly selected two subsets of 100 question-answer pairs from the complete verb and adjective WBSTs. Next, we asked 20 native speakers of Polish to solve both tests. They were instructed to select for each question word only one answer, the closest in meaning to the question. There was no time limit in the task. All test participant were Computer Science students, but the tests were generated on the basis of verbs and adjectives that are quite frequent and have a rather general meaning. Thus, the background of the participants should not influence the results. The average scores achieved by test participants are presented in Table 2. The inter-judge agreement was measured by Fleiss's kappa, which allows the measure of agreement among many participants (Fleiss, 1971). The high value of kappa, supported by the manual evaluation of the test results, shows that the agreement was high, and the raters made similar errors.

Examples of more difficult verb test QA pairs:

q : *nakazywać* (*command*)

A : ***polecać*** (*order*), *pozostawać* (*remain*), *wkroczyć* (*enter*), *wykorzystać* (*utilise*)

PoS	min [%]	avg [%]	max [%]	kappa
Verb	84	88.21	95	0.84
Adjective	82	88.9	95	0.85

Table 2: Results of a manual WBST for Polish verbs and adjectives.

q : *działać* (*act*)

A : *kwitnąć* (*flourish*), *móc* (*can*), ***rzutować*** (*project*), *zrazić* (*alienate*)

Examples of more difficult adjective test QA pairs:

q : *bolesny* (*painful*)

A : *krytyczny* (*critical*), *nieudolny* (*inept*), *portowy* ((*of*) *port*), ***poważny*** (*serious*)

q : *drastyczny* (*drastic*)

A : *azjatycki* (*Asian*), *doroczny* (*annual*), *nadwiślański* (*located by the Vistula*), ***nieprzyzwoity*** (*indecent*)

The result of our best adjective MSR is very close to the result achieved by humans. For verbs, the difference is comparable to that observed for nouns (Piasecki et al., 2007b) (but the result of verb MSR still approaches human performance).

5. Semantic relatedness and wordnet extensions

The constructed MSRs are intended to assist lexicographers in selecting LUs semantically related to the LU being edited. Lexicographers can find missing synonyms or

instances of semantic relations while browsing the lists of k most closely related LUs (according to the MSRs).
Long suggestion lists may preclude careful analysis. We chose $k = 20$ for a small experiment to imitate the future use of the MSRs by lexicographers. We randomly selected two subsets of LUs, verbs and adjectives. The sizes of the samples were determined in such a way that, with the 95% confidence level according to the method discussed in (Israel, 1992), the results of the manual evalutation perfomed on the samples can be ascribed to the whole sets. For every LU in each subset, we generated the list of the $k = 20$ LUs most related to the given one. One of the co-authors manually assessed all elements on all lists, distinguishing any elements that are in some wordnet relation (Derwojedowa et al., 2007) to the head LU.
The evaluated LU lists were classified into:

- *very useful* – a half, or almost a half, of the LUs on the list are in some semantic relation to the given one,
- *useful* – a sizable part of the list is somehow related,
- *neutral* – several LUs on the list are in some relation, but the lexicographer might miss them,
- *useless* – at most a few LUs may be related.

The results of the manual evaluation appear in Table 3.
Selected lists for verbs – first a LU is given, next the number of similar LUs accepted by the lexicographer, and finally the list of the 20 most similar LUs:[6]

ściągnąć (*take off*) (18): ściągać (*take off (habitual)*) 0.640, zdjąć (*take off*) 0.608, ubrać (*clothe*) 0.575, założyć (*put on*) 0.562, włożyć (*put on*) 0.554, przyciągnąć (*draw*) 0.552, nosić (*wear*) 0.550, odziać (*clothe*) 0.548, przyciągać (*draw (habitual)*) 0.542, zrzucić (*drop off*) 0.538, wyegzekwować (*put into effect*) 0.534, sprowadzić (*bring*) 0.534, przywdziać (*don*) 0.532, kupić (*buy*) 0.532, zgromadzić (*gather*) 0.531, pobierać (*collect*) 0.531, ciągnąć (*pull*) 0.531, podrzeć (*tear up*) 0.531, chwycić (*grasp*) 0.531, nałożyć (*put on*) 0.530

graniczyć (*border*) (8): sąsiadować (*neighbour*) 0.575, przylegać (*abut*) 0.548, położyć (*put down*) 0.537, należeć (*belong*) 0.533, zabudować (*build (on)*) 0.532, zaniedbać (*neglect*) 0.531, dotknąć (*touch*) 0.531, okalać (*encircle*) 0.529, administrować (*administer*) 0.527, otaczać (*surround*) 0.526, biec (*run*) 0.525, dzierżawić (*lease*) 0.525, zagrozić (*threaten*) 0.525, znajdować (*find (habitual)*) 0.524, być (*be*) 0.524, zagospodarować (*bring (into cultivation)*) 0.523, wyłączyć (*exclude*) 0.522, stanowić (*constitute*) 0.521, wydzielić (*separate (from)*) 0.520, użytkować (*utilise*) 0.520,

okupować (*occupy*) (1): opuścić (*leave*) 0.556, protestować (*protest*) 0.550, szturmować (*storm*) 0.550, zajmować (*occupy*) 0.543, wyniszczyć (*exterminate*) 0.543, zjednoczyć (*unite*) 0.541, zająć (*occupy*) 0.541, wtargnąć (*invade*) 0.538, maić (*decorate*) 0.537, zabukować (*book*) 0.536, mieszkać (*live*) 0.536, represjonować (*repress*) 0.536, wybudować (*build*) 0.535, przebywać (*stay*) 0.534, położyć (*put*) 0.534, plasować (*place*) 0.534, znaleźć (*find*) 0.533, awansować (*promote*) 0.533, walczyć (*fight*) 0.533, zaadaptować (*adapt*) 0.533,

Selected lists for adjectives:

niezwykły (*unusual*) (13): wyjątkowy (*exceptional*) 0.325, niebywały (*unprecedented*) 0.285, niesamowity (*uncanny*) 0.279, niepowtarzalny (*incomparable*) 0.266, wspaniały (*excellent*) 0.250, niespotykany (*unparalleled*) 0.236, niecodzienny (*uncommon*) 0.222, niesłychany (*unheard of*) 0.213, cudowny (*miraculous*) 0.204, szczególny (*particular*) 0.202, nadzwyczajny (*extraordinary*) 0.196, nieprzeciętny (*uncommon*) 0.196, zdumiewający (*astonishing*) 0.184, nieprawdopodobny (*improbable*) 0.183, mistyczny (*mystical*) 0.182, fantastyczny (*fantastic*) 0.181, ciekawy (*curious*) 0.174, interesujący (*interesting*) 0.170, niezapomniany (*unforgettable*) 0.168, poetycki (*poetic*) 0.163

agresywny (*aggressive*) (6): brutalny (*brutal*) 0.208, odważny (*brave*) 0.203, dynamiczny (*dynamic*) 0.189, aktywny (*active*) 0.189, energiczny (*energetic*) 0.178, napastliwy (*aggressive*) 0.176, ostry (*sharp*) 0.174, arogancki (*arrogant*) 0.173, wulgarny (*vulgar*) 0.170, zdecydowany (*decided*) 0.170, sprytny (*shrewd*) 0.168, ofensywny (*offensive*) 0.167, skuteczny (*effective*) 0.162, waleczny (*brave*) 0.160, uparty (*obstinate*) 0.159, ambitny (*ambitious*) 0.157, nieobliczalny (*unpredictable*) 0.155, nerwowy (*nervous*) 0.154, wrażliwy (*sensitive*) 0.153, chaotyczny (*chaotic*) 0.153

kurtuazyjny (*courteous*) (1): wykrętny (*evasive*) 0.191, kategoryczny (*categorical*) 0.157, oficjalny (*official*) 0.154, urywany (*intermittent*) 0.142, dyskusyjny (*debatable*) 0.139, lakoniczny (*laconic*) 0.138, kawiarniany (*of café*) 0.135, spontaniczny (*spontaneous*) 0.133, retoryczny (*rhetorical*) 0.133, nieoficjalny (*unofficial*) 0.131, towarzyski (*sociable*) 0.126, stanowczy (*resolute*) 0.122, przyjacielski (*friendly*) 0.121, impulsywny (*impulsive*) 0.121, nieprecyzyjny (*imprecise*) 0.120, rozrywkowy (*entertaining*) 0.119, dobitny (*emphatic*) 0.118, górnolotny (*bombastic*) 0.117, cierpki (*tangy*) 0.116, luźny (*loose*) 0.115

In nearly half of the cases, the lexicographer can find valuable hints on the list generated on the basis of MSRs. Suggestions should help notice specific or domain-restricted uses of LUs. The manual evaluation suggests MSR accuracy much lower than for the WBST, but the latter operates on generic relatedness rather than specific semantic relations. To extract instances of semantic relations, we need additional criteria, for example, lexico-syntactic patterns of occurrence contexts.

[6] Many words on these lists are polysemous in both languages. The English translations "select" the meaning common to the grouping that the list suggests.

PoS	very useful	useful	neutral	useless	no relations
Verb [%]	17.8	37.6	20.0	15.6	9.0
Adjective [%]	19.2	26.3	29.7	14.4	10.4

Table 3: Manual evaluation of MSR for verbs and adjectives.

6. Observations and future work

The RWF for nouns (Piasecki et al., 2007a) exhibits comparable performance for verbs and adjectives. A very small number of morphosyntactic constraints resulted in a relatively high accuracy in the WBST. The results of the WBST are well above the random baseline, and better than reported in (Hatzivassiloglou and McKeown, 1993; Freitag et al., 2005), though we worked with many fewer LUs. We also achieved results closer to human performance than those for nouns (Piasecki et al., 2007b).

The method we propose here should be easily adapted to similar (similarly inflected) languages, especially Slavic languages such as Czech or Russian.

7. References

Gemma Boleda, Toni Badia, and Eloi Batlle. 2004. Acquisition of semantic classes for adjectives from distributional evidence. In *Proceedings of the 20th COLING*, pages 1119Ű–1125. ACL.

Gemma Boleda, Toni Badia, and Sabine Schulte im Walde. 2005. Morphology vs. syntax in adjective class acquisition. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 77–86, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Magdalena Derwojedowa, Maciej Piasecki, Stanisław Szpakowicz, and Magdalena Zawisławska. 2007. Polish WordNet on a shoestring. In *Proceedings of Biannual Conference of the Society for Computational Linguistics and Language Technology, Tübingen, April 11Ű13 2007*, pages 169–178. Universität Tübingen.

Magdalena Derwojedowa, Maciej Piasecki, Stanisław Szpakowicz, Magdalena Zawisławska, and Bartosz Broda. 2008. Words, concepts and relations in the construction of Polish WordNet. In A. Tanâcs, D. Csendes, V. Vincze, Ch. Fellbaum, and P. Vossen, editors, *Proceedings of the Global WordNet Conference, Seged, Hungary January 22–25 2008*, pages 162–177. University of Szeged.

Thomas G. Dietterich. 1997. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 25–32, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Maayan Geffet and Ido Dagan. 2004. Vector quality and distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics, COLING2004*, pages 247–254.

V. Hatzivassiloglou and K. R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st ACL,*, pages 172–182. ACL.

G. Israel. 1992. Determining sample size. Tech. rep., University of Florida.

Maria Lapata. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of the NAACL*, pages 63–70. ACL.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL-97)*, pages 64–71, Madrid, Spain. ACL.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING 1998*, pages 768–774. ACL.

Christopher D. Manning and Hinrich Schütze. 2001. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2007a. Automatic selection of heterogeneous syntactic features in semantic similarity of Polish nouns. In *Proceedings of the Text, Speech and Dialog 2007 Conference*, volume 4629 of *LNAI*. Springer.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2007b. Extended similarity test for the evaluation of semantic similarity functions. In Vetulani (Vetulani, 2007), pages 104–108.

Maciej Piasecki. 2006. Handmade and automatic rules for Polish tagger. In Sojka et al. (Sojka et al., 2006).

Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science PAS.

G. Ruge. 1992. Experiments on linguistically-based term associations. *Information Processing and Management*, 28(3):317–332.

Petr Sojka, Ivan Kopecek, and Karel Pala, editors. 2006. *Proceedings of the Text, Speech and Dialog 2006 Conference*, Lecture Notes in Artificial Intelligence. Springer.

Zygmunt Vetulani, editor. 2007. *Proceedings of the 3rd Language and Technology Conference, October 5–7, 2007, Poznań, Poland*. Wydawnictwo Poznańskie Sp. z o.o., Poznań.

Julie Weeds and David Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475.