# Exploration of Semantic Spaces Obtained from Czech Corpora

Lubomír Krčmář, Miloslav Konopík, and Karel Ježek

Department of Computer Science and Engineering,
University of West Bohemia, Plzeň, Czech Republic
{lkrcmar, konopik, jezek_ka}@kiv.zcu.cz

**Abstract.** This paper is focused on semantic relations between Czech words. Knowledge of these relations is crucial in many research fields such as information retrieval, machine translation or document clustering. We obtained these relations from newspaper articles. With the help of LSA[1], HAL[2] and COALS[3] algorithms, many semantic spaces were generated. Experiments were conducted on various settings of parameters and on different ways of corpus preprocessing. The preprocessing included lemmatization and an attempt to use only "open class" words. The computed relations between words were evaluated using the Czech equivalent of the Rubenstein-Goodenough test. The results of our experiments can serve as the clue whether the algorithms (LSA, HAL and COALS) originally developed for English can be also used for Czech texts.

**Keywords:** Information retrieval, Semantic space, LSA, HAL, COALS, Rubenstein-Goodenough test

## 1 Introduction

There are many reasons to create a net of relations among words. As with many other research groups, we are trying to find a way how to facilitate information retrieval. Question answering and query expansion are our main interests. We try to employ nets of words in these fields of research. Not only can people judge whether two words have something in common (they are related) or that they are similar (they describe the same idea). Computers with their computational abilities can also make some conclusions about how words are related with each other. Their algorithms exploit the Harris distributional hypothesis [1], which assumes that terms are similar to the extent to which they share similar linguistic contexts. Algorithms such as LSA, HAL and novel COALS were designed to compute the lexical relations automatically. Our belief is that these methods have not yet been sufficiently explored for other languages than English. A

[1] Latent Semantic Analysis
[2] Hyperspace Analogue to Language
[3] the Correlated Occurrence Analogue to Lexical Semantics

V. Snášel, J. Pokorný, K. Richta (Eds.): Dateso 2011, pp. 97–107, ISBN 978-80-248-2391-1.

great motivation for us was also the S-Space package [2]. The S-Space package is a freely available collection of implemented algorithms dealing with text corpora. LSA, HAL and COALS algorithms are included. Our paper evaluates the applicability of these popular algorithms to Czech corpora.

The rest of the paper is organized as follows. The following section deals with related works. The next section describes the way we created semantic spaces for ČTK[4] corpora. Our experiments and evaluations using the RG benchmark are presented in section 4. In the last section we summarize our experiments and present our future work.

## 2 Related works

The principles of LSA can be found in [3], the HAL algorithm is described in [4]. A great inspiration for us was a paper about the COALS algorithm [5], where the power of COALS, HAL and LSA is compared. The Rubenstein-Goodenough [6] benchmark and some other similar tests such as Miller-Charles [7] or Word-353 are performed. The famous TOEFL[5] or ESL[6] are also included in the evaluation.

We also come from a paper written by Paliwoda [8] where the Rubenstein-Goodenough (RG) test translated into Polish was used. Alternative ways of evaluating semantic spaces can be found in [9] by Bullnaria and Levy.

Different methods which judge how some words are related exploit lexical databases such as WordNet [10]. There are nouns, verbs, adjectives and adverbs grouped in sets of synonyms called synsets in WordNet. Each synset expresses a distinct concept and all the concepts are interlinked with relations including hypernymy, hyponymy, holonymy or meronymy. Although lexical-based methods are popular and still under review, we have decided to follow the fully automatic methods.

## 3 Generation of Semantic Spaces

The final form of semantic space is firstly defined by the quality of the corpus used [9] and secondly by the selection of algorithm. The following chapter applies to the features of our corpus and also describes the ways we preprocessed it. The next chapter is focused on parameter settings of LSA, HAL and COALS.

### 3.1 Corpus and corpus preprocessing

The ČTK 1999 corpus, which consists of newspaper articles, was used for our experiments. The ČTK corpus is one of the largest Czech corpuses we work with in our department. For lemmatization, Hajic's tagger for the Czech language was used [11].

[4] Česká Tisková Kancelář (Czech News Agency)
[5] Test of English as a Foreign Language
[6] English as a Second Language

There was no further preprocessing of input texts performed. Finally, 4 different input files[7] for the S-Space package were used. The first input file contained plain texts of the ČTK corpora. The second one contained plain text without stopwords. Pronouns, prepositions, conjunctions, particles, interjections and punctuation[8] were considered as stopwords in our experiments. That means that removing stopwords from the text in our paper is the same as keeping only open class words in the text. The third file contained lemmatized texts of the ČTK corpora. And the last file contained lemmatized ČTK corpora without stopwords. Statistics on the texts of the corpus are depicted in Table 1, statistics on texts without stopwords are depicted in Table 2 respectively.

**Table 1.** ČTK corpus statistics

| | Plain texts | Lemmatized texts |
|---|---|---|
| Documents' count | 130,956 | |
| Tokens' count | 35,422,517 | |
| Different tokens' count | 579,472 | 291,090 |
| Tokens' count occurring more than once | 35,187,747 | 35,296,478 |
| Different tokens' count occurring more than once | 344,702 | 165,051 |

**Table 2.** ČTK corpus statistics, stopwords removed

| | Plain texts | Lemmatized texts |
|---|---|---|
| Documents' count | 130,956 | |
| Tokens' count | 22,283,617 | |
| Different tokens' count | 577,297 | 290,036 |
| Tokens' count occurring more than once | 22,049,467 | 22,158,048 |
| Different tokens' count occurring more than once | 343,147 | 164,467 |

### 3.2 Settings of algorithms

The LSA principle differs essentially from HAL and COALS. While HAL and COALS are window-based, LSA deals with passages of texts. The passage of text is presented by a whole text of any article of the ČTK corpus in our case.

[7] Each file contained each document of the corpora. One file line corresponds with one distinct document.

[8] Punctuation is rather a token than a word. It was removed while it is not important for the LSA algorithm.

Both LSA and COALS exploit untrivial mathematical operation SVD[9] while HAL does not. COALS simply combines some HAL and LSA principles [5].

The S-Space package provides default settings for its algorithms. The settings are based on previous research. The default settings of parameters are depicted in Table 3. We tried to change some values of parameters because of the Czech language of our texts. The Czech language differs from English especially in the number of forms for one word and in word order, which is as not strictly fixed as in English. Therefore, there are more different terms[10] for Czech language texts. Since the algorithms are sensitive to the term occurrence, this is one of the reasons[11] we tried to remove low occurring words.

Another parameter we observed is HAL's window size. It was expected that the more terms for the Czech language meant the smaller window size would be more appropriate.

The last parameters we changed from defaults were HAL and COALS retained columns' counts. We reduced the dimensionality of spaces in this way by setting the reduction property to the values adopted from [4]. As a consequence, columns with high entropy were retained. To reduce the dimensionality of the COALS algorithm, the impact of SVD was also tested.

**Table 3.** The default settings of algorithms provided by the S-Space package

| Algorithm | property | value |
|---|---|---|
| LSA | term-document matrix transform | log-entropy weighting |
| | the number of dimensions in the semantic space | 300 |
| HAL | window size | 5 |
| | weighting | linear weighting |
| | retain property | retain all columns |
| COALS | retain property | retain 14,000 columns |
| | window size | 4 |
| | reduce using SVD | no |

## 4 Evaluation of Semantic Spaces

Several approaches exist to evaluate semantic spaces as noticed in section 2. Unfortunately, most of the standard benchmarks are suitable only for English. To the best of our knowledge, there is no similar benchmark to the Rubenstein-Goodenough (RG) test or to the Miller-Charles test for the Czech language. Therefore we have decided to translate RG test into Czech.

[9] Singular Value Decomposition
[10] One word in two forms means two terms in this context.
[11] Another reason is to decrease the computation costs.

The following chapter describes the origination of the Czech equivalent of the RG test. The next chapter comprises our results on this test for many generated semantic spaces.

### 4.1 Rubenstein-Goodenough test

The RG test comprises pairs of nouns with corresponding values from 0 to 4 indicating how much words in pairs are related. The powers of relations were judged by 51 humans in 1965. There were 65 word pairs in the original English RG test.

The translation of the original English RG test into Czech was performed by a Czech native speaker. The article by Pilot [12] describing the original meanings of the RG test's words was exploited. The resulting translation of the test was corrected by 2 Czech native speakers who are involved in information retrieval.

After our translation of the RG test into Czech, 62 pairs are left. We had to remove the "midday-noon", "cock-rooster" and "grin-smile" pairs because we couldn't find any appropriate and different translations for both words of these pairs in Czech. Our Czech RG test[12] was evaluated by 24 Czech native speakers with differing education, age and sex. Pearson's correlation between Czech and English evaluators is 0.94.

A particular word we removed from our test before comparing it with semantic spaces is "crane". The Czech translation of this word has 3 different meanings. Furthermore, only one of these meanings was commonly known by the people who participated in our test. Therefore, another 3 pairs disappeared: "bird-crane", "crane-implement" and "crane-rooster". A similar ambiguate word is the Czech translation of "mound", which was also used in a different meaning in the corpus. We removed it with these 4 pairs: "hill-mound", "cemetery-mound", "mound-shore", "mound-stove". In the end, 55 word pairs were left in our test.

Another issue we had to face was the low occurrence of the RG test's words in our corpus. Therefore, we tried to remove the least frequent words of the RG test in sequence and the pairs which they appear in as a consequence. In the end it turned out that especially this step showed us that the relations obtained from S-Space algorithms correlate with human judgments quite well. To evaluate which of the semantic spaces best fits with human judgments the standard Pearson's correlation coefficient was used.

### 4.2 Experiments and results

We created many semantic spaces with the LSA, HAL and COALS algorithms. Cosine similarity was used to evaluate whether two words are related in semantic spaces. Other similarity metrics did not work well.

The obtained results for the different semantic spaces are depicted in Table 4 for plain texts of the ČTK corpora and in Table 5 for lemmatized texts. The

[12] Available at `http://home.zcu.cz/~lkrcmar/RG/RG-ENxCZ.pdf`

best 2 scores in Table 4 and the best 3 scores in Table 5 are highlighted for each tested set of pairs in our RG test.

**Table 4.** Correlation between values for pairs obtained from different semantic spaces and the Czech Rubenstein-Goodenough test. The word pairs containing low occurring words in the corpora were omitted in sequence (o-27 means 27 pairs out of the original 65 were omitted while computing the correlation). N - no stopwords, m2 - words occurring more than once in the corpora are retained for the computation, s1 - window size = 1, d2 - reduce to 200 dimensions using the SVD.

| Semantic space | o-27 | o-29 | o-32 | o-35 | o-37 | o-44 | o-51 |
|---|---|---|---|---|---|---|---|
| LSA_m2 | 0,26 | 0,25 | 0,27 | 0,33 | 0,35 | 0,36 | 0,24 |
| N_LSA | 0,28 | 0,28 | 0,29 | 0,33 | 0,33 | 0,33 | 0,16 |
| N_LSA_m2 | 0,27 | 0,26 | 0,29 | 0,33 | 0,30 | 0,32 | 0,11 |
| HAL_m2 | 0,20 | 0,19 | 0,24 | 0,28 | 0,25 | 0,24 | 0,14 |
| HAL_m2_s1 | 0,12 | 0,11 | 0,18 | 0,19 | 0,14 | 0,06 | 0,04 |
| HAL_m2_s2 | 0,17 | 0,18 | 0,25 | 0,30 | 0,25 | 0,18 | 0,15 |
| N_HAL_m2 | 0,36 | 0,38 | 0,39 | 0,43 | 0,43 | 0,44 | 0,44 |
| N_HAL_m2_s1 | 0,39 | 0,41 | 0,43 | 0,47 | 0,46 | 0,48 | 0,53 |
| N_HAL_m2_s2 | 0,40 | 0,42 | 0,44 | 0,48 | 0,48 | 0,49 | 0,53 |
| COALS_m2 | **0,43** | **0,45** | **0,48** | **0,52** | **0,54** | **0,57** | **0,62** |
| COALS_m2_d2 | 0,28 | 0,30 | 0,30 | 0,35 | 0,38 | 0,39 | 0,42 |
| COALS_m2_d4 | 0,17 | 0,18 | 0,18 | 0,19 | 0,21 | 0,27 | 0,32 |
| N_COALS_m2 | 0,42 | 0,43 | **0,46** | **0,50** | **0,53** | 0,54 | **0,59** |
| N_COALS_m2_d2 | 0,31 | 0,27 | 0,25 | 0,35 | 0,31 | 0,23 | 0,34 |
| N_COALS_m2_d4 | **0,43** | **0,44** | 0,45 | 0,50 | 0,51 | 0,51 | 0,57 |

It turned out we do not have to take into account words which occur only once in our corpora. It saves computing time without a negative impact on the results. This is the reason most of our semantic spaces are computed omitting words which occur only once.

The effect of omitting stopwords is very small for the LSA and the COALS algorithms. However, the HAL algorithm scores are affected a lot (compare HAL and N_HAL in Tables 4 and 5). This difference in results can be caused by the fact that LSA does not use any window and works with whole texts. The COALS algorithm may profit from using the correlation principle[5] that helps it to deal with stopwords.

The scores in our tables show that especially the COALS method is very successful. The best scores are achieved by COALS for the plain texts, and COALS scores for the lemmatized texts are also among the best ones (compare Table 4 and 5).

The HAL method is also very successful. Furthermore, the best score of 0.72 is obtained using the HAL method on lemmatized data without stopwords (see Table 5). It turns out that HAL even outperforms COALS when only pairs

**Table 5.** Correlation between values for pairs obtained from different semantic spaces and the Czech Rubenstein-Goodenough test. The word pairs containing low occurring words in the corpora were omitted in sequence (o-27 means 27 pairs out of the original 65 were omitted while computing the correlation). N - no stopwords, m2 - words occurring more than once in the corpora are retained for the computation, s1 - window size = 1, r14 - only 14,000 columns retained, d2 - reduce to 200 dimensions using the SVD.

| Semantic space | o-14 | o-19 | o-24 | o-27 | o-29 | o-32 | o-35 | o-37 | o-44 | o-51 |
|---|---|---|---|---|---|---|---|---|---|---|
| LSA | 0,19 | 0,22 | 0,25 | 0,33 | 0,35 | 0,35 | 0,44 | 0,47 | 0,48 | 0,47 |
| LSA_m2 | 0,15 | 0,19 | 0,22 | 0,30 | 0,33 | 0,33 | 0,41 | 0,46 | 0,47 | 0,41 |
| N_LSA | 0,16 | 0,18 | 0,20 | 0,30 | 0,33 | 0,36 | 0,44 | 0,46 | 0,47 | 0,37 |
| N_LSA_m2 | 0,17 | 0,19 | 0,21 | 0,32 | 0,36 | 0,37 | 0,43 | 0,47 | 0,47 | 0,39 |
| HAL | 0,35 | 0,44 | 0,45 | 0,47 | 0,48 | 0,53 | 0,57 | 0,54 | 0,57 | 0,41 |
| HAL_m2 | 0,35 | 0,44 | 0,45 | 0,47 | 0,48 | 0,53 | 0,57 | 0,53 | 0,57 | 0,41 |
| HAL_m2_s1 | 0,37 | 0,41 | 0,41 | 0,41 | 0,42 | 0,48 | 0,50 | 0,47 | 0,49 | 0,34 |
| HAL_m2_s2 | 0,45 | 0,51 | 0,52 | 0,54 | 0,57 | **0,62** | **0,68** | **0,64** | **0,67** | 0,56 |
| HAL_m2_s10 | 0,26 | 0,41 | 0,43 | 0,48 | 0,48 | 0,54 | 0,56 | 0,52 | 0,56 | 0,35 |
| HAL_m4 | 0,35 | 0,44 | 0,45 | 0,47 | 0,48 | 0,53 | 0,57 | 0,53 | 0,57 | 0,41 |
| HAL_r14 | 0,40 | 0,47 | 0,48 | 0,50 | 0,52 | **0,58** | **0,61** | 0,57 | **0,62** | 0,48 |
| HAL_r7 | 0,39 | 0,46 | 0,46 | 0,48 | 0,50 | 0,55 | 0,58 | 0,54 | 0,58 | 0,43 |
| N_HAL_m2 | 0,22 | 0,26 | 0,29 | 0,34 | 0,35 | 0,33 | 0,36 | 0,37 | 0,39 | 0,26 |
| N_HAL_m2_s1 | 0,43 | 0,45 | 0,49 | 0,52 | 0,55 | 0,55 | **0,62** | **0,64** | **0,68** | **0,72** |
| N_HAL_m2_s2 | 0,34 | 0,37 | 0,40 | 0,44 | 0,48 | 0,48 | 0,54 | 0,55 | 0,61 | **0,61** |
| COALS | **0,52** | 0,53 | **0,55** | **0,54** | **0,57** | 0,54 | 0,58 | 0,55 | 0,57 | 0,61 |
| COALS_m2 | **0,52** | **0,53** | **0,55** | **0,55** | **0,57** | 0,54 | 0,58 | 0,55 | 0,57 | **0,61** |
| COALS_m2_r7 | 0,52 | **0,53** | 0,53 | 0,52 | 0,54 | 0,53 | 0,56 | 0,55 | 0,56 | 0,59 |
| COALS_m2_d2 | 0,22 | 0,22 | 0,42 | 0,40 | 0,38 | 0,40 | 0,43 | 0,40 | 0,48 | 0,42 |
| COALS_m2_d4 | 0,32 | 0,35 | 0,40 | 0,41 | 0,43 | 0,46 | 0,41 | 0,42 | 0,40 | 0,56 |
| COALS_m4 | 0,48 | 0,48 | 0,50 | 0,50 | 0,52 | 0,50 | 0,54 | 0,52 | 0,53 | 0,55 |
| N_COALS_m2 | **0,53** | **0,54** | **0,57** | **0,56** | **0,59** | **0,56** | 0,60 | **0,59** | 0,59 | 0,60 |
| N_COALS_m2_d2 | 0,26 | 0,27 | 0,22 | 0,28 | 0,31 | 0,32 | 0,41 | 0,45 | 0,45 | 0,55 |
| N_COALS_m2_d4 | 0,32 | 0,34 | 0,38 | 0,43 | 0,46 | 0,45 | 0,51 | 0,51 | 0,56 | 0,53 |

containing only very common words are left. On the other hand, this shows the strength of COALS when also considering low occurring words in our corpora.

It turned out that the LSA algorithm is not as effective as the other algorithms in our experiments. Our hypothesis is that scores of LSA would be better when experimenting with larger corpora such as Rohde [5]. However, the LSA scores also improve when considering only common words. This Figure 1 shows the performance of the 3 tested algorithms for the best settings found.

Our results differ from scores of tests evaluated on English corpora and performed by Rohde [5]. His scores for HAL are much lower than ours. On the other hand, his scores for LSA are higher. Therefore, we believe that the performances of the algorithms are language dependent.

The last Figure 2 in our paper compares human and HAL judgments about the relatedness of 14 pairs containing the most common words from the RG word list in the ČTK corpora. The English equivalents to the Czech word pairs are listed in Table 6. We can notice the pairs which spoil the scores of the tested algorithms in the graph. The graph also shows the difference in human and machine judgments. The pair "automobile-car" is less related than "food-fruit" for the algorithms than for humans. On the other hand, the words of the pair "coast-shore" are more related for our algorithms than for humans.

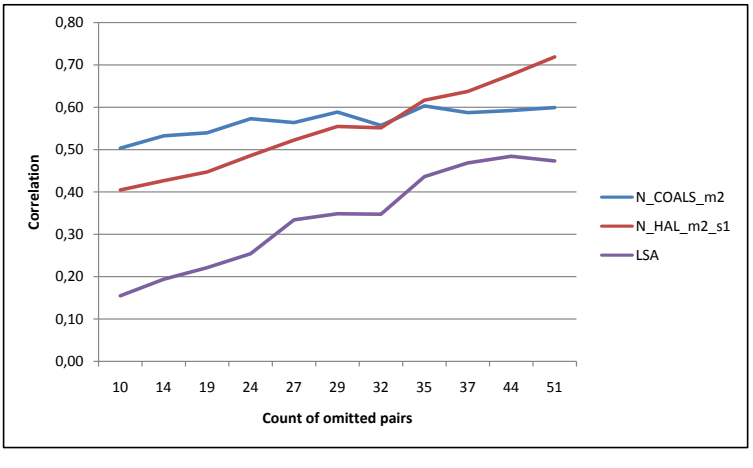


**Fig. 1.** Graph depicting the performances of LSA, HAL and COALS depending on leaving out rare words in the corpora. Our best settings found for algorithms are chosen.

## 5 Conclusion

Our experiments showed that HAL and COALS algorithms performed well and better than LSA on the Czech corpora. Our hypothesis based on our results is

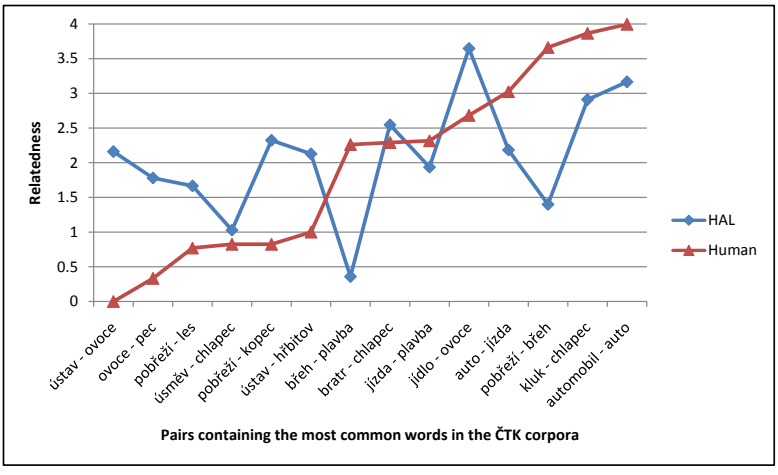


**Fig. 2.** Graph depicting the comparison between human and HAL judgments (value of Cosine similarity of vectors multiplied by 4 is used) about the relatedness of words in pairs. The pairs from the RG test containing only the most common words in the ČTK corpora are left. Our best HAL setting is chosen. The pairs on the X axis are sorted according to the human similarity score.

**Table 6.** The English translation of the Czech word pairs in Figure 2

| Czech word pair | English equivalent | Czech word pair | English equivalent |
|---|---|---|---|
| ústav - ovoce | asylum - fruit | bratr - chlapec | brother - lad |
| ovoce - pec | fruit - furnace | jízda - plavba | journey - voyage |
| pobřeží - les | coast - forest | jídlo - ovoce | food - fruit |
| úsměv - chlapec | grin - lad | auto - jízda | car - journey |
| pobřeží - kopec | coast - hill | pobřeží - beh | coast - shore |
| ústav - hřbitov | asylum - cemetery | kluk - chlapec | boy - lad |
| břeh - plavba | shore - voyage | automobil - auto | automobile - car |

that COALS semantic spaces are more accurate for low occurring words, while semantic spaces generated by HAL are more accurate for pairs of words with higher occurrence. Our experiments show that the lemmatization of corpora is the appropriate approach to improve the scores of algorithms. Furthermore, the best scores of correlation were achieved when only the "open class" words were used.

It turned out that the translation of the original English RG test was not so appropriate for our Czech corpora while it contains words which are not so common in the corpora. However, we believe that when the pairs containing low occurring words were removed, the applicability of the test was improved. The evidence for this is a discovered dependency between the scores of tested algorithms on omitting pairs with low occurring words in them.

We believe that semantic spaces are applicable for the query expansion task which we will focus on in our future work. Apart from this, we are attempting to get some larger Czech corpora for our experiments. We also plan to continue testing the HAL and COALS algorithms, which performed well during our experiments.

## Acknowledgment

The work reported in this paper was supported by the Advanced Computer and Information Systems project no. SGS-2010-028. The access to the MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is also highly appreciated. Finally, we would like to thank the Czech News Agency for providing text corpora.

## References

1. Harris, Z. (1954). Distributional structure. (J. Katz, Ed.) *Word Journal Of The International Linguistic Association, 10*(23), 146-162. Oxford University Press.
2. Jurgens and Stevens, (2010). The S-Space Package: An Open Source Package for Word Space Models. In *System Papers of the Association of Computational Linguistics*.
3. Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2), 259-284. Routledge.
4. Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav Res Methods Instrum Comput, 28*(2), 203-208 203208.
5. Rohde, D. T., Gonnerman, L., & Plaut, D. (2004). An improved method for deriving word meaning from lexical co-occurrence. *Cognitive Science*.
6. Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8*(10), 627-633. ACM Press.
7. Miller, G., & Charles, W. (1991). Contextual Correlates of Semantic Similarity. *Language & Cognitive Processes, 6*(1), 1-28. Psychology Press.
8. Paliwoda-Pękosz, G., Lula, P.: Measures of Semantic Relatedness Based on Wordnet. In: International workshop for PhD students, 2009 Brno. ISBN 978-80-214-3980-1

9. Bullinaria, J., & Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods, 39*(3), 510-526. Psychonomic Society Publications.
10. George A. Miller. 1995. Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41. ACM.
11. J. Hajič, A. Böhmová, E. Hajičová, B. Vidová Hladká, The Prague Dependency Treebank: A Three-Level Annotation Scenario. In A. Abeillé (ed.): *Treebanks Building and Using Parsed Corpora*. pp. 103-127. Amsterdam, The Netherlands: Kluwer, 2000.
12. OShea, J., Bandar, Z., Crockett, K., & McLean, D. (2008). Pilot Short Text Semantic Similarity Benchmark Data Set: Full Listing and Description. *Computing*.