How Similar: Word Similarity Judgments in English and Hungarian

Ágoston Tóth[1]

Abstract: This paper aims to discuss how computers can handle word similarity and how we can measure word similarity in human subjective tests. We compare the results of the subjective human and objective computational experiments and find a significant correlation of r=0.56 at p=0.001. We also compare the results of English and Hungarian human word similarity experiments and find a strong correlation (r>0.93) between the existing English and our Hungarian data.

Keywords: lexical semantics, computational linguistics, distributed semantics

1. Introduction

Word similarity is calculated for many purposes in Computational Linguistics including information retrieval, word sense disambiguation, enriching search operations and automatic thesaurus extraction. Word similarity is usually measured by exploiting existing lexical databases, such as WordNet, or by collecting word co-occurrence data from large text corpora. Both methods have their strengths and weaknesses.

Manually compiled and maintained *lexical databases* (including WordNet, VerbNet and FrameNet) are reliable sources of information. The most widely used database is WordNet, a semantic database that implements a graph with nodes representing senses (synonym sets or synsets) and edges representing semantic relations among senses. The presence of semantic relations *hypernymy* and *hyponymy* make it possible to exploit WordNet as an ontological hierarchy. You can determine the similarity of two words in this hierarchy by measuring the length of the path that connects them. The shorter the path, the more closely they are related and the more similar they are. Intuitively, edges may represent different semantic distances; therefore, a weighted or otherwise enhanced implementation of this technique may be beneficial (see, for instance, Resnik 1995). Also note that the semantic relations in WordNet relate synonym sets (senses) rather than words, therefore *1)* the entire process relies on sense enumeration and on how senses are "split" or "lumped together" while enumerating senses and *2)* you must also find the right sense (synonym set) before carrying out a distance measurement. In

[1] Dr. Ágoston Tóth is Senior Reader at the Institute of English and American Studies of the University of Debrecen. His academic interests include all aspects of computational linguistics, lexicon design, distributed semantics, and he is also genuinely interested in neural, parallel/distributed machine learning. He has conducted research in computational linguistics, psycholinguistics, lexicography, and corpus linguistics. He obtained his Ph.D. degree in Linguistics in 2006. The title of his dissertation is "Perspectives on the lexicon" (Budapest: Akadémiai Kiadó, 2008), which surveys pitfalls and progress in various aspects of lexicon design. E-mail: toth.agoston@arts.unideb.hu

many cases, we see simplistic heuristic solutions to these problems, e.g. Resnik (1995) focuses on the potentially most similar synonym sets.

Methods based on *data acquired from large corpora* are not directly affected by the above problems. The underlying idea is that co-occurrence statistical information is treated as an evidence for a word's potential for replacing another word in the same context and those words that exhibit a greater potential of replacing each other in the context of the same words are said to be more similar. According to the *distributional hypothesis*, this similarity is a semantic phenomenon. The resulting computational semantic approach is called *Distributed Semantics* (DS). A growing literature of DS-based linguistic research has been published in the fields of automatic lexical acquisition, induction of semantic classes/relations and automatic thesaurus extraction. The main Natural Language Processing tasks that can benefit from Distributed Semantics are Word Sense Disambiguation (WSD) and Information Retrieval (IR). For a review of the field, see Turney and Pantel (2010).

Similarity information gathered in this way is linguistically less reliable, less transparent and noisier than the information extracted from manually edited databases. On the other hand, this method is cost and time effective and it may be the only solution if relevant hand-made databases are not available. This method of measuring similarity does not rely on disambiguation or sense enumeration, which can also be a decisive factor.

To measure the precision of a computerized system that calculates the similarity of words, a semantic task is solved and the results are compared to known targets that function as a key. The evaluation task can be a multiple choice test: for an input word, the system selects the most similar word from a list of candidates then the automatically selected answer is compared to the key. A variation of this evaluation method is the TOEFL test, in which the system answers TOEFL exam questions. In many cases, multiple experimental runs are carried out and the parameters of the system are fine-tuned so that the performance indicators (precision and recall) can be maximized (see, for instance Bullinaria and Levi 2007, Dobó and Csirik 2012).

The next section of this paper presents an alternative way of measuring the performance of automated systems – based on human intuition.

2. Human word-similarity judgments

Rubenstein and Goodenough (1965) carried out an experiment in which students judged and scored the similarity of words presented to them on cards. Firstly, they sorted the cards according to similarity. Secondly, they graded the similarity of words on each card by scoring them from 0.0 to 4.0. While the authors refer to *synonymy* as the phenomenon under investigation in their paper, the subjects were instructed to grade 'similarity' and no reference was made to synonymy (or other semantic relation) in the instructions. Note that the method reflects the intuition that word similarity is a continuum rather than a binary phenomenon. The word pairs and the similarity score averages

measured by Rubenstein and Goodenough are shown in the "Word pair" and "R&G (1965)" columns of Table 1, respectively.

In the second part of their investigation, one hundred participants were asked to produce sentences using the words from the cards. Then the target words were characterized by the authors using the context in which they appeared in the elicited sentences. Through comparing these characteristic context patters, similarity scores were computed. These scores correlated well with the results of the human (subjective) test. Note that eliciting text in such a way would be unnecessary (and also: obsolete) today since large corpora are available for the same purpose, but at that time, large machine-readable corpora were unavailable (including the 1-million-word Brown corpus). Of course, processing large corpora, i.e. millions of words would have also required considerable processing power. Nevertheless, Rubenstein and Goodenough's research is a precursor of modern DS-related studies. Importantly, they also introduced an evaluation method for word similarity studies (DS-based or other) that has human empirical basis and relevance.

Miller and Charles (1991) examined the relationship between subjective *similarity* and objective *substitutability of a word in a given textual context*. For measuring similarity, they replicated Rubenstein and Goodenough's experiment by selecting 30 word pairs from the 1965 test suite and carrying out a similar investigation. The pairs included 10 "clear" cases of synonymy, 10 pairs with an obvious lack of similarity and 10 cases in between. The method of collecting human similarity data remained the same. They referred to the process as estimating semantic similarity. Correlation with the original 1965 experiment was very strong (Pearson r=0.97), despite the fact that they changed a word pair from (*cord, smile*) to (*chord, smile*), probably by mistake.

Miller and Charles compared the subjective similarity data with co-occurrence data gained from corpora; for this reason, their study was also a contribution to the literature of Distributed Semantics. They concluded that substitutability (objective, corpus-based similarity) and perceived semantic similarity go hand in hand.

Lexical databases in general, and WordNet in particular, contributed significantly to the area of word similarity measurements (see Meng, Huang and Gu 2013 for a review). In WordNet, a concept is represented by a *synonym set*. Synonym sets store words that are – potentially – semantic equivalents of each other in a given context. Synonym sets are connected by *semantic relations*. These relations include hypernymy and hyponym; they are important because by establishing links between more generic and more specific concepts, they create *is-a* concept hierarchies. For instance, we know that a *robin* is an *animal* by following the hypernym/hyponym links between them. We can exploit these hierarchies (and, of course, the synonyms in the synsets) for quantifying word similarity: the shorter the path between the two concepts, the more similar they are. For instance, the distance between the synonym set containing *train* and the one having *public transport* is one "edge" (one step), the distance between *boy* and *male person* is also one step, while we observe three steps between *train* and *instrumentation* and also three steps between *boy* and *being.* An abstract node

labelled *physical object* dominates both the *being* and the *instrumentation* nodes; therefore, we can also measure the distance between the concepts *boy* and *train*. There are, of course, more sophisticated metrics than this one, e.g. those that do not treat all edges equal: the semantic distances between neighbouring concepts are not necessarily uniform.

Resnik (1995) measured word similarity by calculating WordNet concept-distance information (as described above) and also took information content gained from the Brown corpus into consideration. For testing the model, he repeated the subjective measurement from Miller and Charles (1991) to get his benchmark set. He replicated the experiment with ten students and concluded that his data correlated well with the Miller and Charles data series (r=0.97). The difference between his dataset and the Miller and Charles dataset lied in the omission of the pairs that contained the word *woodland*. The correlation between the subjective and the objective (WordNet-based) similarity data was strong (r=0.79).

Table 1 shows the results of the human similarity measurements from the three papers discussed above. The r=0.96 correlation between the 1965 and the 1991 experiments, as well as the r=0.97 correlation between the 1991 and the 1995 experiments show that the results are repeatable and the evaluation method is reliable. From this perspective, it is also interesting to see how well each individual subject's scores correlated with the average. In Resnik's paper, this correlation was r=0.88 and the standard deviation was very low at 0.08 – this increases the feel of reliability. Resnik also pointed out that this r=0.88 correlation can be treated as the human *upper limit* of precision in solving such a task. We cannot expect better results than this in solving the same task using computers.

#	**Word pair**		**R & G (1965)**	**M & C (1991)**	**Resnik (1995)**
1	car	automobile	3.92	3.92	3.9
2	gem	Jewel	3.84	3.84	3.5
3	journey	Voyage	3.84	3.58	3.5
4	boy	Lad	3.82	3.76	3.5
5	coast	Shore	3.60	3.70	3.5
6	asylum	madhouse	3.04	3.61	3.6
7	magician	Wizard	3.21	3.50	3.5
8	midday	Noon	3.94	3.42	3.6
9	furnace	Stove	3.11	3.11	2.6
10	food	Fruit	2.69	3.08	2.1
11	bird	Cock	2.63	3.05	2.2
12	bird	Crane	2.63	2.97	2.1
13	tool	implement	3.66	2.95	3.4
14	brother	Monk	2.74	2.82	2.4
15	lad	Brother	2.41	1.66	1.2
16	crane	implement	2.37	1.68	0.3

17	journey	Car	1.55	1.16	0.7
18	monk	Oracle	0.91	1.10	0.8
19	cemetery	woodland	1.18	0.95	-
20	food	Rooster	1.09	0.89	1.1
21	coast	Hill	1.26	0.87	0.7
22	forest	graveyard	1.00	0.84	0.6
23	shore	woodland	0.90	0.63	-
24	monk	Slave	0.57	0.55	0.7
25	coast	Forest	0.85	0.42	0.6
26	lad	Wizard	0.99	0.42	0.7
27	cord* / chord**	Smile	0.02 *	0.13 **	0.1 **
28	glass	magician	0.44	0.11	0.1
29	rooster	Voyage	0.04	0.08	0.0
30	noon	String	0.04	0.08	0.0

Table 1. Subjective word similarity scores for English word pairs

3. Hungarian subjective word similarity data

This section of the present paper reports on the findings of a subjective human word similarity experiment. Similarity values for 31 pairs of Hungarian words have been collected, statistically evaluated and compared to the original English data presented in section 2. (In section 4, I will also compare the human results with objective word similarity data collected from a Hungarian corpus.) The words I have used in the subjective experiment are the Hungarian equivalents of the English words selected for the original experiments. Because of the *cord-chord* alternation mentioned above, 31 word pairs have been included in my test suite, which makes it possible to compare my results with all three English data sets presented earlier. 28 subjects have participated in the experiment, 14 male - 14 female, with an age average of 33 years.

The participants have been instructed to

- read all cards,
- sort the cards according to the similarity of meaning beginning with those cards that show the most similar words,
- score the similarity of words on each card (4.0-0.0), where 0.0 stands for no similarity, 4.0 for “complete” similarity – multiple cards can have the same grade, decimals are allowed.

Some of the participants have expressed their wish to make their decision criteria explicit. On the basis of this direct feedback, I can conclude that synonymy was a decisive factor for them – just as expected. If synonymy was not apparent, the participants have looked for associative links. One participant has also added the similarity of the *form* as a third factor. Hypernymy and hyponymy were not mentioned, but also note that hypernyms and hyponyms were almost missing from the test suite

except for the pairs *madár-kakas* [*bird-cock*] and *madár-daru* [*bird-crane*]. On a side note, methods based on lexical databases rely on hypernymy hierarchies and they reach strong correlation with human similarity judgements (cf. Resnik 1995); therefore, it would be interesting to see more superordinate and subordinate terms in human similarity judgements.

Human similarity decisions are potentially influenced by polysemy and homonymy. In the present investigation, examples have included the word-pairs *testvér-szerzetes* [*brother-monk*] and *legény-testvér* [*lad-brother*]. The participants have not been instructed to consider the most relevant senses only or to contemplate other meanings. My data shows that range and deviation are high in these cases.

The results of the experiment are shown in Table 2. Each participant has produced a data series correlating well with the average (r=0.9) and the standard deviation has remained low at 0.04. These values are very similar to what we have seen in Resnik's experiment (r=0.88, std dev=0.08) – Resnik quotes these values as the human upper limit in estimating similarity.

Word pair		**English eqv. (# in table 1)**	**Avg. similarity (n=28)**	**Range**	**Std. deviation**
autó	Személygépkocsi	1	3.92	0.8	0,18
drágakő	Ékkő	2	3.74	1	0.32
dél	Délidő	8	3.72	1	0.32
elmegyógyintézet	Bolondokháza	6	3.69	3	0.65
fiú	Legény	4	3.56	2	0.46
part	Vízpart	5	3.54	1.6	0.39
szerszám	Eszköz	13	3.36	2	0.50
kályha	Tűzhely	9	3.26	3.4	0.65
bűvész	Varázsló	7	3.26	3	0.72
utazás	Út	3	3.24	1.4	0.38
étel	Gyümölcs	10	3.00	1.9	0.51
madár	Kakas	11	2.95	2.2	0.60
madár	Daru	12	2.88	3.5	0.80
testvér	Szerzetes	14	2.78	4	1.08
daru	Eszköz	16	2.74	3.5	0.75
utazás	Autó	17	2.65	3.3	0.87
étel	Kakas	20	2.25	3.5	0.87
legény	Testvér	15	2.17	3.3	0.87
part	Domb	21	1.56	3	1.01
vízpart	Erdő	23	1.37	3	0.96
part	Erdő	25	1.29	3	0.88
erdő	sírkert	22	1.21	2.4	0.71

temető	Erdő	19	1.19	2.3	0.72
szerzetes	Jós	18	0.86	3	0.79
legény	Varázsló	26	0.82	3	0.80
üveg	Bűvész	28	0.80	3.6	0.95
szerzetes	Rabszolga	24	0.60	3	0.69
húr	Mosoly	27 (chord)	0.43	2	0.60
kakas	Utazás	29	0.31	1.7	0.49
fonal	Mosoly	27 (cord)	0.29	2.2	0.57
dél	Kötél	30	0.18	1.8	0.44

Table 2. Subjective word similarity results in the Hungarian experiment

The Hungarian similarity averages correlate with the English original and replica data in the following way (using Spearman rank correlation):

- correlation with Rubenstein and Goodenough (1965): $r=0.959$, $p<0.01$
- correlation with Miller and Charles (1991): $r=0.950$, $p<0.01$
- correlation with Resnik (1995): $r=0.932$, $p<0.01$

We can conclude that the correlation between the Hungarian and the English experiments is significant and very strong in all three cases. This level of similarity among languages, however, cannot be taken for granted as different languages exhibit different approaches to lexicalizing concepts, which introduces differences in lexical ambiguity and other areas.

3. Using the subjective similarity data for evaluating a computerized experiment

The similarity results discussed in the previous section have been collected to establish a relevant and reliable testing method for evaluating automated word similarity measurements. The first state-of-the-art Hungarian DS similarity test that relied on a similar evaluation method was Dobó és Csirik (2012). Note that they did not carry out the Hungarian subjective measurement (they used English similarity scores and the Hungarian equivalents of the original words, therefore the implementation of the evaluation test was not complete), but they put a wide selection of DS algorithms to the test which makes their work an important part of the literature on Distributed Semantics.

The implementation of the computerized part of my experiment is based on my previous research discussed in Tóth (2013). I have implemented custom-made computer software to build a large word co-occurrence matrix using a 100-million-word subcorpus of the lemmatized version of the Hungarian Webcorpus (Halácsy et al. 2004, Kornai et al. 2006). In the *rows* of this matrix, I have collected information ("context vectors") for each *target word* used in the Hungarian subjective similarity experiment described above. Each *column* has stored information proportional to the

number of co-occurrences of the given target word and a *context word*. There were as many columns as context words, 10,000 - 30,000 in this experiment (the most frequent words of the Hungarian Webcorpus). For collecting co-occurrence data, I have used a 1+1 word window around each occurrence of every target word. The elements of the matrix have stored pointwise mutual information instead of raw word frequency. The context vectors (i.e. the target words) have been compared using the cosine similarity measure (c.f. Tóth 2013).

In Distributed Semantics, the similarity of observed context is interpreted as similarity of meaning. For evaluating the corpus-based data, I have used the human subjective similarity scores and computed the correlation of the subjective and the objective (corpus-based) similarity results. Diagrams 1 and 2 show the corresponding Spearman rank correlation values. These diagrams also show how the number of context words (that I have used for characterizing the target words) influences the precision of the model. The correlations are significant ($p<0.05$ without lemmatization, $p<0.01$ with lemmatization), i.e. the subjective and objective data go together. The strongest correlation ($r=0.56$, $p=0.001$) has been measured when maximizing the number of context words in the presence of lemmatization.

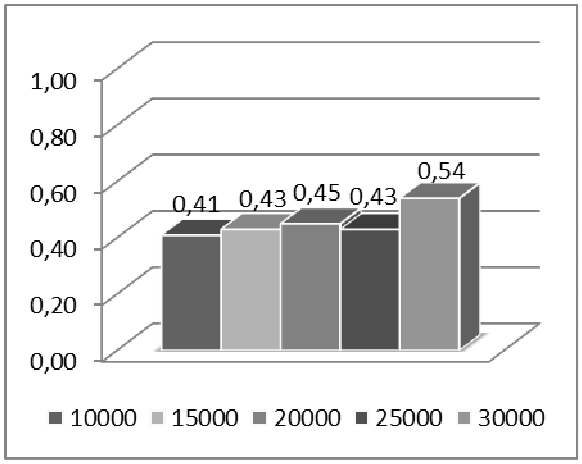

Diagram 1. Correlation – without lemmatization (x-axis: # of context words)

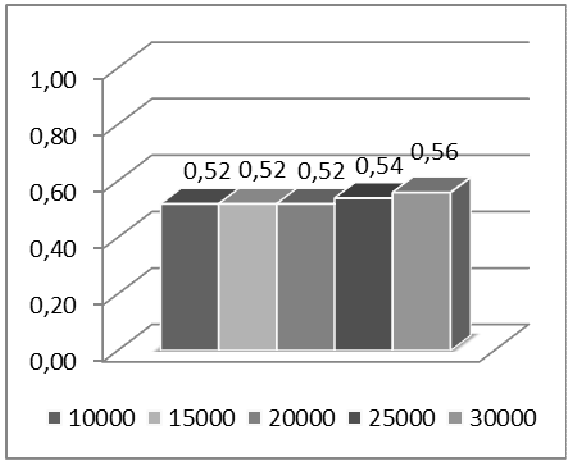

Diagram 2. Correlation – with lemmatization (x-axis: # of context words)

4. Conclusion

In the literature of computing similarity, the goal of carrying out a human subjective similarity test is to provide a benchmark set for evaluating word similarity data gathered automatically from corpora, lexical databases and/or other sources. In the Hungarian subjective word similarity experiment presented in section 3 we have seen that the participants' decisions have been very similar to each other and they have also correlated very strongly with the results of previous English experiments. I have also compared the human decisions with data collected form a 100-million-word Hungarian

corpus and found a significant moderate correlation (r=0.56, p=0.001) between them. “Perfect” results are not to be expected: in Resnik (1995), the upper limit of the human performance was a correlation of r=0.88. We have also highlighted some of the factors that make this task difficult even for humans, including lexical ambiguity. Nevertheless, careful parameter setting helps close the gap between human and machine performance in this semantic task.

Acknowledgements

This research was supported by the *European Union* and the *State of Hungary, co-financed by the European Social Fund* in the framework of TÁMOP-4.2.4.A/2-11/1-2012-0001 ‘National Excellence Program’.

Works Cited

Bullinaria, J. A. and J. P. Levy. “Extracting semantic representations from word co-occurrence statistics: A computational study.” *Behavior Research Methods* 39 (2007): 510-526.

Dobó A. and Csirik J. “Magyar és angol szavak szemantikai hasonlóságának automatikus kiszámítása.” *MSZNY 2013: IX. Magyar Számítógépes Nyelvészeti Konferencia*. Ed. Tanács A. and Vincze V. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2012. 213-224.

Halácsy P., Kornai A., Németh L., Rung A., Szakadát I. and Trón V. “Creating open language resources for Hungarian.” *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004)*, 2004.

Kornai A., Halácsy P., Nagy V., Oravecz Cs., Trón V. and Varga D. “Web-based frequency dictionaries for medium density languages.” *Proceedings of the 2nd International Workshop on Web as Corpus*. Ed. A. Kilgarriff and M. Baroni. 2006. 1-9.

Meng, L., R. Huang and J. Gu. “A Review of Semantic Similarity Measures in WordNet.” *International Journal of Hybrid Information Technology* 6(1), 2013. 1-12.

Miller, G. A. and W. G. Charles. “Contextual correlates of semantic similarity.” *Language and Cognitive Processes* 6(1), 1991. 1-28.

Resnik, P. “Using information content to evaluate semantic similarity.” *Proceedings of the 14th International Joint Conference on Artificial Intelligence.* San Francisco, CA: Morgan Kaufmann Publishers Inc., 1995. 448-453.

Rubenstein, H. and J. B. Goodenough. “Contextual correlates of synonymy.” CACM 8.10 (1965): 627-633.

Tóth Á. “The company that words keep.” *The Round Table: Partium Journal for English Studies* 3 (2013).

Turney, P. D. and P. Pantel. “From Frequency to Meaning: Vector Space Models of Semantics.” *Journal of Artificial Intelligence Research*, 37 (2010): 141-188.