Baltic J. Modern Computing, Vol. 10 (2022), No. 3, pp. 295–306
https://doi.org/10.22364/bjmc.2022.10.3.03

Evaluating Computational Models of Similarity against a Human Rated Dataset

Eduard BARBU[1], Verginica BARBU MITITELU[2]

[1] Institute of Computer Science
University of Tartu
Narva mnt 18, 51009 Tartu, Estonia
[2] Research Institute for Artificial Intelligence
Romanian Academy
13 Calea 13 Septembrie, 050711, Bucharest, Romania

eduard.barbu@ut.ee
vergi@racai.ro

Abstract. The first genuine human-rated similarity set for the Romanian language aligned with an English similarity dataset is presented. Two computational models automatically learn the similarity scores for the word pairs in the similarity set. The first model learns the similarity scores from language corpora. The second one assigns a similarity score based on the taxonomic structure of a semantic network. We studied what model captures human similarity scores best, the language influence on the perception of similarity, and the impact of parts of speech on similarity.

Keywords: similarity, computational models, word embeddings, wordnet

1 Introduction

The similarity relation is considered the basis of organization of objects into categories (Wertheimer, 1938). Three main theories of similarity are studied in Cognitive Psychology. In mental distance theories, such as Latent Semantic Analysis (Landauer and Dumais, 1997), the concepts are mapped onto vector spaces, and the similarity is computed as a distance in that vector space. The featural theories of similarity (Tversky, 1977) assume that the concepts can be described by their properties (called features), and the similarity is measured based on the common and distinct features. The structural theories (Gentner and Markman, 1997) improved on the featural theories, adding the constraint that the common and the distinct features are dependent. Although the study of similarity is prominent in Psychology, other disciplines are equally interested in

studying it: Artificial Intelligence, Computational Linguistics, Semantic Web, to name just a few.

A research line in these disciplines consists of collecting human-annotated sets and testing the computational theories of similarity against them. The most popular sets, at least in Computational Linguistics, are WordSim(WS)-353 (Finkelstein et al., 2001) and MEN (Bruni et al., 2012). Unfortunately, these sets do not distinguish between the relations of similarity and association. The association between two concepts is defined as the propensity of a subject to activate the second concept when the first concept is presented. In contrast, similarity has to do with the proximity of the mental representations of the concepts. Consider the concepts *cup* and *tea*: they are associated, but not similar: there is no perceptual principle to group together an object like a cup and a liquid like tea. Whereas the objects denoted by the concepts *apple* and *pear* are perceptually similar. To remedy this problem pervasive in the constructed similarity sets, the gold standard human similarity set called SimLex-999 containing genuine similarity scores was proposed (Hill et al., 2015).

Although similarity is a universal problem, most of the research is performed on the English language. This fact makes the similarity research partial. It is highly desirable that genuine similarity sets should be constructed for languages other than English. This way, essential questions regarding the relationship between a language and similarity judgment can be answered. Another desirable characteristic of the research is the cross-language comparison of results by building aligned similarity sets.

The Romanian language does not have a genuine similarity set. Although some computational models have been trained for this language, the evaluation was conducted with the WS-353 set (Hassan and Mihalcea, 2009). In this paper, the first genuine Romanian similarity set aligned with an English similarity set is presented.

The research questions addressed in this paper are:

(Q_1) What class of models better correlates with the human scores for the Romanian similarity set? Are these results in line with those obtained for other languages?

(Q_2) Is there a language effect when computing the correlations between the human scores and the computational models' score? We expect the models trained on Romanian resources to correlate better with the Romanian raters' scores than with the scores given by the English raters.

(Q_3) How does the variable part of speech (i.e., the fact that a word pair belongs to a specific part of speech) influence the correlation coefficient's strength?

According to our knowledge the SimLex-999 set was translated into German, Italian and Russian (Leviant and Reichart, 2015), as well as into Estonian (Kittask and Barbu, 2019).

The rest of the paper is organized as follows. The next section introduces the Romanian similarity set and shows how it is different from the original English set. Section 3 presents the distributional and the semantic network similarity models used to assign scores to the word pairs in the two similarity sets. Section 4 presents and discusses the results. The paper ends with the conclusions. The Data Availability Statement shows how one can access the Romanian similarity set and reproduce the experiments reported in this paper.

2 RoSimLex-999 similarity set

The Romanian dataset RoSimLex-999 is the translation of the SimLex-999 set (Hill et al., 2015). The latter contains 999 word pairs (666 noun pairs, 222 verb pairs, 111 adjective pairs) with associated human scores indicating the degree of similarity between the words in each pair. They are content words that have been shown to display various degrees of easiness of learning by children (Gentner, 2006) and are differently conceptualized by the human mind (Gentner, 1978).

We translated the set discussing and resolving the difficult cases. All adjective and verb pairs were translated, but 2 noun pairs could not be translated : *finger – toe* and *taxi – cab*. For the former pair, Romanian lacks specialized words. It uses two phrases made up of the hypernym *deget* (En. *digit*) followed by a modifier indicating the respective body part: *deget de la mână* (En. *digit from hand*) and *deget de la picior* (En. *digit from foot*), respectively. For the latter pair of words, the Romanian language has only one word to express the concept, and this is *taxi*. Hence, the RoSimLex-999 set contains 997 word pairs with the following distribution : 664 noun pairs, 222 verb pairs, and 111 adjective pairs.

The RoSimLex-999 set was presented to human raters for assigning similarity scores for each pair. The instructions for assigning scores were the same as for the original SimLex-999 set. The raters are adult native Romanian speakers. There were (N=53) subjects: 49 were sophomore students in philology at that moment, majoring in the Romanian language, and 4 are experienced linguists having Ph.D. degrees in linguistics and working as linguists in research institutions. The number of raters for RoSimLex-999 is slightly higher than the number of people who annotated the original SimLex-999 dataset (N=50).

The word pairs were uploaded in 53 spreadsheets, equal to the number of raters through the Google Spreadsheet web application. Each spreadsheet contained the instructions for assigning similarity scores and the word pairs to be rated. The Romanian pairs were grouped according to their part of speech and displayed starting with the adjectives and finishing with the verbs. All raters completed the task in one week. The scores assigned range from 0 to 6, with 0 representing the lack of similarity and 6 the highest similarity between words (e.g. synonymy)[3]. A quality assurance checking procedure was implemented as a Python script to ensure that the raters added a score for each word pair and that the score was in the interval [0,6]. The few pairs that lacked the score were sent back to raters to score.

The Spearman correlation coefficient between the human scores for SimLex-999 and RoSimLex-999 is 0.85. The correlation between the two languages' human scores is high, with the best correlation coefficient for the adjectives 0.88, 0.84 for nouns, and 0.83 for verbs.

3 Computational models of similarity

This section gives the theoretical underpinnings of two computational models of similarity and shows how we have trained the models. The first class of models is distri-

[3] The same scores were used in the SimLex-999 experiment.

butional and close to the idea of Latent Semantic Analysis. Large text corpora are used to assign vectors to words. The similarity is computed as the distance between these vectors. The second class of models makes use of semantic networks to calculate the similarity between two concepts.

3.1 Word embeddings

The word embeddings are the latest incarnation of the distributional hypothesis. The hypothesis was formulated in the '50s by the linguists Harris (1954) and Firth (1957). In Firth's formulation, it is stated as "a word is characterized by the company it keeps". Therefore, the distributional hypothesis says its linguistic contexts modulate the meaning of a word. For what concerns the modeling of the similarity, if two words appear in roughly the same context, then they should be very similar.

The word vectors known as word embeddings encode a word's meaning as a vector of numbers. A measure, usually the cosine similarity, is defined on the vector space such that when two words are close in meaning, the measure applied to the word's vectors will return a higher number than when the two words are dissimilar. Word2Vec is a distributional model (Mikolov et al., 2013) implemented as a two-layer neural network. When two words appear in similar contexts in a corpus, the network will output embedding vectors close in the embedding space. Word2Vec implements two architectures called Skip Gram and Continuous Bag of Words (CBOW). In the Skip Gram architecture, a neural network with a hidden layer is trained to output probabilities for all vocabulary words. The probabilities encode how likely it is to find each vocabulary word in a window size centered in the input word. For example, suppose the network receives the word *railway*. In that case, the output probabilities are going to be much higher for related terms like *station* and *ticket* than for unrelated words like *apple* or *dinosaurs*. The CBOW architecture predicts the word's probability in the middle of the sentence when the other words are known. Each architecture has its advantages. Skip Gram performs better when the training set is small and creates better embeddings for rare words, while CBOW is better at capturing the syntactic relations and representing the frequent words. During training, the network adjusts the neuron weights to learn to predict correctly based on positive examples. For an accurate prediction, negatives samples are also generated. The Word2Vec models discussed above learn continuous word representations that ignore the morphology of each word. An improvement of these models represents each word as a bag of n-characters (Bojanowski et al., 2017). Vectors are associated with each character. The word vectors are computed as the sum of the vectors assigned to each word's character.

3.2 Semantic Networks based models

The semantic similarity is also captured by the taxonomic structure of semantic networks. In particular, the meaning of the IS-A relation, the backbone of a taxonomy, involves inheritance of properties. Unlike distributional models, the semantic network structure spells why two concepts are similar. Nowadays, the most used semantic network in computational linguistics is the Princeton WordNet (PWN) (Miller et al., 1990;

Fellbaum, 1998). Its development started in 1985 under the coordination of the cognitive psychologist George Miller. PWN is organized around the notion of synset, which is a set of synonymous words having a definition (called gloss) associated (and sometimes also examples of usage). A word occurs in a number of synsets equal to the number of its senses. Semantic relations connect the synsets for each part of speech. Various similarity measures are defined on the taxonomic noun and verb hierarchies of PWN, but the most used ones are the following three :

1. Path Similarity (PS). Path-based similarity measures compute the shortest path between two concepts in a hierarchical semantic network. A shorter path between two concepts means that they are more similar. The equation for path-based similarity is given in 1, where dist(c1,c2) represents the taxonomic distance between the concepts c1 and c2.

$$sim_{path}(c_1, c_2) = \frac{1}{dist(c_1, c_2) + 1} \tag{1}$$

2. Leacock & Chodorow similarity (LC) (Leacock and Chodorow, 1998). It introduces non-uniform edge weighting measure, that uses logarithmic transformation to normalize the path length with the depth of the graph:

$$sim_{LC}(c_1, c_2) = -\log \frac{l(c_1, c_2)}{2 * depth} \tag{2}$$

   where depth is the length of the longest path from the root node to a leaf node. The length l(c_1,c_2) is measured in nodes attached to the concepts in the graph (Zesch and Gurevych, 2010).
3. Wu & Palmer similarity (WuP). The intuition behind this measure (Wu and Palmer, 1994) is that the concepts that are on the lower levels in the taxonomy are more similar than the concepts in the higher taxonomic levels even if the distance between them is the same (Jurafsky and Martin, 2009). WuP uses the lowest common subsumer (lcs) of two concepts defined as the first shared concept on the paths from the concepts to the root concept. WuP can be computed as in 3.

$$sim_{WuP}(c_1, c_2) = \frac{2 * lcs_{depth}}{l(c_1, lcs) + l(c_2, lcs) + 2 * lcs_{depth}} \tag{3}$$

3.3 Trained models

The Romanian language distributional models were trained on the CoRoLa corpus, Romanian Wikipedia, and the part of Common Crawl corpus[4] containing Romanian text. CoRoLa is the reference corpus for contemporary Romanian (Tufiș et al., 2019), containing 1.2 billion tokens. The semantic network used is the Romanian Wordnet (RoWN) (Tufiş et al., 2013). This is aligned with PWN at synset level.

The English distributional models were trained on Wikipedia and on the Common Crawl corpus containing English text. The semantic network similarity models are computed on PWN.

[4] Common Crawl is a corpus crawled from the web by the Common Crawl Foundation.

Because the wordnets record multiple senses for the words in the human similarity sets, we have used an automatic procedure to choose the most likely sense. The Cartesian product between the word senses in the semantic network corresponding to the words in the human similarity sets is generated. Subsequently, similarity scores are computed for each word sense pair in this set. The word sense pair that maximizes the similarity score is chosen.

We made a manual analysis of a randomly chosen set of pairs for each part of speech in order to evaluate the results of this semantic disambiguation task. Table 1 shows that the same number of noun and verb pairs was chosen for both languages. It can be noticed that for Romanian, the precision for nouns is higher than that for verbs, while for English, the opposite holds true. The difference in precision for the two parts of speech is smaller for English, which means the algorithm gave better results for this language.

Table 1. The precision of the automatic semantic disambiguation of the words in the similarity pairs.

	Nouns		Verbs	
	No. of pairs	% valid	No. of pairs	% valid
Ro	30	75	20	60
En	30	70	20	75

Each computational model of similarity assigns a similarity score to the word pairs in the human similarity sets. For the semantic network models the similarity scores are given by the three similarity metrics reported above. For the distributional models, this score is the cosine similarity between the vectors corresponding to the pair's words. The word embeddings we worked with are:

1. **CoRoLa_300_20** These word embeddings are created from the CoRoLa corpus: the vectors have 300 dimensions, and the minimum frequency of words is set to 20. It is the recommended[5] embedding configuration by the corpus creators.
2. **CoRoLa_400_5** These are the word embeddings (Păiș and Tufiș, 2018) created also from CoRoLa. They contain vectors with 400 dimensions and with the minimum frequency of words set at 5 occurrences. This configuration obtained the same results when the Spearman correlation coefficient is calculated with the WS-353 set (Finkelstein et al., 2001) as the recommended embedding model above.
3. **CoNLL_2017** The embeddings are trained in the CoNLL 2017 Shared Task [6]. A Word2Vec Continuous Skip Gram model is trained on datasets automatically selected with the aid of a language identifier from Common Crawl and Wikipedia. The model is trained with a vector size of 100 dimensions and a window size of 10.
4. **fastText**. These are the pre-trained word embeddings for English and Romanian performed by Facebook team. They are trained on Wikipedia and Common Crawl

[5] http://corolaws.racai.ro/word_embeddings/

[6] http://universaldependencies.org/conll17/data.html

using CBOW with position weights. The vectors have 300 dimensions. They were trained with character n-grams with a length of 5, a window of size 5, and 10 negatives.

4 Results

This section reports the Spearman correlation coefficient between the human estimated similarity and the computational models' similarity. Because not all the words in the human-rated set are found in the corpora or wordnets, we split this section into three subsections. Subsection 4.1 shows the results for all word pairs mapped in each computational model (*the maximal set*). In subsection 4.2 the *common set* is constructed. It includes those word pairs mapped in all computational models. In the last subsection we discuss the results.

4.1 Results for the maximal sets

Table 2 gives the Spearman correlation coefficient between the human similarity scores and the distributional similarity models. In all tables, "RO" stays for the scores assigned by the Romanian raters in the RoSimLex-999 set, and "EN" stays for the English raters' scores in the original SimLex-999 set. For example, in the CoRoLa_400_5 model, 966 word pairs were mapped ("Gb" stays for global). 111 of these are adjectives (A), 635 nouns (N), and 220 verbs (V). The best results are marked bold. The best overall correlation results and the best correlation results for nouns and adjectives are obtained for the fastText model and RoSimLex-999 set, whereas the CoNLL-2017 model gets the best results for verbs.

Table 2. The correlations between the human scores and distributional models for the maximal sets.

	Gb	N	A	V	Gb	N	A	V
Model	**CoRoLa_400_5**				**CoRoLa_300_20**			
#pairs	966	635	111	220	965	635	110	220
RO	.28	. 25	.39	.23	.24	.22	.32	.21
EN	.25	.27	.36	.13	.22	.24	.31	.13
Model	**fastText**				**CoNLL-2017**			
#pairs	967	636	111	220	966	635	111	220
RO	**.37**	**.42**	**.46**	.24	.26	.24	.36	**.25**
EN	.34	.40	.41	.15	.26	.25	.34	.22

Only nouns and verbs are organized hierarchically in wordnets; therefore, we report the similarity measures only for these parts of speech in Table 3, which shows the Spearman correlation coefficient between the RoSimLex-999 scores and SimLex-999 scores and the semantic network models trained on RoWN and PWN, respectively. The WuP model obtains the best overall results and the best results for noun pairs when the

correlation is computed with RoWN. The best results for verb pairs are achieved by the PS model. When the correlation is computed with PWN, the best noun results and the best verb results are obtained with two models: PS and LC. The latter model achieves the best overall results.

Table 3. The correlations between the human scores and semantic network models for the maximal sets.

	RoWN	**RO**	**EN**	**PWN**	**RO**	**EN**
Global	PS	.48	.49	PS	.44	.52
Nouns		.52	.53		.51	**.58**
Verbs		**.42**	.38		.34	**.38**
Global	LC	.47	.48	LC	.48	**.55**
Nouns		.52	.53		.51	**.58**
Verbs		.40	.37		.34	**.38**
Global	WuP	**.51**	.50	WuP	.45	.48
Nouns		**.53**	.54		.47	.55
Verbs		.38	.36		.36	.37

4.2 Results for the common set

The common set is the set of word pairs that map in all models. It contains 788 word pairs, 595 being noun pairs and 193 verb pairs. Table 4 gives the Spearman correlation coefficient between the human similarity scores and the distributional similarity models. The best overall results and the best results for nouns are obtained for the fastText model and RoSimLex-999 set, whereas the best results for verbs are obtained both for the fastText and the CoNLL-2017 models.

Table 4. The correlations between the human scores and distributional models for the common set.

	Gb	N	V	G	N	V
Model	**CoRoLa_400_5**			**CoRoLa_300_20**		
#pairs	788	595	193	788	595	193
RO	.26	.25	.23	.23	.21	.20
EN	.25	.28	.12	.22	.25	.12
Model	**fastText**			**CoNLL-2017**		
#pairs	788	595	193	788	595	193
RO	**.36**	**.42**	**.25**	.24	.23	**.25**
EN	.34	.41	.17	.25	.25	.22

Table 5 shows the Spearman correlation coefficient between the RoSimLex-999 scores (RO) and SimLex-999 scores (EN) and the semantic network models using

RoWN and PWN, respectively. The WuP model obtains the best overall results and the best results for verb pairs when the correlation is computed with the RoSimLex-999 set. The best results for noun pairs are achieved by the WuP model and SimLex-999 human scores. When the correlation is computed with the models trained on PWN, the best noun results are obtained with PS and LC models. The PS model also achieves the best overall results. WuP model gets the best results for verb pairs.

Table 5. Results for the common set.

	RoWN	**RO**	**EN**	**PWN**	**RO**	**EN**
Global	PS	.48	.49	PS	.45	.52
Nouns		.52	.53		.53	**.58**
Verbs		**.42**	.38		.34	.41
Global	LC	.47	.48	LC	.50	**.55**
Nouns		.52	.53		.53	**.58**
Verbs		.40	.37		.34	.41
Global	WuP	**.51**	.50	WuP	.48	.51
Nouns		.53	**.54**		.49	.55
Verbs		.38	.36		.38	**.42**

4.3 Discussion

The correlation coefficients between the similarity scores assigned by the distributional computational models trained on Romanian corpora and the human scores of RoSimLex-999 are better than the same correlation coefficients and the human scores in SimLex-999. The difference ranges from 2 to 5 correlation points. It means that there is a slight language effect on the perception of similarity.The same effect has been observed in (Kittask and Barbu, 2019) for Estonian language, however an opposite effect has been observed in (Leviant and Reichart, 2015). More research is needed for a reasoned answer about the language effect on similarity.

RoWN was built by translating the English synsets, and the semantic relations were imported from PWN. Therefore the differences between the respective correlation coefficients cannot be attributed to a language effect. The effect could be the result of the fact that not all English synsets could be translated into Romanian; thus, there are gaps in the noun taxonomy that are reflected in the computed similarity scores.

Regarding the magnitude of the Spearman correlation coefficient for the Romanian set, the fastText distributional model shows a moderate strength[7] of correlation with the RoSimLex-999 scores. The other distributional models show only a weak correlation with the RoSimLex-999 scores. Therefore, fastText distributional models do capture some of the similarities between the words. We think that the superior performance of this model can be explained by its being trained on Wikipedia, which contains the kind of knowledge one associates with semantic networks. Table 2 shows that the best

[7] In the literature, the strength is moderate if the coefficient is in the range .4-.59.

results are obtained for adjectives, irrespective of the model used. Unlike nouns and verbs whose semantics allow us to organize their meanings in hierarchies, adjectives are considered to occupy a multidimensional hyperspace, to have higher similarity in meaning with other adjectives (to such an extent that they can even share the same antonym) (Fellbaum et al., 1998). They can easily adjust their meanings to the context, i.e. to the modified nouns, which is suggestive of their possibility of collocating with nouns of more or less different meanings. The lowest performance of the models in the case of verbs may be correlated with their high polysemy: in PWN, verbs are the part of speech with the highest polysemy[8].

For the semantic network models, the best global Spearman correlation coefficient is obtained with PWN (55). This coefficient is higher than the best correlation coefficient for the fastText model, but still in the moderate range. Therefore, the best predictor of human similarity is derived from a manually built resource containing clearly defined semantic relations. The correlation coefficient is higher for nouns than for verbs. This fact is explained by the higher density of the noun semantic network.

Regarding the common set, the results are only slightly different from the maximal sets. The best performing distributional model is still fastText, though the results are somewhat lower than for the maximal set. It is also true that the semantic network trained models show better performance than the distributional models.

5 Conclusions

In this paper, we have presented the first genuine similarity set for the Romanian language. RoSimLex-999 is the translation of SimLex-999, thus allowing cross-lingual comparison. Each word pair in RoSimLex-999 has been rated by 53 people. Unlike other sets in the literature for which the annotators were recruited through Mechanical Turk or other services, our raters are students in linguistics or professional linguists. The Romanian similarity scores correlate very well with the English similarity scores. In the Introduction section, we have asked three research questions. At the end of this study, we can answer them. The first question regards the class of models that best correlate with the human scores and the consistency of other languages' results. The models trained on the semantic networks have a higher correlation with the human scores. Among the distributional models, the fastText models have the best results. We have conjectured that this is the case because it is trained on Wikipedia, an encyclopedia containing the kind of knowledge one finds in the semantic networks. In any case, the Spearman correlation coefficient for semantic networks is still in the medium range, meaning that these models do not fully capture the human notion of similarity. The distributional models encode syntactic features in addition to semantic ones. These results are in line with the findings of (Kittask and Barbu, 2019) for the Estonian language, where the semantic networks models also came first. The second question inquired if there is a language effect on similarity. The scores given by the Romanian raters correlate better with the those given by the distributional models trained on Romanian corpora. However more research is needed understand if the higher correlation scores

[8] https://wordnet.princeton.edu/documentation/wnstats7wn

can be attributed to language. In the semantic network's case, this comparison is not conclusive given that the RoWN translates PWN and imports its structure. An identical result for the distributional models and semantic network models have been obtained for the Estonian language. Unlike RoWN, the Estonian Wordnet has been developed without reference to PWN. The third question asked if the part of speech affects the correlation. The answer is positive. The best correlation for the distributional models is obtained for adjectives, then for nouns, and last for verbs. The original English study got the same result. In the Estonian research, the order induced by the correlation coefficient is different: nouns, verbs, and adjectives.

6 Data Availability

The results in this paper could be reproduced following the instructions in the GitHub repository https://github.com/SoimulPatriei/RoSimLex-999. In the repository it is shown:

- how to access the RoSimLex-999 data in spreadsheet and text format;
- how to reproduce the tables in the **Results** section in the paper;
- how to compute the similarities based on the corpora trained embeddings and semantic networks.

7 Acknowledgements

We would like to thank all 53 Romanian raters for helping us in this project.

References

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with subword information.

Bruni, E., Boleda, G., Baroni, M., Tran, N.-K. (2012). Distributional semantics in technicolor, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Jeju Island, Korea, pp. 136–145.
https://www.aclweb.org/anthology/P12-1015

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

Fellbaum, C., Gross, D., Miller, K. (1998). *Adjectives in WordNet*, Princeton University, pp. 26–39.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E. (2001). Placing search in context: The concept revisited, *Proceedings of the 10th international conference on World Wide Web*, ACM, pp. 406–414.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955, *Studies in Linguistic Analysis*, Philological Society, Oxford. reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

Gentner, D. (1978). On relational meaning: The acquisition of verb meaning, *Child Development* pp. 988–998.

Gentner, D. (2006). *Why verbs are hard to learn*, Oxford University Press, pp. 544–564.

Gentner, D., Markman, A. (1997). Structure mapping in analogy and similarity, *American Psychologist* **52**(1), 45–56.

Harris, Z. (1954). Distributional structure, *Word* **10**(2-3), 146–162.

Hassan, S., Mihalcea, R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, pp. 1192–1201.
https://www.aclweb.org/anthology/D09-1124

Hill, F., Reichart, R., Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation, *Computational Linguistics* .

Jurafsky, D., Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*, Pearson Prentice Hall, Upper Saddle River, N.J.

Kittask, C., Barbu, E. (2019). Is similarity visually grounded? computational model of similarity for the Estonian language, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, INCOMA Ltd., Varna, Bulgaria, pp. 541–549.
https://www.aclweb.org/anthology/R19-1064

Landauer, T. K., Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge, *Psychological Review* **104**, 211–240.

Leacock, C., Chodorow, M. (1998). *Combining Local Context and WordNet Similarity for Word Sense Identification*, Vol. 49, pp. 265–283.

Leviant, I., Reichart, R. (2015). Separated by an un-common language: Towards judgment language informed vector space modeling.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013). Distributed representations of words and phrases and their compositionality, *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, Curran Associates Inc., USA, pp. 3111–3119.
http://dl.acm.org/citation.cfm?id=2999792.2999959

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. (1990). Introduction to wordnet: An online lexical database, *International Journal of Lexicography* **4**, 235–244.

Păiș, V., Tufiș, D. (2018). Computing distributed representations of words using the corola corpus, *Proceedings of the Romanian Academy, Series A* (2), 403–409.

Tufiş, D., Barbu Mititelu, V., Ştefănescu, D., Ion, R. (2013). The romanian wordnet in a nutshell, *Language Resources and Evaluation* **47**(4), 1305–1314.

Tufiș, D., Barbu Mititelu, V., Irimia, E., Paiș, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M. (2019). Little strokes fell great oaks. creating CoRoLa, the reference corpus of contemporary romanian, *Revue roumaine de linguistique* pp. 227–240.

Tversky, A. (1977). Features of similarity, *Psychological Review* **84**(4), 327–352.

Wertheimer, M. (1938). Laws of organization in perceptual forms, *in* Ellis, W. (ed.), *A Source Book of Gestalt Psychology*, Routledge and Kegan Paul, London, pp. 71–88.

Wu, Z., Palmer, M. (1994). Verbs semantics and lexical selection, *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 133–138.
https://doi.org/10.3115/981732.981751

Zesch, T., Gurevych, I. (2010). Wisdom of crowds versus wisdom of linguists - measuring the semantic relatedness of words., *Nat. Lang. Eng.* **16**(1), 25–59.

Received August 1, 2022 , accepted August 6, 2022