Introducing two Vietnamese Datasets for Evaluating Semantic Models of (Dis-)Similarity and Relatedness

Kim Anh Nguyen and **Sabine Schulte im Walde** and **Ngoc Thang Vu**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B, 70569 Stuttgart, Germany
{nguyenkh,schulte,thangvu}@ims.uni-stuttgart.de

Abstract

We present two novel datasets for the low-resource language Vietnamese to assess models of semantic similarity: *ViCon* comprises pairs of synonyms and antonyms across word classes, thus offering data to distinguish between similarity and dissimilarity. *ViSim-400* provides degrees of similarity across five semantic relations, as rated by human judges. The two datasets are verified through standard co-occurrence and neural network models, showing results comparable to the respective English datasets.

1 Introduction

Computational models that distinguish between semantic similarity and semantic relatedness (Budanitsky and Hirst, 2006) are important for many NLP applications, such as the automatic generation of dictionaries, thesauri, and ontologies (Biemann, 2005; Cimiano et al., 2005; Li et al., 2006), and machine translation (He et al., 2008; Marton et al., 2009). In order to evaluate these models, gold standard resources with word pairs have to be collected (typically across semantic relations such as synonymy, hypernymy, antonymy, co-hyponymy, meronomy, etc.) and annotated for their degree of similarity via human judgements.

The most prominent examples of gold standard similarity resources for English are the Rubenstein & Goodenough (RG) dataset (Rubenstein and Goodenough, 1965), the TOEFL test questions (Landauer and Dumais, 1997), WordSim-353 (Finkelstein et al., 2001), MEN (Bruni et al., 2012), SimLex-999 (Hill et al., 2015), and the lexical contrast datasets by (Nguyen et al., 2016a, 2017). For other languages, resource examples are the translation of the RG dataset to German (Gurevych, 2005), the German dataset of paradigmatic relations (Scheible and Schulte im Walde,

2014), and the translation of WordSim-353 and SimLex-999 to German, Italian and Russian (Leviant and Reichart, 2015). However, for low-resource languages there is still a lack of such datasets, which we aim to fill for Vietnamese, a language without morphological marking such as case, gender, number, and tense, thus differing strongly from Western European languages.

We introduce two novel datasets for Vietnamese: a dataset of lexical contrast pairs **ViCon** to distinguish between similarity (synonymy) and dissimilarity (antonymy), and a dataset of semantic relation pairs **ViSim-400** to reflect the continuum between similarity and relatedness. The two datasets are publicly available.[1] Moreover, we verify our novel datasets through standard and neural co-occurrence models, in order to show that we obtain a similar behaviour as for the corresponding English datasets *SimLex-999* (Hill et al., 2015), and the lexical contrast dataset (henceforth *LexCon*), cf. Nguyen et al. (2016a).

2 Related Work

Over the years a number of datasets have been collected for studying and evaluating semantic similarity and semantic relatedness. For English, Rubenstein and Goodenough (1965) presented a small dataset (RG) of 65 noun pairs. For each pair, the degree of similarity in meaning was provided by 15 raters. The RG dataset is assumed to reflect similarity rather than relatedness. Finkelstein et al. (2001) created a set of 353 English noun-noun pairs (WordSim-353)[2], where each pair was rated by 16 subjects according to the degree of semantic relatedness on a scale from 0 to 10. Bruni et al. (2012) introduced a large test collec-

---

tion called MEN[3]. Similar to WordSim-353, the authors refer to both similarity and relatedness when describing the MEN dataset, although the annotators were asked to rate the pairs according to relatedness. Unlikely the construction of the RG and WordSim-353 datasets, each pair in the MEN dataset was only evaluated by one rater who ranked it for relatedness relative to 50 other pairs in the dataset. Recently, Hill et al. (2015) presented SimLex-999, a gold standard resource for the evaluation of semantic representations containing similarity ratings of word pairs across different part-of-speech categories and concreteness levels. The construction of SimLex-999 was motivated by two factors, (i) to consistently quantify similarity, as distinct from association, and apply it to various concept types, based on minimal intuitive instructions, and (ii) to have room for the improvement of state-of-the-art models which had reached or surpassed the human agreement ceiling on WordSim-353 and MEN, the most popular existing gold standards, as well as on RG. Scheible and Schulte im Walde (2014) presented a collection of semantically related word pairs for German and English,[4] which was compiled via Amazon Mechanical Turk (AMT)[5] human judgement experiments and comprises (i) a selection of targets across word classes balanced for semantic category, polysemy, and corpus frequency, (ii) a set of human-generated semantically related word pairs (synonyms, antonyms, hypernyms) based on the target units, and (iii) a subset of the generated word pairs rated for their relation strength, including positive and negative relation evidence.

For other languages, only a few gold standard sets with scored word pairs exist. Among others, Gurevych (2005) replicated Rubenstein and Goodenough's experiments after translating the original 65 word pairs into German. In later work, Gurevych (2006) used the same experimental setup to increase the number of word pairs to 350. Leviant and Reichart (2015) translated two prominent evaluation sets, WordSim-353 (association) and SimLex-999 (similarity) from English to Italian, German and Russian, and collected the scores for each dataset from the respective native speakers via crowdflower[6].

---

3 Dataset Design

3.1 Criteria

Semantic similarity is a narrower concept than *semantic relatedness* and holds between lexical terms with similar meanings. Strong similarity is typically observed for the lexical relations of synonymy and co-hyponymy. For example, in Vietnamese "đội" (*team*) and "nhóm" (*group*) represents a synonym pair; "ô_tô" (*car*) and "xe_đạp" (*bike*) is a co-hyponymy pair. More specifically, words in the pair "ô_tô" (*car*) and "xe_đạp" (*bike*) share several features such as physical (e.g. bánh_xe / *wheels*) and functional (e.g. vận_tải / *transport*), so that the two Vietnamese words are interchangeable regarding the kinds of transportation. The concept of semantic relatedness is broader and holds for relations such as meronymy, antonymy, functional association, and other "non-classical relations" (Morris and Hirst, 2004). For example, "ô_tô" (*car*) and "xăng_dầu" (*petrol*) represent a meronym pair. In contrast to similarity, this meronym pair expresses a clearly functional relationship; the words are strongly associated with each other but not similar.

Empirical studies have shown that the predictions of distributional models as well as humans are strongly related to the part-of-speech (POS) category of the learned concepts. Among others, Gentner (2006) showed that verb concepts are harder to learn by children than noun concepts.

Distinguishing antonymy from synonymy is one of the most difficult challenges. While antonymy represents words which are strongly associated but highly dissimilar to each other, synonymy refers to words that are highly similar in meaning. However, antonyms and synonyms often occur in similar context, as they are interchangeable in their substitution.

3.2 Resource for Concept Choice: Vietnamese Computational Lexicon

The Vietnamese Computational Lexicon (VCL)[7] (Nguyen et al., 2006) is a common linguistic database which is freely and easily exploitable for automatic processing of the Vietnamese language. VCL contains 35,000 words corresponding to 41,700 concepts, accompanied by morphological, syntactic and semantic information. The morphological information consists of 8 morphemes

---

including simple word, compound word, reduplicative word, multi-word expression, loan word, abbreviation, bound morpheme, and symbol. For example, "bàn" (*table*) is a simple word with definition "đồ thường làm bằng gỗ, có mặt phẳng và chân đỡ . . ." (*pieces of wood, flat and supported by one or more legs . . .*). The syntactic information describes part-of-speech, collocations, and subcategorisation frames. The semantic information includes two types of constraints: logical and semantic. The logical constraint provides category meaning, synonyms and antonyms. The semantic constraint provides argument information and semantic roles. For example, "yêu" (*love*) is a verb with category meaning "*emotion*" and antonym "ghét" (*hate*).

VCL is the largest linguistic database of its kind for Vietnamese, and it encodes various types of morphological, syntactic and semantic information, so it presents a suitable starting point for the choice of lexical units for our purpose.

3.3 Choice of Concepts

3.3.1 Concepts in ViCon

The choice of related pairs in this dataset was drawn from VCL in the following way. We extracted all antonym and synonym pairs according to the three part-of-speech categories: noun, verb and adjective. We then randomly selected 600 adjective pairs (300 antonymous pairs and 300 synonymous pairs), 400 noun pairs (200 antonymous pairs and 200 synonymous pairs), and 400 verb pairs (200 antonymous pairs and 200 synonymous pairs). In each part-of-speech category, we balanced for the size of morphological classes in VCL, for both antonymous and synonymous pairs.

3.3.2 Concepts in ViSim-400

The choice of related pairs in this dataset was drawn from both the VLC and the *Vietnamese WordNet*[8] (VWN), cf. Nguyen et al. (2016b). We extracted all pairs of the three part-of-speech categories: noun, verb and adjective, according to five semantic relations: synonymy, antonymy, hypernymy, co-hoponymy and meronymy. We then sampled 400 pairs for the ViSim-400 dataset, accounting for 200 noun pairs, 150 verb pairs and 50 adjective pairs. Regarding noun pairs, we balanced the size of pairs in terms of six relations: the five extracted relations from VCL and VWN,

and an "unrelated" relation. For verb pairs, we balanced the number of pairs according to five relations: synonymy, antonymy, hypernymy, co-hyponymy, and unrelated. For adjective pairs, we balanced the size of pairs for three relations: synonymy, antonymy, and unrelated. In order to select the unrelated pairs for each part-of-speech category, we paired the unrelated words from the selected related pairs at random. From these random pairs, we excluded those pairs that appeared in VCL and VWN. Furthermore, we also balanced the number of selected pairs according to the sizes of the morphological classes and the lexical categories.

3.4 Annotation of ViSim-400

For rating ViSim-400, 200 raters who were native Vietnamese speakers were paid to rate the degrees of similarity for all 400 pairs. Each rater was asked to rate 30 pairs on a 0–6 scale; and each pair was rated by 15 raters. Unlike other datasets which performed the annotation via Amazon Mechanical Turk, each rater for ViSim-400 conducted the annotation via a survey which detailed the exact annotation guidelines.

The structure of the questionnaire was motivated by the SimLex-999 dataset: we outlined the notion of similarity via the well-understood idea of the six relations included in the ViSim-400 dataset. Immediately after the guidelines of the questionnaire, a checkpoint question was posed to the participants to test whether the person understood the guidelines: the participant was asked to pick the most similar word pair from three given word pairs, such as kiêu_căng/kiêu_ngạo (*arrogant/cocky*) vs. trầm/bổng (*high/low*) vs. cổ_điển/biếng (*classical/lazy*). The annotators then labeled the kind of relation and scored the degree of similarity for each word pair in the survey.

3.5 Agreement in ViSim-400

We analysed the ratings of the ViSim-400 annotators with two different inter-annotator agreement (IAA) measures, Krippendorff's alpha coefficient (Krippendorff, 2004), and the average standard deviation (STD) of all pairs across word classes. The first IAA measure, IAA-pairwise, computes the average pairwise Spearman's ρ correlation between any two raters. This IAA measure has been a common choice in previous data collections in distributional semantics (Padó et al., 2007; Reisinger and Mooney, 2010; Hill et al., 2015).
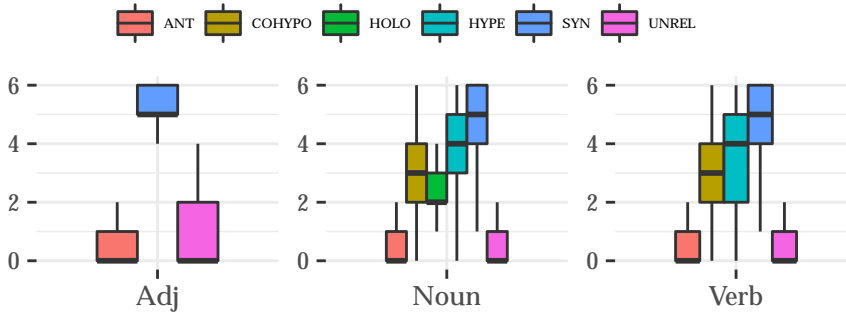
---

[8] http://viet.wordnet.vn/wnms/

Figure 1: Distribution of scored pairs in ViSim-400 across parts-of-speech and semantic relations.

	All	**Noun**	**Verb**	**Adjective**
IAA-Mean ρ	0.86	0.86	0.86	0.78
IAA-Pairwise ρ	0.79	0.76	0.78	0.75
Krippendorff's α	0.78	0.76	0.78	0.86
STD	0.87	0.87	0.90	0.82

Table 1: Inter-annotator agreements measured by Spearman's ρ, Krippendorff's α, and the average standard deviation (STD) of all pairs across word classes.

The second IAA measure, IAA-mean, compares the average correlation of the human raters with the average of all other raters. This measure would smooth individual annotator effects, and serve as a more appropriate *"upper bound"* for the performance of automatic systems than IAA-pairwise (Vulić et al., 2017). Finally, Krippendorff's α coefficient reflects the disagreement of annotators rather than their agreement, in addition to correcting for agreement by chance.

Table 1 shows the inter-annotator agreement values, Krippendorff's α coefficient, and the response consistency measured by STD over all pairs and different word classes in ViSim-400. The overall IAA-pairwise of ViSim-400 is $\rho = 0.79$, comparing favourably with the agreement on the SimLex-999 dataset ($\rho = 0.67$ using the same IAA-pairwise measure). Regarding IAA-mean, ViSim-400 also achieves an overall agreement of $\rho = 0.86$, which is similar to the agreement in Vulić et al. (2017), $\rho = 0.86$. For Krippendorff's α coefficient, the value achieves $\alpha = 0.78$, also reflecting the reliability of the annotated dataset.

Furthermore, the box plots in Figure 1 present the distributions of all rated pairs in terms of the fine-grained semantic relations across word classes. They reveal that –across word classes– synonym pairs are clearly rated as the most similar words, and antonym as well as unrelated pairs

are clearly rated as the most dissimilar words. Hypernymy, co-hyponymy and holonymy are in between, but rather similar than dissimilar.

4 Verification of Datasets

In this section, we verify our novel datasets Vi-Con and ViSim-400 through standard and neural co-occurrence models, in order to show that we obtain a similar behaviour as for the corresponding English datasets.

4.1 Verification of ViSim-400

We adopt a comparison of neural models on SimLex-999 as suggested by Nguyen et al. (2016a). They applied three models, a Skip-gram model with negative sampling SGNS (Mikolov et al., 2013), the dLCE model (Nguyen et al., 2016a), and the mLCM model (Pham et al., 2015). Both the dLCE and the mLCM models integrated lexical contrast information into the basic Skip-gram model to train word embeddings for distinguishing antonyms from synonyms, and for reflecting degrees of similarity.

The three models were trained with 300 dimensions, a window size of 5 words, and 10 negative samples. Regarding the corpora, we relied on Vietnamese corpora with a total of ≈145 million tokens, including the *Vietnamese Wikipedia*,[9] *VNESEcorpus* and *VNTQcorpus*,[10] and the *Leipzig Corpora Collection for Vietnamese*[11] (Goldhahn et al., 2012). For word segmentation and POS tagging, we used the open-source toolkit *UETnlp*[12] (Nguyen and Le, 2016). The antonym and synonym pairs to train the

---

dLCE and mLCM models were extracted from VWN consisting of 49,458 antonymous pairs and 338,714 synonymous pairs. All pairs which appeared in ViSim-400 were excluded from this set.

Table 2 shows Spearman's correlations ρ, comparing the scores of the three models with the human judgements for ViSim-400. As also reported for English, the dLCE model produces the best performance, SGNS the worst.

	SGNS	mLCM	dLCE
ViSim-400	0.37	0.60	**0.62**
SimLex-999	0.38	0.51	**0.59**

Table 2: Spearman's correlation ρ on ViSim-400 in comparison to SimLex-999, cf. Nguyen et al. (2016a).

In a second experiment, we computed the cosine similarities between all word pairs, and used the area under curve (AUC) to distinguish between antonyms and synonyms. Table 3 presents the AUC results of the three models. Again, the models show a similar behaviour in comparison to SimLex-999, where also the dLCE model outperforms the two other models, and the SGNS model is by far the worst.

	Model	Noun	Verb	Adj
ViSim-400	SGNS	0.66	0.63	0.70
	mLCM	0.81	0.92	0.96
	dLCE	**0.92**	**0.95**	**0.98**
SimLex-999	SGNS	0.66	0.65	0.64
	mLCM	0.69	0.71	0.85
	dLCE	**0.72**	**0.81**	**0.90**

Table 3: AUC scores for distinguishing antonyms from synonyms in ViSim-400.

4.2 Verification of ViCon

In order to verify ViCon, we applied three co-occurrence models to rank antonymous and synonymous word pairs according to their cosine similarities: two standard co-occurrence models based on positive point-wise mutual information (PPMI) and positive local mutual information (PLMI) (Evert, 2005) as well as an improved feature value representation $\mathbf{weight^{SA}}$ as suggested by Nguyen et al. (2016a). For building the vector space co-occurrence models, we relied on the same Vietnamese corpora as in the previous section. For inducing the word vector representations via $\mathbf{weight^{SA}}$, we made use of the antonymous and synonymous pairs in VWN, as in the previ-

ous section, and then removed all pairs which appeared in ViCon. Optionally, we applied singular value decomposition (SVD) to reduce the dimensionalities of the word vector representations.

As in Nguyen et al. (2016a), we computed the cosine similarities between all word pairs, and then sorted the pairs according to their cosine scores. Average Precision (AP) evaluated the three vector space models. Table 4 presents the results of the three vector space models with and without SVD. As for English, the results on the Vietnamese dataset demonstrate significant improvements (χ^2,$^*p < .001$) of $\mathbf{weight^{SA}}$ over PPMI and PLMI, both with and without SVD, and across word classes.

	Metric	ADJ		NOUN		VERB	
		SYN	ANT	SYN	ANT	SYN	ANT
ViCon	PPMI	0.70	0.38	0.68	0.39	0.69	0.38
	PLMI	0.59	0.44	0.61	0.42	0.63	0.41
	weight^{SA}	**0.93***	**0.31***	**0.94***	**0.31**	**0.96**	**0.31**
	PPMI + SVD	0.76	0.36	0.66	0.40	0.81	0.34
	PLMI + SVD	0.49	0.51	0.55	0.46	0.51	0.49
	weight^{SA} + SVD	**0.91***	**0.32***	**0.81***	**0.34***	**0.92***	**0.32***
LexCon	PLMI	0.56	0.46	0.60	0.42	0.62	0.42
	weight^{SA}	**0.75**	**0.36**	**0.66**	**0.40**	**0.71**	**0.38**
	PLMI + SVD	0.55	0.46	0.55	0.46	0.58	0.44
	weight^{SA} + SVD	**0.76***	**0.36***	**0.66**	**0.40**	**0.70***	**0.38***

Table 4: AP evaluation of co-occurrence models on ViCon in comparison to LexCon (Nguyen et al., 2016a).

5 Conclusion

This paper introduced two novel datasets for the low-resource language Vietnamese to assess models of semantic similarity: *ViCon* comprises synonym and antonym pairs across the word classes of nouns, verbs, and adjectives. It offers data to distinguish between similarity and dissimilarity. *ViSim-400* contains 400 word pairs across the three word classes and five semantic relations. Each pair was rated by human judges for its degree of similarity, to reflect the continuum between similarity and relatedness. The two datasets were verified through standard co-occurrence and neural network models, showing results comparable to the respective English datasets.

References

Chris Biemann. 2005. Ontology learning from text: A survey of methods. *LDV Forum* 20(2):75–93.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Jeju Island, Korea, pages 136–145.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics* 32(1):13–47.

Philipp Cimiano, Andreas Hotho, and Steffen Staab. 2005. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research* 24(1):305–339.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences*. Ph.D. thesis, University of Stuttgart.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*. Hong Kong, Hong Kong, pages 406–414.

Dedre Gentner. 2006. Why verbs are hard to learn. In Kathryn A. Hirsh-Pasek and Roberta M. Golinkoff, editors, *Action meets word: How Children Learn Verbs*, Oxford University Press, pages 544–564.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*. pages 759–765.

Iryna Gurevych. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*. Jeju Island, Republic of Korea, pages 767–778.

Iryna Gurevych. 2006. Thinking beyond the nouns: Computing semantic relatedness across parts of speech. In *Proceedings of Sprachdokumentation & Sprachbeschreibung, 28. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*. Bielefeld, Germany, page 226.

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. Indirect-HMM-based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii, pages 98–107.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with genuine similarity estimation. *Computational Linguistic* 41(4):665–695.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to its Methodology*. Sage Publications.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104(2):211–240.

Ira Leviant and Roi Reichart. 2015. Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR* abs/1508.00106.

Mu Li, Yang Zhang, Muhua Zhu, and Ming Zhou. 2006. Exploring distributional similarity-based models for query spelling correction. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, pages 1025–1032.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, pages 381–390.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*. Lake Tahoe, Nevada, pages 3111–3119.

Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*. Boston, Massachusetts, pages 46–51.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Distinguishing antonyms and synonyms in a pattern-based neural network. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain, pages 76–85.

Kim-Anh Nguyen, Sabine Schulte im Walde, and Thang Vu. 2016a. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, pages 454–459.

Phuong-Thai Nguyen, Van-Lam Pham, Hoang-An Nguyen, Huy-Hien Vu, Ngoc-Anh Tran, and Thi-Thu-Ha Truong. 2016b. A two-phase approach for building a Vietnamese WordNet. In *Proceedings of the 8th Global WordNet Conference*. Bucharest, Romania, pages 259–264.

Thi Minh Huyen Nguyen, Laurent Romary, Mathias Rossignol, and Xuan Luong Vu. 2006. A lexicon for Vietnamese language processing. *Language Resources and Evaluation* 40(3-4):291–309.

Tuan-Phong Nguyen and Anh-Cuong Le. 2016. A hybrid approach to Vietnamese word segmentation. In *Proceedings of the International Conference on Computing Communication Technologies, Research, Innovation, and Vision for the Future*. Hanoi, Vietnam, pages 114–119.

Sebastian Padó, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of the joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic.

Nghia The Pham, Angeliki Lazaridou, and Marco Baroni. 2015. A multitask objective to inject lexical contrast into distributional semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, pages 21–26.

Joseph Reisinger and Raymond Mooney. 2010. A mixture model with sharing for lexical semantics. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, Massachusetts, pages 1173–1182.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8(10):627–633.

Silke Scheible and Sabine Schulte im Walde. 2014. A database of paradigmatic semantic relation pairs for German nouns, verbs and adjectives. In *Proceedings of the COLING Workshop Lexical and Grammatical Resources for Language Processing*. Dublin, Ireland, pages 111–119.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics* 43(4):781–835.