More than just Frequency?
Demasking Unsupervised Hypernymy Prediction Methods

Thomas Bott, Dominik Schlechtweg and Sabine Schulte im Walde
Institute for Natural Language Processing, University of Stuttgart, Germany
{bottts,schlecdk,schulte}@ims.uni-stuttgart.de

Abstract

This paper presents a comparison of unsupervised methods of hypernymy prediction (i.e., to predict which word in a pair of words such as *fish–cod* is the hypernym and which the hyponym). Most importantly, we demonstrate across datasets for English and for German that the predictions of three methods (Weeds-Prec, invCL, SLQS Row) strongly overlap and are highly correlated with frequency-based predictions. In contrast, the second-order method SLQS shows an overall lower accuracy but makes correct predictions where the others go wrong. Our study once more confirms the general need to check the frequency bias of a computational method in order to identify frequency-(un)related effects.

1 Introduction

Hypernymy represents a major paradigmatic semantic relation between two concepts, a hypernym (superordinate) and a hyponym (subordinate), as in *tree–oak* and *fish–cod*, where the hyponym implies the hypernym, but not vice versa. From a cognitive perspective hypernymy is central to the organisation of the mental lexicon (Deese, 1965; Miller and Fellbaum, 1991; Murphy, 2003), next to further semantic relations such as synonymy, antonymy, etc. From a computational perspective hypernymy is central to solving a number of Natural Language Processing (NLP) tasks such as taxonomy creation (Hearst, 1998; Cimiano et al., 2004; Snow et al., 2006; Navigli and Ponzetto, 2012), textual entailment (Dagan et al., 2006; Clark et al., 2007) and text generation (Biran and McKeown, 2013).

Accordingly, the field has witnessed active research on two subtasks involved in computational models of hypernymy (see Shwartz et al. (2017) for an extensive overview): *hypernymy detection* (i.e., distinguishing hypernymy from other semantic relations) and *hypernymy prediction* (i.e., determining which word in a pair of words is the hypernym and which is the hyponym). The target subtask of the current study is hypernymy prediction: we perform a comparative analysis of a class of approaches commonly refered to as *unsupervised hypernymy methods* (Weeds et al., 2004; Kotlerman et al., 2010; Clarke, 2012; Lenci and Benotto, 2012; Santus et al., 2014). These methods all rely on the distributional hypothesis (Harris, 1954; Firth, 1957) that words which are similar in meaning also occur in similar linguistic distributions. In this vein, they exploit asymmetries in distributional vector space representations, in order to contrast hypernym and hyponym vectors.

While these unsupervised hypernymy prediction methods have been explored and compared extensively on a number of benchmark datasets (Shwartz et al., 2017), this study takes a novel perspective and performs a detailed analysis of whether and where the methods make similar or different decisions. Our prediction experiments on simplex and complex nouns in English and German WordNets and evaluation benchmarks show that most of the methods we investigate overlap in their specific predictions to a surprisingly high degree, and that the predictions strongly correlate with those based on raw frequencies. Our study therefore emphasises the general need to check the frequency bias of a computational method and to distinguish between frequency-related and frequency-unrelated effects.

2 Data and Methods

In the following we describe our gold standard datasets (Section 2.1), our corpora and vector spaces (Section 2.2) and our hypernymy prediction methods (Section 2.3). The code and links to the gold standards are available from https://github.com/Thommy96/hyp-freq-comp.

186

2.1 Gold Standard Datasets

Our study focuses on hypernymy between nouns and uses two types of gold standard resources for hypernymy relations. On the one hand, we rely on WordNets as classical large-scale taxonomies where hypernymy represents one of the core semantic relations for organisation: the English *WordNet*[1] (Miller et al., 1990; Fellbaum, 1998), version 3, and the German *GermaNet*[2] (Hamp and Feldweg, 1997; Kunze and Wagner, 1999; Lemnitzer and Kunze, 2007), version 11. From both WordNets, we extracted all noun–noun pairs with a hypernymy relation and removed duplicates, autohyponyms and space-separated multiword expressions. We also distinguish between compounds (which frequently represent hyponyms of their constituent heads, as in *dog–lapdog*) and non-compounds by applying a simple heuristic, i.e., categorising all hypernym–hyponym pairs as compounds if one is a substring of the other. We expected this subset to exhibit idiosynchratic behaviour in our prediction experiments.

On the other hand, we rely on a number of benchmark datasets for hypernymy evaluation: *BLESS* (Baroni and Lenci, 2011) provides related concepts for 200 English concrete nouns connected through a semantic relation (hypernymy, co-hyponymy, meronymy, attribute, event) or a null-relation. The dataset by Lenci and Benotto (2012) contains a subset of BLESS relation pairs, as created for previous comparisons of hypernymy detection methods. A dataset similar to BLESS, *EVALution*, was induced from ConceptNet and WordNet (Santus et al., 2015). Its semantic relations include hypernymy, synonymy, antonymy and meronymy. For quality reasons, the pairs were filtered by automatic methods and crowd-sourcing to improve consistency and to determine prototypical pairs. Finally, we use the Weeds dataset (Weeds et al., 2004; Weeds and Weir, 2005) which contains word pairs related by hypernymy and co-hyponymy across word classes. From all four benchmark datasets we extracted all noun–noun pairs related by hypernymy.

The first row in Table 1 shows the numbers of hypernymy pairs in the WordNets and in the benchmark datasets.

2.2 Corpora and Vector Spaces

We created our distributional vector spaces based on the *WaCky*[3] corpora (Baroni et al., 2009) for English and for German. The English *PukWaC* corpus is the syntax-annotated version of *ukWaC* (Ferraresi et al., 2008) and contains ≈1.9 billion words; the German *SdeWaC* corpus (Faaß and Eckart, 2013) is a cleaned version of the *WaCky* corpus *deWaC* and contains ≈880 million words; both corpora are pos-tagged with the *TreeTagger* (Schmid, 1994).

For each corpus we created a traditional count vector space[4] based on a co-occurrence window of ± 10 words within sentences (because sentences in the SdeWaC are shuffled, so going beyond sentence border is meaningless). We used a bag-of-words approach only taking into account lemmatised nouns, verbs and adjectives.

2.3 Hypernymy Methods and Baselines

We selected four unsupervised hypernymy methods and defined two baselines. The methods were chosen from different families with regard to how they exploit the distributional hypothesis for hypernymy detection: *WeedsPrec* and *InvCL* rely on the *Distributional Inclusion Hypothesis*, according to which a significant number of distributional features of a word x is included in the distributional features of a word y, if x is semantically more specific than y. *SLQS Row* and *SLQS Sec*[5] rely on the *Distributional Informativeness Hypothesis* using first- and second-order variants of word entropy, respectively. The methods are defined as follows regarding the distributional features f in the two word vectors $\vec{x}$ and $\vec{y}$ for a word pair $\langle x, y \rangle$.

WeedsPrec: An asymmetric precision method suggested by Weeds et al. (2004) that quantifies the weighted inclusion of the features of word x in the features of word y. If $WeedsPrec(x, y) > WeedsPrec(y, x)$, then x is predicted as the hyponym and y as the hypernym, and vice versa.

$$WeedsPrec(x,y) = \frac{\sum_{f \in (\vec{x} \cap \vec{y})} x_f}{\sum_{f \in \vec{x}} x_f}$$

	WordNet		GermaNet		BLESS	EVALution	LB	Weeds
	¬comp	comp	¬comp	comp				
sizes:	106,397	3,366	102,714	35,963	1,337	606	1,747	1,117
Word Length	47.26	**94.65**	56.14	**99.41**	23.19	34.86	52.42	44.76
Word Frequency	**73.19**	92.81	73.66	98.78	62.30	**68.96**	76.48	**76.63**
WeedsPrec	72.22	92.93	74.01	98.87	57.52	64.22	**77.02**	74.22
InvCL	72.97	92.84	73.92	98.78	63.05	68.86	76.48	76.45
SLQS Row	71.82	93.02	**74.40**	98.79	58.56	55.91	71.27	72.43
SLQS Sec	65.05	74.66	70.38	90.15	**71.80**	59.63	62.66	65.71

Table 1: Sizes of datasets and overall prediction results across datasets.

InvCL: An asymmetric precision method suggested by Lenci and Benotto (2012) that takes both feature inclusion as well as feature non-inclusion (originally suggested as *ClarkDE (cde)* by Clarke (2012)) into account. If $invCL(x,y) > invCL(y,x)$, then x is predicted as the hyponym and y as the hypernym, and vice versa.

$$cde(x,y) = \frac{\sum_{f \in (\vec{x} \cap \vec{y})} min(x_f, y_f)}{\sum_{f \in \vec{x}} x_f}$$

$$invCL(x,y) = \sqrt{cde(x,y) \cdot (1 - cde(y,x))}$$

SLQS Row: An asymmetric method suggested by Shwartz et al. (2016) which relies on the word entropy $H(w)$ for a word w, taking all context words as features into account: w_f. If $SLQS_{Row}(x,y) > 0$, then x is predicted as the hyponym and y as the hypernym, and vice versa.

$$SLQS_{Row}(x,y) = 1 - \frac{H(x)}{H(y)}$$

$$H(w) = -\sum_{w_f} p(w_f|w) \cdot log_2(p(w_f|w))$$

SLQS Sec: An asymmetric method suggested by Santus et al. (2014) which relies on second-order word entropy $E(w)$ and is calculated as the median entropy Med of a word's most strongly associated context words w_f. We use the 50 strongest contexts in our vector spaces, as determined by weighted co-occurrence scores using *positive local mutual information* (Evert, 2005). If $SLQS_{Sec}(x,y) > 0$, then x is predicted as the hyponym and y as the hypernym, and vice versa.

$$SLQS_{Sec}(x,y) = 1 - \frac{E(x)}{E(y)}$$

$$E(w) = Med_{w_f} H(w_f)$$

Baselines: In comparison to the hypernymy methods we applied two baselines, cf. Zipf's principles of least effort (Zipf, 1949):

- ***Word Length***: Given that hyponyms refer to more specific concepts than their hypernyms, and assuming that more specific concepts tend to have a longer word length, this baseline predicts the longer word in a word pair (as measured by the number of characters) as the hyponym.

- ***Word Frequency***: Given that hyponyms refer to more specific concepts than their hypernyms and assuming that more specific concepts appear less often in a corpus, this baseline predicts the less frequent word in a word pair (as measured by corpus frequency) as the hyponym.

3 Prediction Results and Comparisons

3.1 Overall results

Table 1 shows the overall accuracy results of the predictions across methods and datasets (best results in bold fonts). Accuracy is defined by the proportion of correct predictions given that we know which word in a word pair is the hypernym and which is the hyponym.

For each WordNet we list two results, one for the non-compound pairs (in blue, as the benchmark results) and one for the compound pairs (in grey). For compound pairs word length is an almost perfect predictor,[6] as expected, and all unsupervised methods are also above 90%, with SLQS Sec as an exception. In all other columns we can see that word length is generally a poor baseline. Word frequency, however, is a very powerful baseline; across datasets it keeps up or even outperforms the respective best methods, which are SLQS Sec

---

[6]The prediction does not reach 100% because our heuristic included non-compound pairs, such as *selection–election*.

	Length	Freq	WeedsPrec	InvCL	SLQS Row
Freq	0.592				
WeedsPrec	0.604	0.947			
InvCL	0.592	0.992	0.951		
SLQS Row	0.599	0.914	0.938	0.912	
SLQS Sec	0.547	0.715	0.719	0.714	0.727

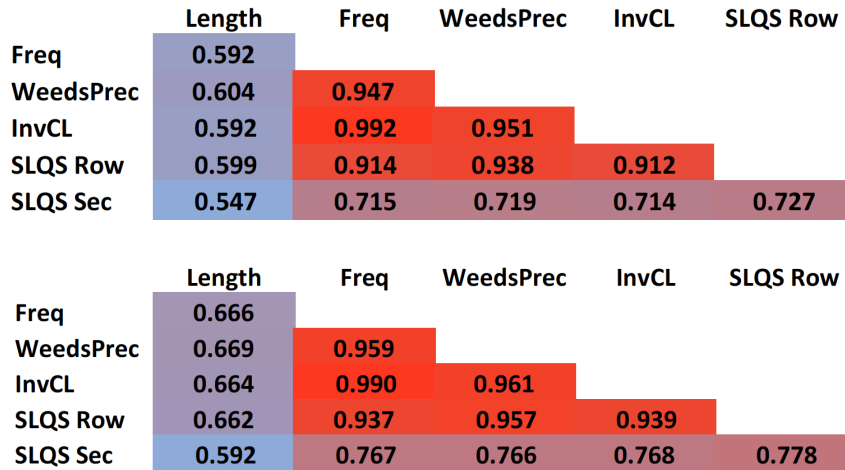|	Length	Freq	WeedsPrec	InvCL	SLQS Row
Freq	0.666				
WeedsPrec	0.669	0.959			
InvCL	0.664	0.990	0.961		
SLQS Row	0.662	0.937	0.957	0.939	
SLQS Sec	0.592	0.767	0.766	0.768	0.778

Figure 1: SMC correlations between methods for WordNet (above) and GermaNet (below) non-compound pairs.

on BLESS; InvCL on EVALution and Weeds; and WeedsPrec on LB. Across datasets, the best results vary between 68.96% and 77.02% for non-compounds; compounds obviously represent "easy" cases of hypernymy.

3.2 Correlations between predictions

To explore similarities in predictions across methods, we applied the *Simple Matching Coefficient (SMC)* (Sokal, 1958) to determine for each two methods to which degree their decisions overlap, by comparing the number of matching decisions (i.e., where both methods predicted the same noun in a noun pair as the hypernym) against the number of decisions (i.e., the total number of noun pairs). The heatmaps in Figure 1 show the results for the non-compounds in the English WordNet (left) and in GermaNet (right). They clearly demonstrate that word length makes very different decisions to word frequency and the unsupervised methods, and that word frequency and all unsupervised methods but SLQS Sec highly correlate in their predictions.

3.3 Role of frequency

We go one step further to explore the role of frequency. Figure 2 presents the prediction results on 10 equally-sized subsets of the non-compound pairs in the WordNets after the target pairs were sorted by decreasing difference in hypernym corpus frequency minus hyponym corpus frequency. I.e., in the left-most subset on the x-axis we see the results on the subset with largest differences in hypernym–hyponym frequencies.

We can clearly see that up to subset 7 (up to which the hyponym frequencies are all below the hypernym frequencies), decisions based on word frequency, WeedsPrec, invCL and SLQS Row predict the hypernym almost perfectly; for subset 8 (where the hyponym frequencies start to become larger than the hypernym frequencies) their predictions are becoming worse; and for subsets 9–10 the predictions are mostly wrong. Results by relying on word length and SLQS Sec are clearly worse for the first seven subsets but also better for the last two subsets, thus confirming that they make different predictions.

3.4 Correctness of predictions

While SMC in Section 3.2 informed us about overlap in decisions, it did not tell us whether one of the methods is qualitatively superior, so we analysed whether some methods are simply worse than others, according to their lower accuracy in prediction, or whether the methods all have their own strengths. We calculated for each pair of methods which proportion of wrongly predicted pairs of one method was predicted correctly by the other method. Figure 3 illustrates for the English WordNet how many pairs wrongly predicted by word frequency are predicted correctly by another method (see x-axis).

As we can see, while word length and SLQS Sec are often worse in performance than frequency, they still manage to make correct predictions when frequency fails, which is much less the case for the frequency-alike methods WeedsPrec, invCL and SLQS Row. In particular, invCL seems to make almost identical predictions as frequency, which was already indicated by their almost perfectly overlapping lines in Figure 2.
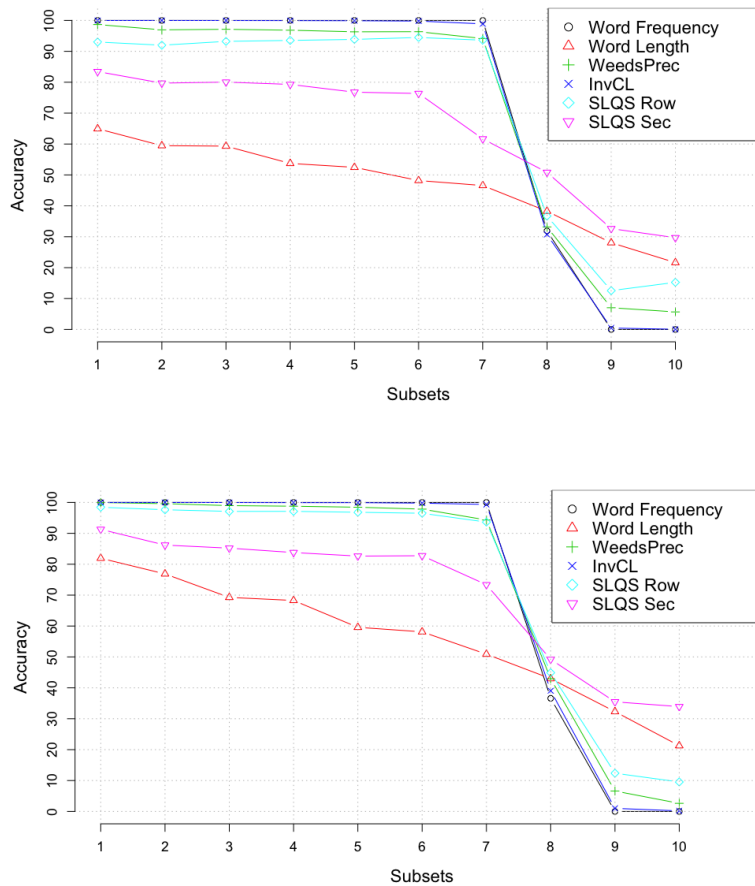
189

Figure 2: Prediction results for WordNet (above) and GermaNet (below) non-compound pairs across equally-sized subsets of target pairs sorted by difference in hypernym frequency and hyponym frequency.
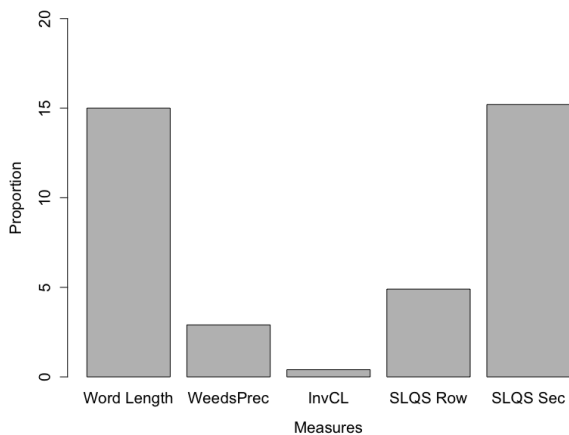

Figure 3: Proportions of pairs predicted wrongly by word frequency but correctly by the given method.

4 Conclusion

This study performed a series of hypernymy predictions by unsupervised methods. We demonstrated that across datasets for English and for German the predictions of three methods (WeedsPrec, invCL and SLQS Row) are highly correlated and also mostly identical with frequency-based predictions. In contrast, word length and SLQS Sec show an overall lower accuracy but at the same time make correct predictions where the others go wrong. Our study once more confirms the general need to check the frequency bias of a computational method in order to identify frequency-(un)related effects.

Acknowledgements

190

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed Distributional Semantic Evaluation. In *Proceedings of the EMNLP Workshop on Geometrical Models for Natural Language Semantics*, pages 1–10, Edinburgh, UK.

Or Biran and Kathleen McKeown. 2013. Classifying Taxonomic Relations between Pairs of Wikipedia Articles. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 788–794, Nagoya, Japan.

Philipp Cimiano, Lars Schmidt-Thieme, Aleksander Pivk, and Steffen Staab. 2004. Learning Taxonomic Relations from Heterogeneous Evidence. In *Proceedings of the ECAI Workshop on Ontology Learning and Population*.

Peter Clark, William R. Murray, John Thompson, Phil Harrison, Jerry Hobbs, and Christiane Fellbaum. 2007. On the Role of Lexical and World Knowledge in RTE3. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, pages 54–59, Prague, Czech Republic.

Daoud Clarke. 2012. A Comparison of Vector-based Representations for Semantic Composition. *Computational Linguistics*, 38(1):41–71.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. *Lecture Notes in Computer Science*, 3944:177–190.

James Deese. 1965. *The Structure of Associations in Language and Thought*. The John Hopkins Press, Baltimore, MD, USA.

Stefan Evert. 2005. *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC – A Corpus of Parsable Sentences from the Web. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, pages 61–68, Darmstadt, Germany.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA, USA.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and Evaluating *ukWaC*, a Very Large Web-Derived Corpus of English. In *Proceedings of the 4th Web as Corpus Workshop*, pages 47–54, Marrakech, Morocco.

John R. Firth. 1957. *Papers in Linguistics 1934-51*. Longmans, London, UK.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – A Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Marti Hearst. 1998. Automated Discovery of WordNet Relations. In (Fellbaum, 1998).

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional Distributional Similarity for Lexical Inference. *Natural Language Engineering*, 16(4):359–389.

Claudia Kunze and Andreas Wagner. 1999. Integrating GermaNet into EuroWordNet, a Multilingual Lexical-Semantic Database. *Sprache und Datenverarbeitung*, 23(2):5–19.

Lothar Lemnitzer and Claudia Kunze. 2007. *Computerlexikographie*. Gunter Narr Verlag, Tübingen, Germany.

Alessandro Lenci and Giulia Benotto. 2012. Identifying Hypernyms in Distributional Semantic Spaces. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*, pages 75–79, Montréal, Canada.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to Wordnet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

George A. Miller and Christiane Fellbaum. 1991. Semantic Networks of English. *Cognition*, 41:197–229.

M. Lynne Murphy. 2003. *Semantic Relations and the Lexicon*. Cambridge University Press.

Roberto Navigli and Simone Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014. Chasing Hypernyms in Vector Spaces with Entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 38–42, Gothenburg, Sweden.

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. EVALution 1.0: an Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics*, pages 64–69, Beijing, China.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging using Decision Trees. In *Proceedings of the 1st International Conference on New Methods in Language Processing*.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving Hypernymy Detection with an Integrated Path-based and Distributional Method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2389–2398, Berlin, Germany.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under Siege: Linguistically-motivated Artillery for Hypernymy Detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–75, Valencia, Spain.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic Taxonomy Induction from Heterogenous Evidence. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 801–808, Sydney, Australia.

Robert R. Sokal. 1958. A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Julie Weeds and David Weir. 2005. A Flexible Framework for Lexical Distributional Similarity. *Computational Linguistics*, 31(4):439–476.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising Measures of Lexical Distributional Similarity. In *Proceedings of the 20th International Conference of Computational Linguistics*, pages 1015–1021, Geneva, Switzerland.

George K. Zipf. 1949. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge.