

A Corpus of Metaphor Novelty Scores for Syntactically-Related Word Pairs

Natalie Parde and Rodney D. Nielsen
Department of Computer Science and Engineering
University of North Texas
{natalie.parde, rodney.nielsen}@unt.edu

Abstract

Automatically scoring metaphor novelty is an unexplored topic in natural language processing, and research in this area could benefit a wide range of NLP tasks. However, no publicly available metaphor novelty datasets currently exist, making it difficult to perform research on this topic. We introduce a large corpus of metaphor novelty scores for syntactically related word pairs, and release it freely to the research community. We describe the corpus here, and include an analysis of its score distribution and the types of word pairs included in the corpus. We also provide a brief overview of standard metaphor detection corpora, to provide the reader with greater context regarding how this corpus compares to other datasets used for different types of computational metaphor processing. Finally, we establish a performance benchmark to which future researchers can compare, and show that it is possible to learn to score metaphor novelty on our dataset at a rate significantly better than chance or naïve strategies.

Keywords: metaphor, metaphor novelty, figurative language, crowdsourcing

1. Introduction

Metaphors exist along a novelty continuum, with highly conventional metaphors (those that a person is likely to encounter on a day-to-day basis, e.g., “I *spent* an *hour* on my homework.”) at one end, and highly novel metaphors (those that stand out as being particularly uncommon or creative, e.g., “She *frowned* like a *thunderstorm*.”) at the other, with many falling somewhere in between. Despite this, research on computational metaphor processing to date has largely viewed metaphor as a binary phenomenon. This simplification has at least partially arisen due to a lack of metaphor novelty-annotated data, as pointed out by Haagsma and Bjerva (2016).

Resource scarcity notwithstanding, having a reliable means of automatically scoring metaphor novelty could be of benefit to many tasks. Word sense disambiguation, for example, can be quite adept at handling conventional metaphors (simply treating them as additional word senses), but novel metaphors require more complex processing (Shutova, 2015; Haagsma and Bjerva, 2016). Automatically grading essays, assessing text difficulty levels, and identifying interesting topics for book discussion systems could all benefit from an ability to automatically score metaphors. Metaphor novelty can also be leveraged for psychological and cognitive health assessment. For example, Gutierrez et al. (2017) applied standard, binary metaphor detection to open-ended patient interviews to aid in the detection of schizophrenia, motivated by clinical research indicating that schizophrenic patients tend to produce particularly bizarre figurative speech (Kuperberg, 2010; Billow et al., 1997). Although the binary metaphor labels were useful, one can hypothesize that performance could be further magnified with continuous novelty scores. Likewise, research in cognitive neuroscience has found that patients with Alzheimer’s disease often struggle with comprehending novel metaphors but not conventional metaphors (Amanzio et al., 2008). Systems capable of distinguishing between different grades of metaphor novelty could thus learn ways to assess cognitive health based on a user’s per-

ceived comprehension of different metaphors.

However, data scarcity currently acts as a barrier to research activity in these promising application areas. In this work, we remove that barrier by presenting a large (18,000+ instances) corpus of syntactically-related word pairs from four domains annotated for metaphor novelty, and make it publicly available to researchers under the Creative Commons Attribution ShareAlike 3.0 Unported License. We first provide an overview of existing metaphor datasets. Then, we describe our data collection process, and present an analysis of this new corpus. Finally, as a proof-of-concept we develop a metaphor novelty scoring approach and train it on this new dataset to establish a performance benchmark for this task.

2. Related Work

Although no datasets annotated for metaphor novelty currently exist, a number of datasets have been released in the past for research on traditional metaphor detection. Details about currently-existing free, publicly-available, English-language metaphor datasets are provided in Table 1.

The most widely-used of these has been the Vrije Universiteit Amsterdam Metaphor Corpus (VUAMC) (Steen et al., 2010). The VUAMC is a subset of the BNC Baby corpus (Consortium, 2005), comprised of document fragments from four domains: news, academic text, fiction, and transcribed conversations. Individual words in those documents are labeled as metaphors. Those metaphors are of all part-of-speech types, ranging from conventional to novel; however, novelty is not specified in the annotations.

Other existing metaphor datasets have filled more specialized needs for different tasks. Some of these have included data from multiple languages. Levin et al. (2014), Tsvetkov et al. (2014), and Mohler et al. (2016) all included instances in English, Spanish, Russian, and Farsi in their data. Levin et al.’s dataset contains conventional metaphors from each of those languages belonging to three target domains: poverty, wealth, and taxation. Tsvetkov et al. (2014) included only English instances in their training set, but built

Name	Citation	Instance Type	Label Type	Word Type(s)	Size	Link
VUAMC	(Steen et al., 2010)	Word	Binary	All	187,570	http://ota.ahds.ac.uk/headers/2541.xml
Levin	(Levin et al., 2014)	Word Pair	Binary	Verb-Sub., Verb-Obj., Verb-Adv., Noun-Adj., Noun-Noun, Noun-Compound	7657 (English) 24,277 (All)	Available upon request from the authors.
TroFi	(Birke and Sarkar, 2006)	Word	Binary	Verb	3737	http://natlang.cs.sfu.ca/software/trofi.html
Tsvetkov	(Tsvetkov et al., 2014)	Word Pair, Word Triple	Binary	Adj.-Noun, Sub.-Verb-Obj.	1968 (Eng., AN) 1831 (Eng., SVO) 2608 (All, AN) 2335 (All, SVO)	https://github.com/ytsvetko/metaphor
Mohammad	(Mohammad et al., 2016)	Word	Binary	Verb	1639	http://saifmohammad.com/WebPages/metaphor.html
LCC Metaphor Dataset	(Mohler et al., 2016)	Word Pair	Discrete Range (0-3)	Pairs with nouns, verbs, multi-word expr., adjs., advs.	8724 (Eng., Free) 15,837 (All, Free)	Available upon request from the authors.
Master Metaphor List	(Lakoff, 1994)	Concept Pair	List	All	208	https://tinyurl.com/mastermetaphor
Metaphors of Mind	(Pasanek, 2015)	Sentence	List	All	14,000+	http://metaphors.iath.virginia.edu/
Our Corpus	—	Word Pair	Contin. Range (0-3)	Pairs with nouns, verbs, adjs., advs., personal pronouns	18,439	http://hilt.cse.unt.edu/resources.html

Table 1: Metaphor Datasets

small test sets in Spanish, Russian, and Farsi (as well as English) to study whether metaphor detection models trained on English samples could be applied to other languages. Mohler et al. (2016) produced *metaphoricity* annotations on a scale from 0-3; theirs is the only existing metaphor dataset to use non-binary labels. The free version of their dataset contains English and Spanish instances, and the paid version contains additional instances for those languages as well as instances in Russian and Farsi.

Some metaphor datasets have been principally concerned with the metaphoric and literal uses of different verbs. The TroFi Metaphor Dataset (Birke and Sarkar, 2006) is one of these, comprised of sentences originating in the Wall Street Journal Corpus (Charniak et al., 2000). The sentences contain metaphoric or literal uses of 50 verbs. Similarly, Mohammad et al. (2016) created a dataset containing sentences with metaphoric or literal uses of 440 verbs. Their data also includes emotionality annotations, and an-

notations indicating relative metaphoricness and emotionality (*more*, *less*, or *equal*) for paired uses of the same verb.

Finally, a few metaphor datasets are simply lists of metaphors. These datasets were developed for research in linguistics or the humanities, rather than for computational analysis. Perhaps the most famous collection of metaphors, the Master Metaphor List (Lakoff, 1994), falls under this category. The Master Metaphor List contains conceptual metaphor mappings that guide the generation of linguistic metaphors (e.g., the conceptual metaphor, TIME is MONEY, gives rise to the linguistic metaphor, “I *spent* an hour on my homework.”). Also a member of this category is the large list of metaphors created by Pasanek (2015). This dataset is comprised of sentences from primarily 18th-century British literature that convey metaphors of the mind (a sample from the dataset is Jane Austen’s “*Astonishment and doubt first seized them; and a shortly succeeding ray of common sense added some bitter emotions of shame.*”).

Instructions

Rate the extent to which the word relations specified are novel metaphors. A novel metaphor is a metaphor that you will have only rarely (or never) seen in the past. It should seem "interesting" or creative.

If you think the word relation is literal (for example, "stood" and "Queen" in the first example below), select "Not Metaphoric." Also select "Not Metaphoric" if you do not think the words relate to one another in the sentence.

Note: All completed HITs will be run through an automatic quality-check program prior to acceptance. Responses that the program identifies as spam will be automatically rejected. Please respond to all questions, and read through the instructions and examples carefully before selecting your answers.

Examples (from Alice in Wonderland):

Sentence: Alice looked up, and there stood the Queen in front of them, with her arms folded, *frowning* like a **thunderstorm**.

Rate the novelty of the metaphor formed between *frowning* and **thunderstorm**:

Metaphor Novelty

Low
High

0: Not Metaphoric
 1
 2
 3

Figure 1: Annotation Instructions

The Metaphors of Mind dataset is still growing, as Pasanek continues to add to it.

Of the datasets described, the closest to ours is the LCC Metaphor Dataset, since it labels word pairs with a range of scores rather than with binary labels. However, the focus of the LCC Metaphor Dataset is on *metaphoricity* (ranging from *no metaphoricity* to *clear metaphor*) rather than *metaphor novelty*. We found upon analysis of the dataset that instances rated with high metaphoricity were not necessarily novel metaphors. The following sample from the LCC Metaphor Dataset illustrates this:

A measure of the protection provided to an industry by the entire *structure* of *tariffs*, taking into account the effects of tariffs on inputs as well as on outputs.

In the LCC dataset, the word pair *structure* of *tariffs* is scored as a “3” (high metaphoricity), and this is a reasonable metaphoricity score; clearly, a tariff cannot have a physical structure. However, this metaphor is also quite conventional, and unlikely to strike any reader as particularly creative (in fact, many readers may not notice that the expression is figurative at all unless it is explicitly pointed out to them). Thus, its *novelty* score should be low.

3. Data Collection

Our dataset contains continuous metaphor novelty annotations for syntactically-related word pairs extracted from the VUAMC. Word pairs are comprised either of two *content words* (nouns, verbs, adverbs, and adjectives, excluding stopwords, proper nouns, and some auxiliary verbs) or a content word and a personal pronoun. Syntactic relations were identified using Stanford CoreNLP (Manning et al., 2014). We required that each word pair contain at least one content word labeled as a metaphor using the original binary annotations provided with the VUAMC, thus entailing that each word pair was *potentially metaphoric*.¹ An-

¹A word’s metaphoric usage is rarely conveyed in all word pair contexts: for example, in “Her *laughter* **waltzed** through the *courtyard*,” the pair {*laughter*, **waltzed**} is a novel metaphor but the pair {**waltzed**, *courtyard*} is non-metaphoric.

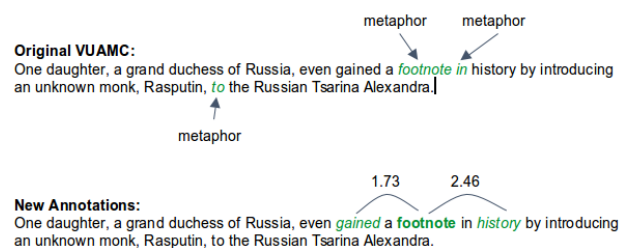


Figure 2: Original VUAMC Corpus and Our Annotations

notators were asked to score the metaphor novelty of each word pair, given the enclosing sentence as context, on a discrete scale from 0 (non-metaphoric) to 3 (highly novel metaphor); the multiple annotations collected for a given word pair were later aggregated to a single continuous label. A sample of the annotation instructions is shown in Figure 1, and Figure 2 illustrates both the annotations included in the original VUAMC and the final, continuous labels included in our dataset.

We collected five annotations from crowd workers for all instances (18,439 word pairs) using Amazon Mechanical Turk (<https://www.mturk.com>). Additionally, we collected annotations from two trained annotators for each instance in the test set (3162 word pairs). The trained annotators used the same labeling interface as the crowd workers. Instances were grouped into Human Intelligence Tasks (HITs), each containing all of the instances associated with 10 sentences, and crowd workers were paid \$0.20 per HIT. Overall, 479 crowd workers participated in annotating the dataset. Workers were required to have an overall HIT approval rate (across all HITs on Amazon Mechanical Turk) greater than or equal to 90%.

Crowdsourced annotations were filtered using the algorithm described in Parde and Nielsen (2017), which identifies substandard workers based on their poor correlation with other workers. Workers who were filtered by the algorithm for a given batch of HITs were also disqualified from accepting future HITs.

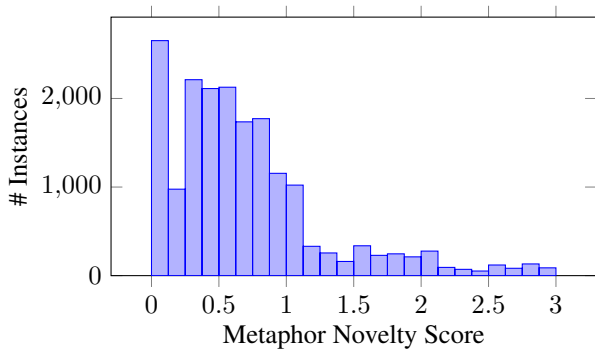


Figure 3: Score Distribution

3.1. Adjudication

The crowdsourced annotations for the training instances were automatically aggregated to continuous labels using the regression-based approach developed by Parde and Nielsen (2017). Briefly, this approach trains a random subspace regression model on data that has been labeled by both crowd workers and experts, using features based on annotation distribution and estimated annotator quality, to predict optimal aggregations of crowd labels.

The annotations provided by trained annotators for the test set were averaged, unless the annotators disagreed strongly (e.g., a 0 and a 3) or if one of the annotators did not agree with the score produced by averaging. In those cases (111 total), instances were forwarded to a third-party adjudicator to make the final decision. The adjudicated labels comprise the dataset’s *gold standard* test labels. Inter-annotator agreement between trained annotators was measured across all 3162 test instances, prior to any adjudication or annotator discussion, using kappa (κ) with quadratic weights between the four “classes” of 0, 1, 2, and 3, resulting in $\kappa = 0.435$. This highlights the fact that scoring metaphor novelty is a difficult task even for humans; however, most of the annotators’ disagreements were just between two adjacent (similar) scores. We also computed κ with the more relaxed constraints established by Mohler et al. (2016) (considering scores within a distance of 1 from one another to agree), and this resulted in $\kappa = 0.897$. In our published dataset, we include the original crowdsourced annotations for each instance, the aggregated label for each training instance, and the gold standard label for each test instance.

4. Corpus Analysis

4.1. Score Distribution

The corpus contains many more conventional metaphors than novel metaphors, which mirrors the distribution seen in naturally-occurring text. The score distribution across the full corpus, with scores binned in 0.125 intervals, is shown in Figure 3. Aggregated training scores were scaled to fit the 0-3 range (the aggregation model originally predicted scores within a more narrow range). Scores from 2.0-3.0 represent novel metaphors, scores from 0.0-1.0 represent conventional (fossilized and non-) metaphors, and scores from 1.0-2.0 are metaphors that fall somewhere between conventional and novel.

		Metaphor Word				
		Noun	Verb	Adj.	Adv.	Total
Other Word	Noun	4352	5630	1533	98	11613
	Verb	2227	1282	275	124	3908
	Adj.	1618	512	83	32	2245
	Adv.	148	418	83	5	654
	Pro.	3	13	3	0	19
	Total	8348	7855	1977	259	18439

Table 2: Word Pair Distribution

4.2. Pair Type Distribution

Unlike many existing metaphor detection corpora, our corpus contains a wide variety of word pairs. This makes it possible to capture metaphors that may be missed by more constrained datasets. Table 2 lists the frequencies with which different types of word pairs occur in our corpus, with the horizontal axis corresponding to the pair’s focus word (the word labeled as a metaphor in the VUAMC), and the vertical axis corresponding to the word paired with that term. The most common pair type in our dataset is verb metaphors paired with nouns, followed closely by noun metaphors paired with nouns. Other common pair types include noun metaphors paired with verbs, noun metaphors paired with adjectives, adjective metaphors paired with nouns, and verb metaphors paired with verbs (e.g., “She *aims* to *explore* the Denton coffee scene.”).

4.3. Dependency Type Distribution

Finally, we analyze the various syntactic relations included in our corpus. Table 3 shows the frequency of each dependency type, the average score associated with it, and the standard deviation of the scores associated with it. The most common relation in our corpus is *nmod* (nominal modifier), followed by *amod* (adjectival modifier), *doobj* (direct object), and *nsubj* (nominal subject). Instances with dependency *amod* had the highest average score, and instances with dependency *nmod:tmod* (temporal modifier) had the lowest. The largest standard deviation among scores was found with instances of type *amod*.

5. Benchmark

Although the focus of this paper is on the corpus itself, as a proof of learnability and to establish a benchmark we also train a metaphor novelty scoring model using our data. To do so, we construct a simple neural network with one hidden layer using Keras (<https://keras.io/>) with a TensorFlow (<https://www.tensorflow.org/>) backend. We apply a random normal initialization function, *tanh* activation, and dropout of 0.1 to both the input layer and hidden layer, and set the number of output nodes to 256 for the input layer and 32 for the hidden layer. We optimize the network using *RM-Sprop*, an optimization technique that updates learning rates for weights based on running averages of the magnitudes of their recent gradients (Hinton et al., 2012), and use a mean squared error loss function. We train the network (5 epochs, with a training batch size of 32) using the following features:

Relation Type	Freq.	Avg.	SD
acl	599	0.62	0.50
acl:relel	573	0.51	0.45
advcl	689	0.31	0.32
advmod	709	0.80	0.63
amod	2668	1.07	0.70
appos	123	0.42	0.54
compound	761	0.93	0.66
csubj	42	0.33	0.30
csubjpass	3	0.35	0.09
dep	423	0.45	0.49
dobj	2276	0.73	0.48
iobj	2	0.72	0.01
nmod	4807	0.69	0.59
nmod:npmmod	14	0.47	0.37
nmod:poss	62	0.59	0.43
nmod:tmod	45	0.29	0.27
nsubj	1460	0.68	0.54
nsubjpass	444	0.68	0.44
reln-obj ²	693	0.33	0.31
reln-subj ²	344	0.42	0.37
subj-obj ²	949	0.37	0.41
xcomp	753	0.53	0.39

Table 3: Dependency Type Distribution

- **Word Embeddings:** 300-dimensional vectors for the word pair’s governor and modifier, extracted from the pretrained Google News word embeddings (Mikolov et al., 2013).
- **Dependency Relation:** One-hot encoded vector indicating the relation between the words in the pair.
- **Word Distance:** Absolute distance between the words in the sentence from which the pair was extracted.

We compare our benchmark approach to several baselines, described in Table 4. Our results are shown in Table 5; we ran our approach 10 times (weights from our approach were randomly initialized) and report average correlation (r) and root mean squared error (RMSE). Differences between our approach and others are statistically significant ($p < 0.0001$).

The results illustrate that it is possible to learn to score metaphor novelty from our dataset at a rate significantly better than chance or naïve strategies like predicting the mean training label. They also provide a performance benchmark for future researchers who use this dataset to develop their own approaches. We explore the task of automatically scoring metaphor novelty further in Parde and Nielsen (2018).

6. Conclusion

In this work, we describe a new, large dataset (18,000+ instances) annotated for metaphor novelty, built on top

²Extracted from subject-relation-object triples generated by Stanford OpenIE (Angeli et al., 2015). Note that upon manual inspection of these pairs, we found many of the reln-obj and reln-subj pairs to be syntactically invalid.

Baseline	Description
RANDOM	Predicts a random continuous value in the range of 0-3 for each instance.
DISTRIBUTION-AWARE RANDOM	Learns a probability density function from the training set and predicts random continuous values from 0-3 with that same distribution.
AVERAGE	Computes the average value from the training set and then predicts that value for each instance.

Table 4: Baseline Approaches

Approach	r	RMSE
RANDOM	0.0048	1.4658
DISTRIBUTION-AWARE RANDOM	0.0007	0.8143
AVERAGE	0.0000	0.7191
OURS	0.4380	0.6528

Table 5: Results

of the VUAMC. We make this dataset publicly available³ under the Creative Commons Attribution ShareAlike 3.0 Unported License. We provide an overview of existing metaphor datasets, and a comprehensive analysis of our new corpus. Finally, we establish a performance benchmark on this dataset to which other researchers may compare their work⁴. In the future, we plan to improve upon the performance of our metaphor novelty scoring system. It is our hope that the availability of this dataset stimulates further research in this area by others as well.

7. Acknowledgements

This material is based upon work supported by the NSF Graduate Research Fellowship Program under Grant 1144248, and the NSF under Grant 1262860. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. Bibliographical References

- Amanzio, M., Geminiani, G., Leotta, D., and Cappa, S. (2008). Metaphor comprehension in alzheimer’s disease: Novelty matters. *Brain and Language*, 107(1):1 – 10.
- Angeli, G., Johnson Premkumar, M. J., and Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

³<http://hilt.cse.unt.edu/resources.html>

⁴<https://github.com/natalieparde/metaphor-novelty-scoring>

- Billow, R. M., Rossman, J., Lewis, N., Goldman, D., and Raps, C. (1997). Observing expressive and deviant language in schizophrenia. *Metaphor and Symbol*, 12(3):205–216.
- Birke, J. and Sarkar, A. (2006). Clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, April.
- Charniak, Eugene and Blaheta, Don and Ge, Niyu and Hall, Keith and Hale, John and Johnson, Mark. (2000). *BLLIP 1987-89 WSJ Corpus*. Linguistic Data Consortium, Penn Treebank, Release 1, ISLRN 233-420-716-637-7.
- BNC Consortium. (2005). *The BNC Baby*. Oxford University Computing Services, British National Corpus, version 2.
- Gutierrez, E. D., Cecchi, G., Corcoran, C., and Corlett, P. (2017). Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2913–2920.
- Haagsma, H. and Bjerva, J. (2016). Detecting novel metaphor using selectional preference information. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 10–17, San Diego, California, June. Association for Computational Linguistics.
- Hinton, G., Srivastava, N., and Swersky, K. (2012). Neural networks for machine learning-lecture 6a-overview of mini-batch gradient descent.
- Kuperberg, G. R. (2010). Language in schizophrenia part 1: an introduction. *Language and linguistics compass*, 4(8):576–589.
- Lakoff, G. (1994). *Master Metaphor List*. University of California.
- Levin, L., Mitamura, T., Macwhinney, B., Fromm, D., Carbonell, J., Feely, W., Frederking, R., Gershman, A., and Ramirez, C. (2014). Resources for the detection of conventionalized metaphors in four languages. In *Proc. of the Ninth International Conf. on Language Resources and Evaluation*. European Language Resources Association, may.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.
- Mohammad, S. M., Shutova, E., and Turney, P. D. (2016). Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*Sem)*, Berlin, Germany.
- Mohler, M., Brunson, M., Rink, B., and Tomlinson, M. (2016). Introducing the lcc metaphor datasets. In *Proc. of the Tenth International Conf. on Language Resources and Evaluation*, May.
- Parde, N. and Nielsen, R. (2017). Finding patterns in noisy crowds: Regression-based annotation aggregation for crowdsourced data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1908–1913, Copenhagen, Denmark, September. Association for Computational Linguistics.
- Parde, N. and Nielsen, R. D. (2018). Exploring the terrain of metaphor novelty: A regression-based approach for automatically scoring metaphors. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana.
- Pasanek, B. (2015). *Metaphors of Mind: An Eighteenth-Century Dictionary*. Johns Hopkins University Press.
- Shutova, E. (2015). Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623, December.
- Steen, Gerard J. and Dorst, Aletta G. and Hermann, J. Berenike and Kaal, Anna and Krennmayr, Tina and Pasma, Trijntje. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins Publishing Company, Converging evidence in language and communication research.
- Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *Proc. of the 52nd Annual Meeting of the ACL*, pages 248–258, June.