Different Texts, Same Metaphors: Unigrams and Beyond

Beata Beigman Klebanov, Chee Wee Leong, Michael Heilman, Michael Flor
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541
{bbeigmanklebanov,cleong,mheilman,mflor}@ets.org

Abstract

Current approaches to supervised learning of metaphor tend to use sophisticated features and restrict their attention to constructions and contexts where these features apply. In this paper, we describe the development of a supervised learning system to classify all content words in a running text as either being used metaphorically or not. We start by examining the performance of a simple unigram baseline that achieves surprisingly good results for some of the datasets. We then show how the recall of the system can be improved over this strong baseline.

1 Introduction

Current approaches to supervised learning of metaphor tend to (a) use sophisticated features based on theories of metaphor, (b) apply to certain selected constructions, like adj-noun or verb-object pairs, and (c) concentrate on metaphors of certain kind, such as metaphors about governance or about the mind. In this paper, we describe the development of a supervised machine learning system to classify all content words in a running text as either being used metaphorically or not – a task not yet addressed in the literature, to our knowledge. This approach would enable, for example, quantification of the extent to which a given text uses metaphor, or the extent to which two different texts use similar metaphors. Both of these questions are important in our target application – scoring texts (in our case, essays written for a test) for various aspects of effective use of language, one of them being the use of metaphor.

We start by examining the performance of a simple unigram baseline that achieves surprisingly good results for some of the datasets. We then show how the recall of the system can be improved over this strong baseline.

2 Data

We use two datasets that feature full text annotations of metaphors: A set of essays written for a large-scale assessment of college graduates and the VUAmsterdam corpus (Steen et al., 2010),[1] containing articles from four genres sampled from the BNC. Table 1 shows the sizes of the six sets, as well as the proportion of metaphors in them; the following sections explain their composition.

Data	#Texts	#NVAR tokens	#metaphors (%)
News	49	18,519	3,405 (18%)
Fiction	11	17,836	2,497 (14%)
Academic	12	29,469	3,689 (13%)
Conversation	18	15,667	1,149 (7%)
Essay Set A	85	21,838	2,368 (11%)
Essay Set B	79	22,662	2,745 (12%)

Table 1: Datasets used in this study. NVAR = Nouns, Verbs, Adjectives, Adverbs, as tagged by the Stanford POS tagger (Toutanova et al., 2003).

2.1 VUAmsterdam Data

The dataset consists of 117 fragments sampled across four genres: Academic, News, Conversation, and Fiction. Each genre is represented by approximately the same number of tokens, although the number of texts differs greatly, where the news archive has the largest number of texts.

We randomly sampled 23% of the texts from each genre to set aside for a blind test to be carried out at a later date with a more advanced system; the current experiments are performed using cross-validation on the remaining 90 fragments: 10-fold on News, 9-fold on Conversation, 11 on Fiction, and 12 on Academic. All instances from the same text were always placed in the same fold.

[1] http://www2.let.vu.nl/oz/metaphorlab/metcor/search/index.html

Proceedings of the Second Workshop on Metaphor in NLP, pages 11–17,
Baltimore, MD, USA, 26 June 2014.

The data is annotated using MIP-VU procedure. It is based on the MIP procedure (Praglejaz, 2007), extending it to handle metaphoricity through reference (such as marking *did* as a metaphor in *As the weather broke up, so did their friendship*) and allow for explicit coding of difficult cases where a group of annotators could not arrive at a consensus. The tagset is rich and is organized hierarchically, detecting various types of metaphors, words that flag the presense of metaphors, etc. In this paper, we consider only the top-level partition, labeling all content words with the tag "function=mrw" (metaphor-related word) as metaphors, while all other content words are labeled as non-metaphors.[2]

2.2 Essay Data

The dataset consists of 224 essays written for a high-stakes large-scale assessment of analytical writing taken by college graduates aspiring to enter a graduate school in the United States. Out of these, 80 were set aside for future experiments and not used for this paper. Of the remaining essays, 85 essays discuss the statement "High-speed electronic communications media, such as electronic mail and television, tend to prevent meaningful and thoughtful communication" (**Set A**), and 79 discuss the statement "In the age of television, reading books is not as important as it once was. People can learn as much by watching television as they can by reading books." (**Set B**). Multiple essays on the same topic is a unique feature of this dataset, allowing the examination of the effect of topic on performance, by comparing performance in within-topic and across-topic settings.

The essays were annotated using a protocol that prefers a reader's intuition over a formal definition, and emphasizes the connection between metaphor and the arguments that are put forward by the writer. The protocol is presented in detail in Beigman Klebanov and Flor (2013). All essays were doubly annotated. The reliability is $\kappa = 0.58$ for Set A and $\kappa = 0.56$ for Set B. We merge the two annotations (union), following the observation in a previous study Beigman Klebanov et al. (2008) that attention slips play a large role in accounting for observed disagreements.

We will report results for 10-fold cross-validation on each of sets A and B, as well as across prompts, where the machine learner would be trained on Set A and tested on Set B and vice versa.

[2]We note that this top-level partition was used for many of the analyses discussed in (Steen et al., 2010).

3 Supervised Learning of Metaphor

For this study, we consider each content-word token in a text as an instance to be classified as a metaphor or non-metaphor. We use the logistic regression classifier in the SKLL package (Blanchard et al., 2013), which is based on scikit-learn (Pedregosa et al., 2011), optimizing for F_1 score (class "metaphor"). We consider the following features for metaphor detection.

- **Unigrams (U)**: All content words from the relevant training data are used as features, without stemming or lemmatization.
- **Part-of-Speech (P)**: We use Stanford POS tagger 3.3.0 and the full Penn Treebank tagset for content words (tags starting with A, N, V, and J), removing the auxiliaries *have*, *be*, *do*.
- **Concreteness (C)**: We use Brysbaert et al. (2013) database of concreteness ratings for about 40,000 English words. The mean ratings, ranging 1-5, are binned in 0.25 increments; each bin is used as a binary feature.
- **Topic models (T)**: We use Latent Dirichlet Allocation (Blei et al., 2003) to derive a 100-topic model from the NYT corpus years 2003–2007 (Sandhaus, 2008) to represent common topics of public discussion. The NYT data was lemmatized using NLTK (Bird, 2006). We used the gensim toolkit (Řehůřek and Sojka, 2010) for building the models, with default parameters. The score assigned to an instance w on a topic t is $log\frac{P(w|t)}{P(w)}$ where $P(w)$ were estimated from the Gigaword corpus (Parker et al., 2009). These features are based on the hypothesis that certain topics are likelier to be used as source domains for metaphors than others.

4 Results

For each dataset, we present the results for the unigram model (**baseline**) and the results for the full model containing all the features. For cross-validation results, all words from the same text were always placed in the same fold, to ensure that we are evaluating generalization across texts.

	M	Unigram			UPCT		
Data	F	P	R	F	P	R	F
Set A	.20	.72	.43	.53	.70	.47	.56
Set B	.22	.79	.54	.64	.76	.60	.67
B-A	.20	.58	.45	.50	.56	.50	.53
A-B	.22	.71	.28	.40	.72	.35	.47
News	.31	.62	.38	.47	.61	.43	.51
Fiction	.25	.54	.23	.32	.54	.24	.33
Acad.	.23	.51	.20	.27	.50	.22	.28
Conv.	.14	.39	.14	.21	.36	.15	.21

Table 2: Summary of performance, in terms of precision, recall, and F_1. Set A, B, and VUAmsterdam: cross-validation. B-A and A-B: Training on B and testing on A, and vice versa, respectively. Column M: F_1 of a pseudo-system that classifies all words as metaphors.

4.1 Performance of the Baseline Model

First, we observe the strong performance of the unigram baseline for the cross-validation within sets A and B (rows 1 and 2 in Table 2). For a new essay, about half its metaphors will have been observed in a sample of a few dozen essays on the same topic; these words are also consistently used as metaphors, as precision is above 70%. Once the same-topic assumption is relaxed down to related topics, the sharing of metaphor is reduced (compare rows 1 vs 3 and 2 vs 4), but still substantial.

Moving to VUAmsterdam data, we observe that the performance of the unigram model on the News partition is comparable to its performance in the cross-prompt scenario in the essay data (compare row 5 to rows 3-4 in Table 2), suggesting that the News fragments tend to discuss a set of related topics and exhibit substantial sharing of metaphors across texts.

The performance of the unigram model is much lower for the other VUAmsterdam partitions, although it is still non-trivial, as evidenced by its consistent improvement over a pseudo-baseline that classifies all words as metaphor, attaining 100% recall (shown in column M in Table 2). The weaker performance could be due to highly divergent topics between texts in each of the partitions. It is also possible that the number of different texts in these partitions is insufficient for covering the metaphors that are common in these kinds of texts – recall that these partitions have small numbers of long texts, whereas the News partition has a larger number of short texts (see Table 1).

4.2 Beyond Baseline

The addition of topic model, POS, and concreteness features produces a significant increase in recall across all evaluations ($p < 0.01$), using McNemar's test of the significance of differences between correlated proportions (McNemar, 1947). Even for Conversations, where recall improvement is the smallest and F_1 score does not improve, the UPCT model recovers all 161 metaphors found by the unigrams plus 14 additional metaphors, yielding a significant result on the correlated test.

We next investigate the relative contribution of the different types of features in the UPCT model by ablating each type and observing the effect on performance. Table 3 shows ablation results for essay and News data, where substantial improvements over the unigram baseline were produced.

We observe, as expected, that the unigram features contributed the most, as removing them results in the most dramatic drop in performance, although the combination of concreteness, POS, and topic models recovers about one-fourth of metaphors with over 50% precision, showing non-trivial performance on essay data.

The second most effective feature set for essay data are the topic models – they are responsible for most of the recall gain obtained by the UPCT model. For example, one of the topics with a positive weight in essays in set B deals with visual imagery, its top 5 most likely words in the NYT being *picture*, *image*, *photograph*, *camera*, *photo*. This topic is often used metaphorically, with words like *superficial*, *picture*, *framed*, *reflective*, *mirror*, *capture*, *vivid*, *distorted*, *exposure*, *scenes*, *face*, *background* that were all observed as metaphors in Set B. In the News data, a topic that deals with hurricane Katrina received a positive weight, as words of suffering and recovery from distaster are often used metaphorically when discussing other things: *starved*, *severed*, *awash*, *damaged*, *relief*, *victim*, *distress*, *hits*, *swept*, *bounce*, *response*, *recovering*, *suffering*.

The part-of-speech features help improve recall across all datasets in Table 3, while concreteness features are effective only for some of the sets.

5 Discussion: Metaphor & Word Sense

The classical "one sense per discourse" finding of Gale et al. (1992) that words keep their senses within the same text 98% of the time suggests that

	Set A cross-val.			Set B cross-val.			Train B : Test A			Train A : Test B			News		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
M	.11	1.0	.20	.12	1.0	.22	.11	1.0	.20	.12	1.0	.22	.18	1.0	.31
U	.72	.43	.53	.79	.54	.64	.58	.45	.50	.71	.28	.40	.62	.38	.47
UPCT	.70	.47	.56	.76	.60	.67	.56	.50	.53	.72	.35	.47	.61	.43	.51
– U	.58	.21	.31	.63	.28	.38	.44	.21	.29	.59	.18	.27	.55	.23	.32
– P	.71	.46	.56	.76	.58	.66	.57	.48	.52	.70	.33	.45	.61	.41	.49
– C	.70	.46	.55	.77	.58	.66	.56	.50	.53	.71	.34	.46	.61	.43	.50
– T	.71	.43	.53	.78	.55	.65	.57	.45	.51	.71	.29	.41	.62	.41	.49

Table 3: Ablation evaluations. Model M is a pseudo-system that classifies all instances as metaphors.

if a word is used as a metaphor once in a text, it is very likely to be a metaphor if it is used again in the same text. Indeed, this is the reason for putting all words from the same text in the same fold in cross-validations, as training and testing on different parts of the same text would produce inflated estimates of metaphor classification performance.

Koeling et al. (2005) extend the notion of discourse beyond a single text to a domain, such as articles on Finance, Sports, and a general BNC domain. For a set of words that each have at least one Finance and one Sports sense and not more than 12 senses in total, guessing the predominant sense in Finance and Sports yielded 77% and 76% precision, respectively. Our results with the unigram model show that guessing "metaphor" based on a sufficient proportion of previously observed metaphorical uses in the given domain yields about 76% precision for essays on the same topic. Thus, metaphoricity distinctions in same-topic essays behave similarly to sense distinctions for polysemous words with a predominant sense in the Finance and Sports articles, keeping to their domain-specific predominant sense $\frac{3}{4}$ of the time.

Note that a domain-specific predominant sense may or may not be the same as the most frequent sense overall; similarly, a word's tendency to be used metaphorically might be domain specific or general. The results for the BNC at large are likely to reflect general rather than domain-specific sense distributions. According to Koeling et al. (2005), guessing the predominant sense in the BNC yields 51% precision; our finding for BNC News is 62% precision for the unigram model. The difference could be due to the mixing of the BNC genres in Koeling et al. (2005), given the lower precision of metaphoricity prediction in non-news (Table 2).

In all, our results suggest that the pattern of metaphorical and non-metaphorical use is in line with that of dominant word-sense for more and less topically restricted domains.

6 Related Work

The extent to which different texts use similar metaphors was addressed by Pasanek and Sculley (2008) for corpora written by the same author. They studied metaphors of mind in the oeuvre of 7 authors, including John Milton and William Shakespeare. They created a set of metaphorical and non-metaphorical references to the mind using excerpts from various texts written by these authors. Using cross-validation with unigram features for each of the authors separately, they present very high accuracies (85%-94%), suggesting that authors are highly self-consistent in the metaphors of mind they select. They also find good generalizations between some pairs of authors, due to borrowing or literary allusion.

Studies using political texts, such as speeches by politicians or news articles discussing politically important events, documented repeated use of words from certain source domains, such as rejuvenation in Tony Blair's speeches (Charteris-Black, 2005) or railroad metaphors in articles discussing political integration of Europe (Musolff, 2000). Our results regarding settings with substantial topical consistency second these observations.

According to the Conceptual Metaphor theory (Lakoff and Johnson, 1980), we expect certain basic metaphors to be highly ubiquitous in any corpus of texts, such as TIME IS SPACE or UP IS GOOD. To the extent that these metaphors are realized through frequent content words, we expect some cross-text generalization power for a unigram model. Perhaps the share of these basic metaphors in all metaphors in a text is reflected most faithfully in the peformance of the unigram model on the non-News partitions of the VUAms-

terdam data, where topical sharing is minimal.

Approaches to metaphor detection are often either rule-based or unsupervised (Martin, 1990; Fass, 1991; Shutova et al., 2010; Shutova and Sun, 2013; Li et al., 2013), although supervised approaches have recently been attempted with the advent of relatively large collections of metaphor-annotated materials (Mohler et al., 2013; Hovy et al., 2013; Pasanek and Sculley, 2008; Gedigan et al., 2006). These approaches are difficult to compare to our results, as these typically are not whole texts but excerpts, and only certain kinds of metaphors are annotated, such as metaphors about governance or about the mind, or only words belonging to certain syntactic or semantic class are annotated, such as verbs[3] or motion words only.

Concreteness as a predictor of metaphoricity was discussed in Turney et al. (2011) in the context of concrete adjectives modifying abstract nouns. The POS features are inspired by the discussion of the preference and aversion of various POS towards metaphoricity in Goatly (1997). Heintz et al. (2013) use LDA topics built on Wikipedia along with manually constructed seed lists for potential source and target topics in the broad target domain of governance, in order to identify sentences using lexica from both source and target domains as potentially containing metaphors. Bethard et al. (2009) use LDA topics built on BNC as features for classifying metaphorical and non-metaphorical uses of 9 words in 450 sentences that use these words, modeling metaphorical vs non-metaphorical contexts for these words. In both cases, LDA is used to capture the topical composition of a sentence; in contrast, we use LDA to capture the tendency of words belonging to a topic to be used metaphorically in a given discourse.

Dunn (2013) compared algorithms based on various theories of metaphor on VUAmsterdam data. The evaluations were done at sentence level, where a sentence is metaphorical if it contains at least one metaphorically used word. In this accounting, the distribution is almost a mirror-image of our setting, as 84% of sentences in News were labeled as metaphorical, whereas 18% of content words are tagged as such. The News partition was very difficult for the systems examined in Dunn (2013) – three of the four systems failed to predict any non-metaphorical sentences, and the one system that did so suffered from a low recall of metaphors, 20%. Dunn (2013) shows that the different systems he compared had relatively low agreement ($\kappa < 0.3$); he interprets this finding as suggesting that the different theories underlying the models capture different aspects of metaphoricity and therefore detect different metaphors. It is therefore likely that features derived from the various models would fruitfully complement each other in a supervised learning setting; our findings suggest that the simplest building block – that of a unigram model – should not be ignored in such experiments.

[3] as in Shutova and Teufel (2010)

7 Conclusions

We address supervised learning of metaphoricity of words of any content part of speech in a running text. To our knowledge, this task has not yet been studied in the literature. We experimented with a simple unigram model that was surprisingly successful for some of the datasets, and showed how its recall can be further improved using topic models, POS, and concreteness features.

The generally solid performance of the unigram features suggests that these features should not be neglected when trying to predict metaphors in a supervised learning paradigm. Inasmuch as metaphoricity classification is similar to a coarse-grained word sense disambiguation, a unigram model can be thought of as a crude predominant sense model for WSD, and is the more effective the more topically homogeneous the data.

By evaluating models with LDA-based topic features in addition to unigrams, we showed that topical homogeneity can be exploited beyond unigrams. In topically homogeneous data, certain topics commonly discussed in the public sphere might not be addressed, yet their general familiarity avails them as sources for metaphors. For essays on communication, topics like sports and architecture are unlikely to be discussed; yet metaphors from these domains can be used, such as *leveling of the playing field* through cheap and fast communications or *buildling bridges* across cultures through the internet.

In future work, we intend to add features that capture the relationship between the current word and its immediate context, as well as add essays from additional prompts to build a more topically diverse set for exploration of cross-topic generalization of our models for essay data.

References

Beata Beigman Klebanov and Michael Flor. 2013. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, Georgia, June. Association for Computational Linguistics.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing disagreements. In *COLING 2008 workshop on Human Judgments in Computational Linguistics*, pages 2–7, Manchester, UK.

Steven Bethard, Vicky Tzuyin Lai, and James Martin. 2009. Topic model analysis of metaphor frequency for psycholinguistic stimuli. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, CALC '09, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Steven Bird. 2006. NLTK: The natural language toolkit. In *Proceedings of the ACL, Interactive Presentations*, pages 69–72.

Daniel Blanchard, Michael Heilman, and Nitin Madnani. 2013. *SciKit-Learn Laboratory*. GitHub repository, https://github.com/EducationalTestingService/skll.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, pages 1–8.

Jonathan Charteris-Black. 2005. *Politicians and rhetoric: The persuasive power of metaphors*. Palgrave MacMillan, Houndmills, UK and New York.

Jonathan Dunn. 2013. What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 1–10, Atlanta, Georgia, June. Association for Computational Linguistics.

Dan Fass. 1991. Met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

William Gale, Kenneth Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the Speech and Natural Language Workshop*, pages 233–237.

Matt Gedigan, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pages 41–48, New York.

Andrew Goatly. 1997. *The Language of Metaphors*. Routledge, London.

Ilana Heintz, Ryan Gabbard, Mahesh Srivastava, Dave Barner, Donald Black, Majorie Friedman, and Ralph Weischedel. 2013. Automatic Extraction of Linguistic Metaphors with LDA Topic Modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66, Atlanta, Georgia, June. Association for Computational Linguistics.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57, Atlanta, GA. Association for Computational Linguistics.

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of HLT-EMNLP*, pages 419–426, Vancouver, Canada. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago.

Hongsong Li, Kenny Q. Zhu, and Haixun Wang. 2013. Data-driven metaphor recognition and explanation. *Transactions of the ACL*, 1:379–390.

James Martin. 1990. *A computational model of metaphor interpretation*. Academic Press Professional, Inc., San Diego, CA, USA.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35, Atlanta, GA. Association for Computational Linguistics.

Andreas Musolff. 2000. *Mirror images of Europe: Metaphors in the public debate about Europe in Britain and Germany*. München: Iudicium. Annotated data is available at http://www.dur.ac.uk/andreas.musolff/Arcindex.htm.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English Gigaword Fourth Edition LDC2009T13. Linguistic Data Consortium, Philadelphia.

Bradley Pasanek and D. Sculley. 2008. Mining millions of metaphors. *Literary and Linguistic Computing*, 23(3):345–360.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Group Pragglejaz. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. LDC Catalog No: LDC2008T19.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In *Proceedings of HLT-NAACL*, pages 978–988.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source - target domain mappings. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 3255–3261, Valletta, Malta, May. European Language Resources Association (ELRA).

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1002–1010.

Gerard Steen, Aletta Dorst, Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification*. Amsterdam: John Benjamins.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of NAACL*, pages 252–259.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Stroudsburg, PA, USA. Association for Computational Linguistics.