Topic Model Analysis of Metaphor Frequency for Psycholinguistic Stimuli

Steven Bethard
Computer Science Department
Stanford University
Stanford, CA 94305
bethard@stanford.edu

Vicky Tzuyin Lai
Department of Linguistics
University of Colorado
295 UCB, Boulder CO 80309
vicky.lai@colorado.edu

James H. Martin
Department of Computer Science
University of Colorado
430 UCB, Boulder CO 80309
james.martin@colorado.edu

Abstract

Psycholinguistic studies of metaphor processing must control their stimuli not just for word frequency but also for the frequency with which a term is used metaphorically. Thus, we consider the task of *metaphor frequency estimation*, which predicts how often target words will be used metaphorically. We develop metaphor classifiers which represent metaphorical domains through Latent Dirichlet Allocation, and apply these classifiers to the target words, aggregating their decisions to estimate the metaphorical frequencies. Training on only 400 sentences, our models are able to achieve 61.3% accuracy on metaphor classification and 77.8% accuracy on HIGH vs. LOW metaphorical frequency estimation.

1 Introduction

Psycholinguistic studies of metaphor try to understand metaphorical language comprehension by presenting subjects with linguistic stimuli and observing their responses. Recent work has observed such responses at the electrophysiological level, measuring brain electrical activity as the stimuli are read (Coulson and Petten, 2002; Tartter et al., 2002; Iakimova et al., 2005; Arzouan et al., 2007; Lai et al., 2007). All these studies have attempted to make comparisons across different types of stimuli (e.g. literal vs. metaphorical) by holding the frequencies of the target words constant across experimental conditions. For example, Tartter et al. (2002) compared the metaphorical and literal sentences *his face was contorted by an angry cloud* and *his face was contorted by an angry frown*, where the two sentences end in different words, but where the final words *cloud* and *frown* had similar word frequencies. As another example, Lai et al. (2007) compared the metaphorical and literal sentences *Their theories have collapsed* and *The old building has collapsed*, where the two sentences end in exactly the same words, so the target word frequencies across conditions were perfectly matched. In both designs, controlling for word frequency allowed the researchers to attribute the differences in experimental conditions to interesting factors, like figurativity, rather than simple word frequency.

However, word frequency is not the only type of frequency relevant to such experiments. In particular, *metaphorical frequency*, that is, how inherently metaphorical one word is as compared to another, may also play an important role in explaining the psycholinguistic results. For example, if *collapsed* is usually used literally, a greater processing effort may be observed when a metaphorical instance of *collapsed* is presented. Likewise, if *collapsed* is usually used metaphorically, greater effort may be observed when a literal instance is presented. Psycholinguistic studies of metaphor have not, to date, controlled for such metaphorical frequency because there were no corpora or algorithms which could provide the needed metaphorical frequencies.

The present study aims to address this deficiency by producing models which can automatically estimate how often a word is used metaphorically. We build these models using only 50 examples each of a small number of target words (< 10), rather than requiring 50 or more examples of every target word

Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity, pages 9–16,
Boulder, Colorado, June 2009.

(100+) in the stimuli, as would be required by standard corpus linguistics methods. Our approach is also novel in that it combines metaphor classification with statistical topic models. Topic models are intuitively promising for our task because they produce topics that seem to translate well to the theory of *conceptual domains*, which suggests that, for example, conceptual domains such as THEORIES and BUILDINGS are used to understand *Their theories have collapsed*. These topic models also show some promise for distinguishing conventional metaphors from novel metaphors.

2 Prior Work

Two types of prior research inform our current study: corpus analyses investigating metaphor frequency by hand, and machine learning models that classify text as either literal or metaphorical. The latter could be used to estimate metaphor frequencies by applying the classifier to a corpus and aggregating the classifications.

2.1 Metaphor Frequency

Researchers have manually estimated several different kinds of metaphor frequency. Pollio et al. (1990) looked at overall metaphorical frequency, performing an exhaustive analysis of a variety of texts, and concluding that there were about five metaphors for every 100 words of text. Martin (1994) looked at the frequency of different types of metaphor, using a sample of 600 sentences from the Wall Street Journal (WSJ), and concluded among other things that the most frequent type of WSJ metaphor was VALUE is LOCATION, e.g. *Spain Fund* ***tumbled*** *23%*. Martin (2006) looked at conditional probabilities of metaphor, for example noting that in 2400 WSJ sentences, the probability of seeing an instance of a metaphor was greatly increased after a first instance had already been observed. However, none of these studies provided the metaphorical frequencies of individual words needed for our research.

Sardinha (2008) performed what is probably closest to the type of analysis we are interested in. Using a corpus of Portuguese conference calls, Berber Sardinha identified 432 terms that were used metaphorically. He then took 100 instances of each of these terms in a general Brazilian corpus and manually annotated them as being either literal or metaphorical. Berber Sardinha found that on average these terms were used metaphorically 70% of the time, and provided analysis of the metaphorical frequencies of a number of individual terms. While it is exactly these kinds of individual term frequencies that we are after, we cannot use Berber Sardinha's data because his corpus was in Portuguese while we are interested in English. This brings out one of the main drawbacks of the corpus annotation approach: moving to a new language (or even a new genre) requires an extensive manual annotation project. Our goal is to avoid such costs by taking advantage of machine learning techniques for automatically identifying metaphorical text.

2.2 Metaphor Classification

Recent years have seen a rising interest in metaphor classification systems. Birke and Sarkar (2006) took a semi-supervised approach, collecting noisy examples of literal and non-literal sentences from both WordNet and metaphor dictionaries, and using a word-based measure of sentence similarity to group sentences into literal and non-literal clusters. They evaluated on hand-annotated sentences for 25 target words and reported an F-score of 0.538, a substantial improvement over the 0.294 majority class baseline.

Gedigian et al. (2006) approached metaphor identification as supervised classification, annotating around 4000 WSJ motion words as literal or metaphorical, and training a maximum entropy classifier using as features based on named entities, WordNet and semantic roles. They achieved an accuracy of 95.1%, a decent improvement over the very high majority class baseline of 93.8%.

Krishnakumaran and Zhu (2007) focused on three syntactically constrained sub-types of metaphors: nouns joined by *be*, nouns following verbs, and nouns following adjectives. They combined WordNet hypernym information with bigram statistics and a threshold, and evaluated their algorithm on the Berkeley Master Metaphor List (Lakoff, 1994), achieving an accuracy of around 46%.

All of these approaches produced models which could be applied to new text to identify metaphors, but each has some drawbacks for our task. The WSJ study of Gedigian et al. (2006) found 94% of their target words to be metaphorical, a vastly differ-

Target	L	M	M%
attacked	32	18	36%
born	45	5	10%
budding	16	34	68%
collapsed	10	40	80%
digest	7	43	86%
drifted	16	34	68%
floating	25	25	50%
sank	31	19	38%
spoke	47	3	6%
Total	229	221	49%

Table 1: Metaphorical (M) and literal (L) counts, and metaphorical percentage (M%), for the annotated verbs.

ent number from the 49% for our target words (see Section 3). Krishnakumaran and Zhu (2007) considered only a few different syntactic constructions, but we need to consider all the ways a metaphor may be expressed to evaulate overall metaphor frequency. Birke and Sarkar (2006) did consider a variety of target words in unrestricted text, but relied on large scale language resources like WordNet and metaphor dictionaries, while we are interested in approaches that are less resource intensive.

Thus, rather than basing our models on these prior systems, we develop a novel approach to metaphor frequency estimation based on using topic models to operationalize metaphorical domains.

3 Data

The first step in building models of metaphorical frequency is obtaining data for training and evaluation. In one of the post-hoc analyses of the Lai et al. (2007) experiment, 50 sentences from the British National Corpus (BNC, 2007) were gathered for each of nine of their target words. They annotated each instance as either literal or metaphorical, and then used these annotations to calculate metaphorical frequencies for analysis.

This data served as our starting point for exploring computational approaches to estimating metaphorical frequency. Table 1 shows the nine verbs and their metaphorical frequencies. Table 2 shows some examples. Some verbs, such as *digest*, are almost always used metaphorically (86% of the time), while other verbs, such as *spoke*, are almost always used literally (94% of the time). Annotation of just 50 instances of each of these nine verbs was time consuming, and yet to fully re-analyze the ERP results, metaphorical frequencies would be needed for all of the over 100 target words. Thus our goal was to automate this process.

L	Aye, that's where I was *born* and reared.
M	VATman threatens our *budding* entrepreneurs.
M	Suddenly all her bravado *collapsed*.
L	This makes it easier for us to *digest* the wheat.
L	Gulls *drifted* lethargically on the swell.
M	My heart *sank* as I looked around.

Table 2: Examples of sentences with metaphorical (M) and literal (L) target words.

T#	Most frequent words
00	book (4%) write (2%) read (2%) english (2%)
17	record (3%) music (2%) band (2%) play (2%)
42	social (3%) history (2%) culture (1%) society (1%)
58	film (3%) play (2%) theatre (1%) women (1%)
82	dog (9%) rabbit (2%) ferret (1%) pet (1%)

Table 3: Example topics (T#) from the BNC and their most frequent words. Numbers in parentheses indicate the percent of the topic each word represents.

4 Topic Models

Our approach to estimating metaphorical frequencies was first to classify words in unrestricted text as literal or metaphorical, and then to aggregate those decisions to estimate a frequency. Thus, we first needed to build a model which could identify metaphorical expressions. Our approach to this problem was based on the theory of conceptual domains, in which metaphors are seen as taking terms from one domain (e.g. *attacked*) and applying them to another domain (e.g. *argument*).

To operationalize these domains, we employed statistical topic models, in particular, Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Intuitively, LDA looks at how words co-occur in the documents of a large corpus, and identifies *topics* or groups of words that are semantically similar. For example, Table 3 shows a few topics from the BNC. These topics can be thought of as grouping words by their semantic domains. For example, we might think of topic 00 as the *Book* domain and topic 42 as the *Society* domain. Because LDA generates topics that look

much like the source and target domains associated with metaphors, we expect that LDA can provide a boost to metaphor identification models.

The LDA algorithm is usually presented as a generative model, that is, as an imagined process that someone might go through when writing a text. This generative process looks something like:

1. Decide what topics you want to write about.
2. Pick one of those topics.
3. Think of words used to discuss that topic.
4. Pick one of those words.
5. To generate the next word, go back to 2.

This is a somewhat unrealistic description of the writing process, but it gets at the idea that the words in a document are topically coherent. Formally, the process above can be described as:

1. For each document d select a topic distribution $\theta^d \sim Dir(\alpha)$
2. Select a topic $z \sim \theta^d$
3. For each topic select a word distribution $\phi^z \sim Dir(\beta)$
4. Select a word $w \sim \phi^z$

The goal of the LDA learning algorithm then is to maximize the likelihood of our documents, where for one document $p(d|\alpha, \beta) = \prod_{i=1}^{N} p(w_i|\alpha, \beta)$. Estimating these probabilities can be done in a few different ways, but in this paper we use Gibbs sampling as it has been widely implemented and was available in the LingPipe toolkit (Alias-i, 2008).

Gibbs sampling starts by randomly assigning topics to all words in the corpus. Then the word-topic distributions and document-topic distributions are estimated using the following equations:

$$P(z_i|z_{i-}, w_i, d_i, w_{i-}, d_{i-}, \alpha, \beta) = \frac{\phi_{ij}\theta_{jd}}{\sum_{t=1}^{T}\phi_{it}\theta_{td}}$$

$$\phi_{ij} = \frac{C_{word_{ij}}+\beta}{\sum_{k=1}^{W} C_{word_{kj}}+W\beta} \quad \theta_{jd} = \frac{C_{doc_{dj}}+\alpha}{\sum_{k=1}^{T} C_{doc_{dk}}+T\alpha}$$

$C_{word_{ij}}$ is the number of times word i was assigned topic j, $C_{doc_{dj}}$ is the number of times topic j appears in document d, W is the total number of unique words in the corpus, and T is the number of topics requested. In essence, we count the number of times that a word is assigned a topic and the number of times a topic appears in a document, and we use these numbers to estimate word-topic and document-topic probabilities. Once topics have been assigned and distributions have been calculated, Gibbs sampling repeats the process, this time selecting a new topic for each word by looking at the calculated probabilities. The process is repeated until the distributions become stable or a set number of iterations is reached.

We ran LDA over the documents in the BNC, extracting 100 topics after 2000 iterations of Gibbs sampling. We left the α and β parameters at their LingPipe defaults of 0.1 and 0.01, respectively. Table 3 shows some of the resulting topics.

5 Metaphor Frequency

Our primary goal was to use the topics produced by LDA to help characterize words in terms of their metaphorical frequency. We approached this problem by first training metaphor classifiers based on LDA topics to identify target words in text as literal or metaphorical. Then we ran these classifiers over unseen data, and aggregated the individual decisions. The result is an approximate metaphorical frequency for each word. The following sections detail this process and discuss our preliminary results.

5.1 Metaphor Classification

Our data is composed of 50 sentences for each of nine target words, with each sentence annotated as either metaphorical or literal. We treated this as a classification task, where the classifier took as input a sentence containing a target word, and produced as output either LITERAL or METAPHORICAL.

We trained support vector machine (SVM) classifiers on this data, using LDA topics as features. For each of the sentences in our data, we used the LDA topic models to assign topic probability distributions to each of the words in the sentence. We then summed the topic distributions over all the words in the sentence to produce a sentence-wide topic distribution. The result was that for each sentence we could say something like "this sentence was composed of 5% topic 00, 2% topic 01, 8% topic 02, etc." We used these sentence-level topic probability distributions as features for an SVM classifier, in particular, SVM$^{\text{perf}}$ (Joachims, 2005).

We compared this SVM-LDA model against two baselines. The first was the standard majority class

classifier, which simply assigns all instances in the test data whichever label (metaphorical or literal) was most comon in the training data.

The second baseline was an SVM based on TF-IDF features, a well known document classification model (Joachims, 1998; Sebastiani, 2002; Lewis et al., 2004). Under this approach, there is a numeric feature for each of the 3000+ words in the training data, and each word feature is assigned the weight:

$$\frac{|\{w \in doc : w = word\}|}{|\{w \in doc\}|} \cdot \log \frac{|\{d \in docs\}|}{|\{d \in docs : w \in d\}|}$$

Essentially, this formula means that the weight increases with the number of times the word occurs in the document, and decreases with the number of documents in the corpus that contain that word. The vectors of TF-IDF features are then normalized to have Euclidean length 1.0, using the formula:

$$weight(word) = \frac{\text{tf-idf}(word)}{\sqrt{\sum_{word'} \text{tf-idf}(word')^2}}$$

To evaluate our model against both the majority class and the TF-IDF baselines, we ran nine-fold cross-validations, where each fold corresponded to a single target word. Note that this means that we trained our models on the sentences of eight target words, and tested on the sentences of the ninth target word. This is a harder evaluation than a stratified cross-validation where all target words would have been observed during training. But it is a much more realistic evaluation for our task, where we want to learn enough about metaphors from nine target words that we can automatically classify instances of the remaining 95.

Table 4 compares the performance of our SVM-LDA model and the baseline models[1]. The majority class classifier performs poorly, achieving only 26.4% accuracy[2]. The TF-IDF based model performs much better, at 50.7% accuracy. However, our SVM based on LDA features outperforms both baseline models, achieving 54.9% accuracy.

[1]For all models, hyper parameters (the cost parameter, the loss function, etc.) were set using only the training data of each fold by running an inner eight-fold cross validation.

[2]This might be initially surprising since our corpus was 49% metaphorical. Consider, however, that during cross validation, holding out a more metaphorical target word for testing means that our training data is more literal, and vice versa.

Model	Accuracy
Majority Class	26.4%
SVM + TF-IDF	50.7%
SVM + LDA topics	54.9%
SVM + LDA topics + LDA groups	61.3%

Table 4: Model performance on the literal vs. metaphorical classification task.

Type	Most frequent words
CONCRETE	book write read english novel
ABSTRACT	god church christian jesus spirit
MIXED	sleep dream earth theory moon
OTHER	many time only number large

Table 5: Examples of annotated topics.

5.2 Annotating Topics

The metaphor classification results showed the benefit of operationalizing metaphor domains as LDA topics. But metaphors are typically viewed as mapping a *concrete* source domain onto an *abstract* target domain, and our LDA topics had no direct notion of this concrete/abstract distinction. To try to represent this distinction, we manually annotated[3] the 100 LDA topics with one of four labels: CONCRETE, ABSTRACT, MIXED or OTHER. Table 5 shows examples of the annotated topics.

We then used the annotated topics to generate new features for our classifiers. In addition to the original 100 topic probability features, we provided four new probability features, one for each of our labels, calculated by taking the sum of the probabilities of the corresponding topics. For example, since topics 07, 13, 37 and 77 were identified as ABSTRACT topics, the probability of the new ABSTRACT feature was just the sum of the probabilities of the topic features 07, 13, 37 and 77. The last row of Table 4 shows the performance of the SVM model trained with the augmented feature set. This model outperforms all our other models, achieving an accuracy of 61.3% on the literal vs. metaphorical distinction.

These results are interesting because they show that human analysis of LDA topics can add substantial value for machine learning models at a low cost. Annotating the entire set of 100 topics took under

[3]All annotation was performed by a single annotator. Future work will measure inter-annotator agreement.

Model	Accuracy
Majority Class	0.0%
SVM + TF-IDF	22.2%
SVM + LDA topics	55.6%
SVM + LDA topics + LDA groups	77.8%

Table 6: Model performance on the HIGH vs. LOW metaphor frequency prediction task.

an hour, and yet provided a 6% gain in model accuracy. The speed of annotation suggests that LDA topics are conceptually accessible to humans, and the performance boost suggests that manual grouping of LDA topics may be a fruitful area for feature engineering.

5.3 Predicting Metaphorical Frequencies

Having constructed successful metaphor classification models, we return to our question of metaphorical frequency. Given a target word, can we predict the frequency with which that word will be used metaphorically? Our models are not accurate enough that we can expect the frequencies derived from them to be exact predictions of metaphorical frequency. But we may be able to distinguish, for example, words with high metaphorical frequency from words with low metaphorical frequency.

Thus, we evaluate our models on the binary task of assigning target words an overall metaporical frequency, either HIGH ($\geq 50\%$) or LOW ($< 50\%$). We can perform this evaluation using the same data and cross validation technique as before, this time examining each testing fold (which corresponds to a single target word) and aggregating the metaphor classifications to get a metaphorical frequency estimate of that target. Table 6 shows how the models fared on this task. The majority class model misclassified all the words, and the TF-IDF model managed to get only two of the nine correct. The LDA models performed better, with the model including the grouped topic features achieving 77.8% accuracy. This suggests that our model may already be good enough to use for analysis of the original Lai experimental data. Of course, this evaluation was carried out only over the nine available target words, so additional evaluation will be necessary to confirm these trends.

To further analyze our model performance, we looked at the metaphorical frequency estimates for each target word. Table 7 shows the estimates of our best model along with the true metaphorical frequencies. The three target words with the largest differences between true and predicted accuracies are *spoke*, *floating* and *digest*, with *spoke* and *floating* predicted to be much more metaphorical than they actually are, and *digest* predicted to be much less.

Word	True	Predicted	Difference
attacked	36%	24%	-12%
born	10%	2%	-8%
budding	68%	98%	+30%
collapsed	80%	98%	+18%
digest	86%	40%	-46%
drifted	68%	92%	+24%
floating	50%	100%	+50%
sank	38%	26%	-12%
spoke	6%	62%	+56%

Table 7: Model performance on the HIGH vs. LOW metaphor frequency prediction task.

We also performed some analysis of the model errors. In many cases it was difficult to judge why the model succeeded or failed in identifying a metaphor, but a couple of things stood out. First, 70% of the *digest* instances our model misclassified were *Digest* (capitalized), e.g. *Middle East Economic Digest*. Our topic models were trained on all lowercased words, so *Digest* and *digest* were not distinguished. Re-training the models without collapsing the case distinctions might address this problem. Second, *spoke* seems to be an inherently harder term to classify because it co-occurs with so many other topics. About 40% of the *spoke* instances occurred as *spoke of* or *spoke about*, where speaking about a metaphorical topic caused *spoke* to be interpreted metaphorically, and speaking about a literal topic caused *spoke* to be interpreted literally. Addressing this problem would probably require some understanding of argument structure, perhaps akin to what was done by Gedigian et al. (2006).

6 Metaphor Novelty

As a final exploration of topic models for metaphorical domains, we considered metaphorical novelty, as used in the original Lai experiment. In particular, we were interested in how LDA topics might reflect

Type	Stimulus Sentence
LIT	Every soldier in the frontline was attacked
CON	Every point in my argument was attacked
NOV	Every second of our time was attacked
ANOM	Every drop of rain was attacked
LIT	The old building has collapsed
CON	Their theories have collapsed
NOV	Their compromises have collapsed
ANOM	The apples have collapsed

Table 8: Example stimuli: literal (LIT), conventional metaphor (CON), novel metaphor (NOV) and anomalous (ANOM).

more conventional or more novel metaphors. In the Lai experiment, conventional and novel metaphors for a particular target word shared the same source domain (e.g. WAR) but differed in the target domain (e.g. ARGUMENT vs. TIME). If LDA topics are a good operationalization of such domains, then it should be possible use LDA topics to distinguish between conventional and novel metaphors.

To explore this area, we employed the stimuli from the Lai experiment, and looked in particular at the conventional and novel conditions. The Lai experiment used 104 different target words, so these data included 104 conventional metaphors and 104 novel metaphors. Novel metaphors were generated for the Lai experiment by considering a conventional source-target mapping and selecting a new target domain. For example, the conventional metaphor *Every point in my argument was attacked* maps the source domain WAR to the target domain ARGUMENT, while the novel metaphor *Every second of our time was attacked* maps the source domain WAR to the target domain TIME. Table 8 shows example stimulus sentences from the Lai experiment. Though these experimental stimuli have the drawback of being manually constructed, not collected from a corpus, they have the advantage of being already annotated with a definition of novelty that clearly distinguishes the two types of metaphors.

We performed a simple correlational analysis using the conventional and novel metaphors from the Lai experiment. We produced topic distributions for each stimulus, using our topic models trained on the BNC. We then labeled conventional metaphors as -1 and novel metaphors as +1, and identified the top-

-0.19	like house old shop door look street room
-0.18	darlington programme club said durham hall
-0.15	film play theatre women actor work perform
-0.14	area local plan develop land house rural urban
-0.14	any sale good publish custom product price

Table 9: Top 5 topics correlated with conventionality.

0.20	freud sexual sophie male joanna people female
0.17	doctor leed rory dalek fergus date subject aug
0.13	book write read english novel publish reader
0.11	lorton kirov dougal jed manville vologski celia
0.09	war british france britain french nation europe

Table 10: Top 5 topics correlated with novelty.

ics that correlated best with this distinction. Table 9 shows the most negatively correlated (conventional) topics and Table 10 shows the most positively correlated (novel) topics.

Though even the best correlations are somewhat low, there seem to be some trends in this analysis. Conventional metaphors seem to correspond more to concrete terms, like *house*, *club*, *play* and *sale*. Novel metaphors have less of a coherent theme, including terms like *freud* and *sexual* as well as names like *Rory*, *Kirov* and *Britain*. This may reflect a real distinction in the use of conventional and novel metaphors, or it may be an artifact of how the experimental stimuli were created. A deeper investigation into the relations between LDA topics and metaphor novelty will probably require annotating sentences from some naturally occuring data.

7 Conclusions

We presented a novel two-phase approach to the task of *metaphorical frequency estimation*. First, examples of a target word were automatically classified as literal or metaphorical, and then these classifications were aggregated to estimate how often the target word was used metaphorically. Our classifiers operationalized metaphorical source and target domains using topics derived from Latent Dirichlet Allocation. Support vector machine classifiers took these topic probability distributions and learned to classify sentences as literal or metaphorical. These models achieved 61.3% accuracy on the classifiation task, and their aggregated classifications produced an accuracy of 77.8% on the task of distinguishing

between target words with high and low metaphorical frequencies.

Future work will perform a larger scale evaluation, and will use our model's metaphorical frequency estimates to analyze psycholinguistic data. In particular, we will split the conventional metaphorical sentences of Lai et al. (2007) into low and high-frequency items. If the low and high frequency items display significantly different brainwave patterns, then this could suggest that metaphorical frequency of a given word plays a critical role in metaphor comprehension.

Future work will also explore frequency effects that consider the sentential context in the stimulus items. For example, a context like "*Their theories have ____*" probably gives a higher expectation of a metaphorical word filling in the blank than a context like "*The old building has ____*". Having a measure of how much the words in the preceding context predict an upcoming metaphor would provide another useful stimulus control.

References

Alias-i. 2008. LingPipe 3.7.0. http://alias-i.com/lingpipe/, October.

Yossi Arzouan, Abraham Goldstein, and Miriam Faust. 2007. Brainwaves are stethoscopes: ERP correlates of novel metaphor comprehension. *Brain Research*, 1160:69–81, July.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *European Chapter of the ACL (EACL)*.

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

BNC. 2007. The british national corpus, version 3 (BNC XML edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/.

Seana Coulson and Cyma Van Petten. 2002. Conceptual integration and metaphor: an event-related potential study. *Memory & Cognition*, 30(6):958–68, September. PMID: 12450098.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Workshop On Scalable Natural Language Understanding*.

Galina Iakimova, Christine Passerieux, Jean-Paul Laurent, and Marie-Christine Hardy-Bayle. 2005. ERPs of metaphoric, literal, and incongruous semantic processing in schizophrenia. *Psychophysiology*, 42(4):380–390.

Thorsten Joachims, 1998. *Text categorization with Support Vector Machines: Learning with many relevant features*, pages 137–142. Springer Berlin / Heidelberg.

Thorsten Joachims. 2005. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pages 377–384, Bonn, Germany. ACM.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Workshop on Computational Approaches to Figurative Language*.

Vicky Tzuyin Lai, Tim Curran, and Lise Menn. 2007. The comprehension of conventional and novel metaphors: An ERP study. In *13th Annual Conference on Architectures and Mechanisms for Language Processing*, August.

George Lakoff. 1994. Conceptual metaphor WWW server. http://cogsci.berkeley.edu/lakoff/.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: a new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397.

James H. Martin. 1994. MetaBank: a Knowledge-Base of metaphoric language conventions. *Computational Intelligence*, 10(2):134–149.

James H. Martin. 2006. A rational analysis of the context effect on metaphor processing. In Stefan Th. Gries and Anatol Stefanowitsch, editors, *Corpus-Based Approaches to Metaphor and Metonymy*. Mouton de Gruyter.

Howard R. Pollio, Michael K. Smith, and Marilyn R. Pollio. 1990. Figurative language and cognitive psychology. *Language and Cognitive Processes*, 5:141–167.

Tony Berber Sardinha. 2008. Metaphor probabilities in corpora. In Mara Sofia Zanotto, Lynne Cameron, and Marilda do Couto Cavalcanti, editors, *Confronting Metaphor in Use*, pages 127–147. John Benjamins.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47.

Vivien C. Tartter, Hilary Gomes, Boris Dubrovsky, Sophie Molholm, and Rosemarie Vala Stewart. 2002. Novel metaphors appear anomalous at least momentarily: Evidence from N400. *Brain and Language*, 80(3):488–509, March.