Mining Online Discussion Forums for Metaphors

Andrew Gargett, John Barnden
School of Computer Science
University of Birmingham
A.D.Gargett@cs.bham.ac.uk, J.A.Barnden@cs.bham.ac.uk

Abstract

We present an approach to mining online forums for figurative language such as metaphor. We target in particular online discussions within the illness and the political conflict domains, with a view to constructing corpora of Metaphor in Illness Discussion, and Metaphor in Political Conflict Discussion. This paper reports on our ongoing efforts to combine manual and automatic detection strategies for labelling the corpora, and present some initial results from our work showing that metaphor use is not independent of illness domain.

Keywords: metaphor, corpus linguistics, illness and political conflict discourse

1. Background

Corpus-based studies of metaphor are well-established (e.g. Shutova et al. 2013). Such studies have used a variety of manual but also more automated techniques. The role of surface linguistic patterns in manually detecting metaphor use has been well-described (e.g. (Skelton et al., 2002; Deignan, 2008; Deignan, 2006; Hidalgo Downing and Kraljevic Mujic, 2009; Li and Sporleder, 2010)). State-of-the-art automatic techniques typically look to less direct features for detecting metaphorical use of lexical items that are syntactically related (e.g. in a head-dependency relation), including techniques for determining preferential selections (e.g. (Mason, 2004)), as well as those for determining degrees of abstractness and of similarity of such lexical items (e.g. (Birke and Sarkar, 2006; Tsvetkov et al., 2013)).
One of the most comprehensive manual approaches to identifying metaphorical language use is the *Metaphor Identification Procedure*, or MIP(VU),[1] (Pragglejaz Group, 2007), which employs the following procedure for detecting metaphor:

1. Read the entire text to establish a general understanding of the meaning.
2. Determine the lexical units in the text.
3. Label each lexical unit in the text:
   (a) For each lexical unit, establish its meaning in context, that is, how it applies to an entity, relation, or attribute in the situation evoked by the text (contextual meaning). Take into account what comes before and after the lexical unit.
   (b) For each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. Basic meanings tend to be:
      i. More concrete;
      ii. Related to bodily action;
      iii. More precise (as opposed to vague);
      iv. Historically older;
   (c) Basic meanings are not necessarily the most frequent meanings of the lexical unit.
   (d) If the lexical unit has a more basic current-contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.
4. If yes to all of the above, mark the lexical unit as metaphorical.

A contentious aspect of this procedure is the requirement in (3b) to determine whether the lexical unit under consideration is being used in its "most basic meaning" or not (e.g. see discussion in (Shutova et al., 2013)). Resolving this issue is crucial for automating the MIP procedure ((Dorst et al., 2013)).
There is a wealth of online data involving highly metaphorical language, and online discussion forums are particularly notable for this. Such online forms of language are not only notably figurative in expression, but are also often remarkably similar to the kind of language used during dialogue, yet such aspects of this form of language are relatively understudied, although there is growing interest in such language, typically from more computational approaches (Balahur et al., 2013).
The rest of this paper is structured as follows: Section (2.) outlines the approach we have taken to address the issues identified in this section. Section 3 outlines our ongoing work on the data we have collected, and presents some preliminary results we have obtained in our efforts to develop tools for supporting annotators using procedures like the MIP. Section 4 concludes the paper.

2. Method

In response to issues such as those outlined in Section (3.2.), we are developing a semi-supervised annotation scheme for metaphor detection. We aim to use this approach to facilitate a particular metaphor annotation procedure, namely the MIP(VU) (Pragglejaz Group, 2007).

2.1. Data

Online discussion forums are inherently "dialogic", being relatively informal in register and displaying nested struc-

[1] Actually an updated version of the earlier MIP.

ture. Similar to multi-party conversation, a contribution to a forum typically receives multiple replies, and many of these replies are in turn replied to, so that the resulting structure of original forum post plus various responses (and responses to responses, and so on), are highly reminiscent of the structure of everyday conversation (Levinson, 2006). The corpora we are building incorporate information about such structure, which is useful for capturing patterns of metaphor across participants engaged in such embedded interactions.

A forum is also typically organised into topics, which generally coincide with large divisions within the domain itself. For example, a forum on illness might be divided into topics on diabetes, cancer, infectious diseases and stress. This sort of division allows investigation of the different role of metaphor across different domains, and we have made use of this fact in our work (see below).

2.2. Annotation

In order to carry out annotation of each type of forum, illness discussion and conflict discussion, we combine manual and automatic techniques. Manual annotation is carried out using a modified version of MIP(VU), along the lines suggested in (Shutova et al., 2013), and we present preliminary results of a study employing this annotation procedure below. As noted above, (Shutova et al., 2013) draw attention to the difficulty of determining whether an item has a so-called "more basic meaning" (part (3b) of the MIP(VU) procedure), and in response suggest avoiding the requirement that annotators strictly employ dictionaries to find evidence to support a judgement that there is a *more basic meaning* of the lexical unit under consideration. Rather, (Shutova et al., 2013) propose that annotators should "imagine the contexts in which the verb has a more basic meaning." They fully acknowledge that with this approach there is considerable risk that results will become more unconstrained with the greater flexibility (and indeed they reported only moderate agreement between annotators for their study).

On our approach, manual annotation is supported by an automatically generated *confidence measure*, which is a real number value drawn from the interval $[0, 1]$, regarding whether a particular lexical item should be labelled metaphorical, 0 meaning **no confidence that this item is metaphorical** through to 1 meaning **complete confidence that this item is metaphorical**. A confidence measure c is a function:

$$c : X \times Y \times K \to I \tag{1}$$

where X and Y are input and output domains respectively, K is a set of features for determining confidence,[2] and I is the range of the confidence measure (e.g. the interval $[0, 1]$).

(Gandrabur et al., 2006) refer to two subclasses for this general model of confidence, *posterior probabilities* and *correctness probabilities*. For correctness probabilities, the general situation is reduced to the probability that the specific output $y \in Y$ is correct, given input $x \in X$, and features $k \in K$: $P(C = 1|x, y, k)$ for $c \in C$. However, correctness probabilities are more relevant for situations where there are numerous *equally* correct candidates (e.g. machine translation tasks), whereas we are interested in accurately judging whether *x is a metaphor-related word*, abbreviated to x_{MRW}.[3] More relevant for our approach is the posterior probability that this judgement is correct,[4] given some input and some set of features (which is to say, based on prior evidence), and a function $label(.)$ for labelling items as **mrw/not**. For the judgment that *x is a metaphor-related word*, or x_{MRW} for short, the probability associated function can be transformed into a binary decision in terms of some threshold t as follows:

$$x_{MRW} = \begin{cases} \text{correct if } P(label(x) = mrw|x, k) \geq \text{t} \\ \text{incorrect otherwise} \end{cases} \tag{2}$$

During the annotation task, the annotator receives from the automatic annotation module a judgment as to whether specific words of the text are *metaphorically related words* or not, as well as a level of confidence about this judgment. Using a procedure such as (2), the confidence measure defined in (1) is converted into a judgment about how the lexical unit being considered should be labeled (e.g. *yes, this is a metaphorically related word*, or *no, this is not a metaphorically related word*). The annotator my then use this judgment together with the original confidence measure, to help them in arriving at decision about the correct label to apply to the linguistic expression being considered. Working out the set of features K for equation (1) is an empirical matter, and in Section (3.), we report preliminary results for ongoing corpus-based investigations which we are currently undertaking. For example, restricting judgement of metaphoricity to predicates, we are testing whether it can be reliably judged that an individual predicate is being used non-literally, based on the relative concreteness of its arguments (e.g. (Tsvetkov et al., 2013)). For example, for the sentences *The cars were racing* vs. *Her mind was racing*, which have predicate-argument sets $\langle race, car\rangle$ and $\langle race, mind\rangle$, respectively, the arguments are significantly less concrete than the predicate in the latter vs. the former sentence. This difference in relative concreteness of predicates vs. arguments has been argued to indicate metaphorical meanings ((Tsvetkov et al., 2013)), and we are currently investigating this claim, as part of our ongoing work toward an optimal set of features for detecting metaphor.

[2] For example, as pointed out later in this section, relative concreteness of predicates vs. arguments has been reported in the literature (Tsvetkov et al., 2013) to correlate with metaphorical forms of expression.

[3] Note that we are closely following the MIP(VU) procedure in formulating the metaphor detection task in terms of labelling a specific lexical item as a *metaphor-related word* or not.

[4] As (Gandrabur et al., 2006) note, there is a formal requirement that $\sum_y P(x|y) = 1$. However, they point out that this requirement models probability of correctness as dependent on the sheer number of correct answers y for a given input x, yet a specific machine translation task may have many *equally* correct solutions. Rather, what they need is the kind of *uniform* interpretation, "independent of x", given by correctness probabilities.

Presently for our illness data set, the initial manual annotation study has been completed, results have been analysed and a follow-up study is currently underway. Automatic annotation for this initial data set is currently underway, to be completed by mid-2014. Work on collecting the conflict data set is finished, and the data is currently being manually annotated. The MIP(VU) procedure has given rise to the Amsterdam Metaphor Corpus,[5] and we will employ this corpus to evaluate the results of our procedure.

3. Results

3.1. Corpus collection

Searching the Illness Discussion corpus, we found metaphor types such as those reported by Skelton et al. (2002), such as BODY AS MACHINE, DOCTOR AS CONTROLLER, ILLNESS AS ATTACK, but also novel modifications of these types, such as PATIENT IS A CONTROLLER (e.g. *What most people do to control type 2 diabetes actually makes their blood sugar get worse!*), as well as completely fresh metaphors, such as what might be dubbed ILLNESS IS A RIDE (e.g. *the diabetes roller-coaster*).

For our initial study across illness topics, we selected four conceptual metaphors which have relevance to the illness domain (Hidalgo Downing and Kraljevic Mujic, 2009; Skelton et al., 2002): (a) BODY AS MACHINE, (b) DOCTOR AS CONTROLLER, (c) PATIENT AS CONTROLLER, (d) ILLNESS AS ATTACK. In Table (1), we compare the occurrence of these metaphors crossed by domains.

Table (1) reports initial results for frequency of metaphor types for different illnesses. A Pearson chi-squared test on this data yields $\chi^2 = (p < .005, df = 9)$, suggesting metaphor is not independent of domain of illness.[6] To interpret these results, note that relative size of the value in a cell in Table I (indicated by standardized residuals in brackets) suggests relative contribution to the overall chi-squared value. For example, comparing standardized residuals for table cells, we could say that while we can be confident that a natural-seeming metaphor about stress is ILLNESS IS AN ATTACK (e.g. *stress attack* is quite common), this is not the case for diabetes (e.g. *diabetic attack* is far less common).

3.2. Toward an automatic annotation procedure

Another aspect of our work is the development of tools for at least partially automating the task of identifying metaphorical expressions in a corpus. The work in the broader area of metaphor identification is vast.[7] While the diversity and scope of such approaches is daunting, one common thread in much previous work in this area has been that metaphor is essentially a word-level phenomena, which is to say, individual words are seen to be the bearers of metaphorical meaning. However, there is reason to think that perhaps metaphor may also emerge as a sentence-level phenomena in at least some cases (e.g. (Dunn, 2013)). In our work, one key question we have been pursuing is what, indeed, is the optimal level on which to focus our efforts to detect metaphorical phenomena.

In line with others (e.g. (Dunn, 2013), (Shutova et al., 2013)), we are making use of the Amsterdam Metaphor Corpus, from VU University Amsterdam (hereafter, VUAMC) (Steen et al., 2010), one of the larger collections of manually annotated metaphorical expressions. This corpus consists of around 190,000 lexical units, drawn from academic texts, conversation, fiction, and news texts, and was built using the MIP(VU), outlined in Section (). While the VUAMC labels metaphors directly at word-level, it presents these words within their sentential contexts, and we have been investigating whether there might be a way of factoring in such contextual information.

When carrying out detection kinds of tasks, such as the one we are engaged in developing a solution for, a well-established approach is to build models for predicting new instances of the phenomena in question, which in our case are metaphorical expressions. This may often involve in-depth investigation of the features characteristic of the target phenomena, and then somehow building a model (manually or automatically) to detect this phenomena via such features. However, assembling the "correct" set of features is rather difficult. Fortunately, some very good work on this already exists, in particular, we have been exploring a proposal made by (Tsvetkov et al., 2013), who draw on a range of interesting resources for doing metaphor detection work. One particular suggestion they have made is to employ features such as level of concreteness, given an established insight is that metaphor frequently involves the combination of concrete predicates with more abstract arguments (recall discussion of this in Section (2.2.) above). Following such suggestions by (Tsvetkov et al., 2013), we have begun making use of the MRC Psycholinguistic Database[8] (Wilson, 1988), a dictionary of 150,837 words, with different subsets of these words having been rated by human subjects in psycholinguistic experiments. Of special note, the database includes 4,295 words rated with degrees of abstractness, these ratings ranging from 158 (meaning *highly abstract*) to 670 (meaning *highly concrete*), and also 9,240 words rated for degrees of imageability, which is taken to indicate how easily a word can evoke mental imagery, these ratings also ranging between 100 and 700 (a higher score indicating greater imageability).

We have used the MRC concreteness and imageability scores, to mark up words from the VUAMC. Figure (1) shows the distribution of concreteness and imageability scores for literal vs. nonliteral words from the VUAMC, and Table (2) shows global averages for literal and nonliteral words which we have are using in our work on the VUAMC. Note that Figure (1) suggests concreteness and imageability do indeed interact with nonliteral meaning, with many of the VUAMC words annotated as metaphorical also having MRC scores indicating lower concreteness and imageability. Further, the results in Table (2), indicate

[5] See: http://www2.let.vu.nl/oz/metaphorlab/metcor/search/. Downloadable from: http://ota.ahds.ac.uk/desc/2541.

[6] With H0 *metaphor is independent of domain of illness*.

[7] Good representative collections of recent work includes (Feldman and Lu, 2007) and (Ekaterina Shutova et al., 2013).

[8] See: http://ota.oucs.ox.ac.uk/headers/1054.xml

Metaphor	Diabetes	Infection	Cancer	Stress	ROWS
A	28(-.4)	9(2.3)	11(2.5)	5(-2.3)	53
B	3(.5)	0(-.6)	1(1.)	0(-1.)	4
C	117(3.8)	4(-2.2)	7(-1.9)	18(-3.1)	146
D	8(-5.2)	9(1.4)	8(.4)	47(6.7)	72
COLUMNS	156	22	27	70	**275**

Table 1: Frequency of metaphor types for different illnesses (including standardised residuals in brackets), in online discussion forums (see text for details).

	Nonliteral			**Literal**	
P.O.S.	Imageability	Concreteness	P.O.S.	Imageability	Concreteness
Adverb:	342	310	Adverb:	344	294
Past participle:	463	475	Past participle:	431	419
Conjunction:	227	206	Conjunction:	226	214
Preposition:	314	266	Preposition:	314	266
Verb:	289	241	Verb:	303	264
Interjection:	346	304	Interjection:	346	304
Adjective:	437	370	Adjective:	409	337
Noun:	402	275	Noun:	464	335

Table 2: Global averages for concreteness and imageability by parts-of-speech, for nonliteral and literal words from the VUAMC.

that nouns and verbs typically have lower concreteness and imageability than other parts of speech. Putting both these results together, we decided to focus initially on nouns and verbs for metaphor detection.

However, we found that relatively few of the words in the VUAMC are represented in the MRC database, not unexpected given the relatively small numbers of words in both. Working toward an initial strategy for smoothing such sparsity, we have tried to factor in the sentential context. Two main ideas informed the strategy we developed: (1) we reasoned that metaphorical meanings in a sentence are likely to some degree to affect the meaning of most, if not all, words in that sentence, and (2) we recalled that there are often striking similarities between literal/metaphorically words from the same part-of-speech.[9] Consequently, for those words from the VUAMC lacking MRC scores, we assign a value drawn from the global average for their respective part-of-speech. While this particular smoothing strategy is in some respects rather crude, and we are refining this and other strategies based on ongoing empirical work, it has in practice turned out to be surprisingly effective.

In order to determine whether MRC scores could be used as predictor features, for carrying out metaphor detection, we are currently evaluating models for predicting metaphorical expressions from the VUAMC. By way of reporting on progress with this, the most useful approach we have tried so far was in some respects the simplest: a simple logistic regression model, for making the binary decision that some word encountered is **either** literal **or else** it is nonliteral. For an initial study, we built 5 models using this learning algorithm, based on distinct linear combinations of predictor variables, trained over a representative sample of 90% of a set of data from the VUAMC, and the resulting models were tested on the remaining 10% of the data, which was held out for this purpose. The combinations of predictor variables used, as well as the results for the percentage success of these models in predicting our test set is as follows (citing accuracy scores only): (1) “concreteness + imageability + part-of-speech” = 66%, (2) “concreteness + part-of-speech” = 69%, (3) “imageability + part-of-speech” = 72%, (4) “concreteness” = 56%, (5) “imageability” = 59%. While these early results are not overly impressive, our initial investigation has served the purpose of discovering a baseline, i.e. model (3) combining imageability and part-of-speech. We intend following this study up with a full-scale learning experiment in Spring 2014.

4. Conclusion

The results reported here demonstrate the value of the database of metaphors in Illness and Conflict Discussion we are building. The key component of these results is that domain of illness and type of metaphor are not independent, which is to say, some metaphors are more typically used in certain domains than other domains. We have already begun exploiting this result to support our work on generating metaphor (Gargett and Barnden, 2014). Yet these results are also of broader relevance to work on detecting metaphor, especially in discourse on illness, and early results for our work on conflict discourse suggests similarly promising results. The yield of this component of our overall project will be a searchable database of metaphorical expressions for illness and conflict discourse.

[9] There is some detailed discussion of this latter idea in (Pragglejaz Group, 2007).

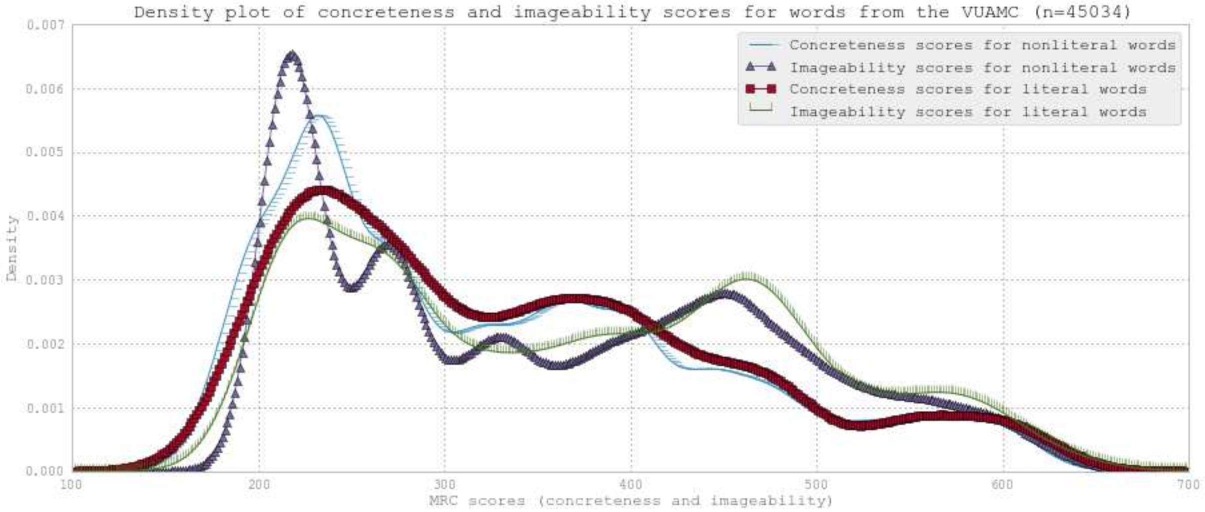

Figure 1: Density graph of number of occurrences of literal/metaphorical words in the VUAMC vs. their concreteness/imageability (estimating the probability density function of concreteness and imageability scores)

A key problem we faced in developing this corpus has been the use of conceptual metaphors, and such difficulties have been noted elsewhere (e.g. (Shutova et al., 2013)). To address these and other issues, we are exploring the formulaic expression of metaphorical utterances, and so the close link between the conceptual and linguistic levels posited by conceptual metaphor theory, seems to us very much worth pursuing.

Finally, while at this stage we have only examined concreteness and imageability as features for automatically detecting metaphor, we are exploring a range of such features (including similarity measures, patterns of polysemy, etc). However, it would seem that certainly imageability, and perhaps also concreteness, are useful features for predicting metaphorical use of language. While we are employing this for detecting metaphor, we are also anticipating being able to redeploy it within our broader project of generating metaphor.[10]

5. Acknowledgements

We acknowledge financial support through a Marie Curie International Incoming Fellowship (project 330569) awarded to both authors (A.G. as fellow, J.B. as P.I.).

6. References

Balahur, A., van der Goot, E., and Montoyo, A., editors. (2013). *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Atlanta, Georgia, June.

Birke, J. and Sarkar, A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL*.

Deignan, A. (2006). The grammar of linguistic metaphors. In Stefanowitsch, A. and Gries, S., editors, *Corpus-based approaches to metaphor and metonymy*, pages 106–122. Walter de Gruyter.

Deignan, A. (2008). Corpus linguistics and metaphor. In Gibbs, R., editor, *The Cambridge Handbook of Metaphor and Thought*, Cambridge Handbooks in Psychology, pages 280–294. Cambridge University Press, Cambridge.

Dorst, A. G., Reijnierse, W. G., and Venhuizen, G. (2013). One small step for mip towards automated metaphor identification?: Formulating general rules to determine basic meanings in large-scale approaches to metaphor. *Metaphor and the Social World*, 3(1):77–99.

Dunn, J. (2013). How linguistic structure influences and helps to predict metaphoric meaning. *Cognitive Linguistics*, 24(1):33–66.

Ekaterina Shutova, University of California at Berkeley, U., Beata Beigman Klebanov, Educational Testing Service, U., Joel Tetreault, Nuance, U., and Zornitsa Kozareva, USC Information Sciences Institute, U., editors. (2013). *Proceedings of the First Workshop on Metaphor in NLP*. Association for Computational Linguistics, Atlanta, Georgia, June.

Feldman, A. and Lu, X. (2007). *Proceedings of the North-American Chapter of Association for Computational Linguistics -Human Language Technologies (NAACL-HLT 2007) Workshop on Computational Approaches to Figurative Language*. The Association for Computational Linguistics, Rochester, NY, USA.

Gandrabur, S., Foster, G., and Lapalme, G. (2006). Confidence estimation for nlp applications. *ACM Transactions on Speech and Language Processing (TSLP)*, 3(3):1–29.

Gargett, A. and Barnden, J. (2014). Gen-meta: Generating metaphors by combining ai and corpus-based modeling. In *Web Intelligence and Agent Systems: An International Journal*.

Hidalgo Downing, L. and Kraljevic Mujic, B. (2009). Infectious diseases are sleeping monsters: Conventional and culturally adapted new metaphors in a corpus of abstracts on immunology. *Ibérica*, 17:61–82.

Levinson, S. C. (2006). On the human ‘interaction engine’.

[10] For details of our project, Gen-Meta, see: http://www.cs.bham.ac.uk/ gargetad/genmeta-index.html.

Roots of human sociality: Culture, cognition and interaction, pages 39–69.

Li, L. and Sporleder, C. (2010). Linguistic cues for distinguishing literal and non-literal usages. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 683–691. Association for Computational Linguistics.

Mason, Z. J. (2004). Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.

Pragglejaz Group. (2007). Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Shutova, E., Teufel, S., and Korhonen, A. (2013). Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Skelton, J. R., Wearn, A. M., and Hobbs, F. R. (2002). A concordance-based study of metaphoric expressions used by general practitioners and patients in consultation. *The British Journal of General Practice*, 52(475):114–118.

Steen, G., Dorst, A., Herrmann, J., Kaal, A., and Krennmayr, T. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Convering Evidence in Language and Communication Research. John Benjamins Publishing Company.

Tsvetkov, Y., Mukomel, E., and Gershman, A. (2013). Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, Atlanta, Georgia, June. Association for Computational Linguistics.

Wilson, M. (1988). Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.