"Computational Metaphor Identification"

by

Eric P.S. Baumer, David Hubin
and Bill Tomlinson

Technical Report LUCI-2010-002
http://luci.ics.uci.edu

Computational Metaphor Identification

Eric P. S. Baumer
Department of Informatics, University of California, Irvine

David Hubin
Department of Computer Science, University of California, Irvine

Bill Tomlinson
Department of Informatics, University of California, Irvine

Conceptual metaphors are pivotal to human cognition, but most previous computational linguistics treatments of metaphor focus on discerning a metaphor's literal meaning. Instead, this article presents computational metaphor identification (CMI), a technique for identifying potential conceptual metaphors in written text. This technique draws on and extends previous related work in cognitive linguistics and computational linguistics. CMI hinges on mapping selectional preferences from a source corpus to a target corpus in order to identify metaphorical mappings. Example results are presented and then evaluated via two methods: comparison with expert linguistic analysis, and assessment by non-expert human subjects. The results show that CMI is an effective means for identifying conceptual metaphors; computationally identified metaphors are shown to be conceptually similar to those identified in previous expert linguistic analysis, and confidence scores assigned by the system to identified metaphors correlate significantly with non-expert subjects' assessments. This work represents a novel direction, both for computational linguistics research on metaphor, and for artificial intelligence research more broadly.

1. Introduction

Metaphor plays a crucial role in human understanding of the world. The term metaphor here refers to the partial framing of a target concept or set of experiences in terms of a source concept (Lakoff and Johnson 1980; Lakoff 1993). For example, consider the ways in which an argument might be described: "he *attacked* your *position*," "her points are well *defended*," "I *destroyed* his claims." Such phrases are instances of the conceptual metaphor ARGUMENT IS WAR[1], that is, we often frame our experience of being in an argument with someone as if we were at war with that person. Such metaphors pervade myriad aspects of human thought and activity, but their ubiquity can make them difficult to notice (Lakoff and Johnson 1980).

This article presents **computational metaphor identification** (CMI), a technique for identifying potential conceptual metaphors in written text. In contrast to most previous computational linguistics work on metaphor, the goal here is not to discern metaphorical from non-metaphorical phrases, but rather to identify overarching conceptual metaphors that permeate a text. It is designed to draw attention to linguistic

[1] This article uses SMALL CAPS for conceptual metaphors, *italics* for source or target concepts within a metaphor, ALL CAPS for source or target domains from which concepts are drawn, and "quotes" for quotation of example words or phrases.

patterns indicative of potential metaphors, thereby helping to foster critical and creative thinking about those metaphors.

While the work presented in this article draws on and extends techniques from computational linguistics, machine learning, and artificial intelligence, the approach here represents an inversion of classical AI. Most artificial intelligence research asks, "Can people make computers think?" That is, can humans "creat[e] machines that perform functions that require intelligence when performed by people" (Kurzweil 1990). Similar definitions include "the science of making machines do things that would require intelligence if done by men" (Minsky 1968, page v), and "the study of how to make computers do things at which, at the moment, people are better" (Rich and Knight 1991). In contrast, the research presented here offers an inversion of the classical AI approach, instead asking, "Can computers make people think?" Rather than using humans as a model for how to build intelligent computational systems, we might instead be able to develop computational systems that encourage human users to think in new and different ways or approach familiar concepts from novel alternative perspectives.

The remainder of this article reviews the relevant literature on conceptual metaphor from cognitive linguistics and from computational linguistics (Section 2); describes the technical details of the computational metaphor identification implementation (Section 3); provides example results of metaphors identified in a corpus of political blogs (Section 4); evaluates CMI by comparing computationally identified metaphors with expert linguistic analysis (Section 5.1), and by subjective human rating of identified metaphors (Section 5.2); and provides a number of potential future directions, both in terms of improvements to the CMI technique and in terms of possible applications of CMI or components thereof (Section 6).

2. Related Work

2.1 Conceptual Metaphor Theory

The technique presented in this article draws largely on the work of Lakoff and colleagues (Lakoff 1993; Lakoff and Turner 1989; Lakoff and Johnson 1980), who argue that metaphor is not a linguistic or poetic device, but rather is fundamental to human cognition. For example, when discussing time, one might say "you're *wasting* my time," "this gadget will *save* you hours," "that flat tire *cost* me an hour," or "you need to *budget* your time" (Lakoff & Johnson, 1980, pp. 7-8). Lakoff and Johnson claim that such linguistic patterns evidence the conceptual metaphor TIME IS A RESOURCE, that we understand the abstract concept of time partially in terms of our everyday experiences with physical resources. We use words from our experiences with physical resources to talk about time because the cognitive structure of the metaphor "sanctions the use of source domain language and inference patterns for the target domain" (Lakoff and Turner 1989, page 208). This is not to say that conceptual metaphor is primarily a linguistic phenomenon. Rather, the linguistic patterns serve as evidence for the cognitive phenomenon.

The metaphor TIME IS A RESOURCE "emerged naturally in our culture because of the way we view work, our passion for quantification, and our obsession with purposeful ends" (Lakoff and Johnson 1980, page 67). However, in other cultures where this metaphor is either less common or nonexistent, it would be absurd, if not impossible, to speak of "wasting time," because time does not belong to a category of things that can be wasted. This possibility of multiple, potentially conflicting metaphors for the same target concept is referred to as metaphorical pluralism, and is a central tenet is conceptual metaphor theory. As another example, a cornucopia of metaphors can be used from the concept of *love*, such as such as LOVE IS A JOURNEY: "this relationship is[n't] *going anywhere*;" LOVE IS MADNESS: "I'm just *wild* about Harry;" or LOVE IS MAGIC: "she is *bewitching*" (Lakoff and Johnson 1980, page 44,49). Each of these metaphors simultaneously highlights certain aspects of the experiences surrounding love while downplaying others. Lakoff and Johnson argue that "successful function in our daily lives seems to require a constant shifting of [many] metaphors... that are inconsistent with one another... to comprehend the details of our daily existence" (Lakoff and Johnson 1980, page 221). Moreover, suggestion of an alternative, novel metaphor can provide a reconceptualization that draws our attention to different aspects of the situation, that can "cause us to try to understand how [the novel metaphor] could be true, [and that] makes possible a new understanding of our lives" (Lakoff and Johnson 1980, page 175). CMI is designed to promote such critical reflection by identifying particular linguistic patterns and presenting a reader with the metaphors those patterns might imply.

One area where metaphorical framings can carry significant consequence is in politics. In analyzing political discourse from popular media, Howe (1988) describes how political metaphors often draw on the source domains of SPORTS and WAR. For example, a political party can be seen as a team, with individual politicians "joining" the team, "captaining" the team, being a "team player," etc. Similarly, a political party can be seen as an army, where individual politicians are soldiers or officers on their respective armies, and elections are battles or wars between these armies. Such sports and war metaphors emphasize conflict, but they also downplay the importance of negotiation and compromise. Despite their importance and influence, such metaphors often go unquestioned; "because they can be used so automatically and effortlessly, we find it hard to question them, if we can even notice them" in the first place (Lakoff and Turner 1989, page 65). One of the primary goals of the work presented here, then, is to draw potential metaphors to readers' attention in order to encourage consideration of what a given metaphor highlights, what it hides, and what alternative metaphors might be frame the situation differently.

2.2 Metaphor in Computational Linguistics

While a significant amount of previous computational linguistics research has dealt with metaphor, most such work focused on discerning whether individual phrases were literal or metaphorical so as to apply additional processing to the metaphorical phrases to determine their literal meaning. I argue here that such techniques, while valuable in accomplishing the goals for which they were intended, are not as useful for the purposes of the system described in this article.

For example, MIDAS (Martin 1990) describes a help system for the UNIX command line that can properly interpret metaphorical language in questions asked by users. For example, if a user types, "how do I enter Emacs?" MIDAS would determine that "enter" in this case is an instance of a container metaphor, and that to enter a program means to invoke the program. Furthermore, MIDAS can learn new metaphors by extending its current knowledge, for example, determining that when a user asks, "how do I kill a process?" s/he wants to know how to terminate a process. In such cases, MIDAS first attempts a literal interpretation, then, when no literal sense of "kill" can be used to derive a satisfactory interpretation, it searches for an alternate interpretation. This approach was extended to the development of Metabank (Martin 1994), a large database of metaphorical interpretations for English metaphors.

Another system, met* (Fass 1991), can detect both metonymies and metaphors using violations of selectional restriction rules encoded in a knowledge representation. For example, "eat" and "drink" both select for *animal* as the agent; similarly, "eat" prefers edible foods as the object, and "drink" prefers potable liquids. The phrase, "my car drinks gasoline," violates the preference rules for drink, i.e., no suitable literal interpretation can be derived. Using its knowledge representation, met* initially tries to find some metonymic relationship between *car* and *animal*; for example, if a car were a part-of an animal, the phrase could be interpreted as a PART for WHOLE metonymy. When unable to find such a metonymic relationship, met* derives an analogical mapping from its knowledge about cars and about animals.

Some more recent approaches do not require the use of extensive knowledge representations. For example, Gedigian et al. (2006) utilize corpora in which metaphorical uses are annotated by hand to train an automatic classifier. Krishmakumaran and Zhu (2007) examine the relationships of verbs and adjectives with nouns to find instances that violate standard expectations in WordNet (Fellbaum 1998) for example, "he is a brave lion," would be considered metaphorical, because "he," taken to mean a "person," is not a WordNet hyponym of "lion."

Each of the above cases subscribe in various ways to the literal meaning hypothesis (Reddy and Ortony 1979). This hypothesis asserts that every sentence has a literal meaning, as derived from the literal meanings contained in each of the sentence's constituent words. However, some sentences also have a figurative meaning, drawing on a larger context beyond the words themselves. In this view, a figurative interpretation of a sentence is sought only after a literal interpretation has been formed and found inconsistent, nonsensical, or otherwise faulty (Black 1962; Searle and Ortony 1979). However, Gibbs (1984) and Gentner et al. (2001) review several studies suggesting not only that people do not attempt to derive a literal meaning before a metaphorical meaning, but that, in many cases, a literal meaning is not sought at all. Even distinguishing whether a given expression is literal or metaphorical can be difficult at best (Fass 1991; Gibbs 1984; Ortony 1980). For example, "the rock is becoming brittle with age" (Reddy 1969, page 242), has "a literal interpretation when uttered about a stone and a metaphorical one when said about a decrepit professor emeritus" (Fass 1991, pgae 54).

Consider further the following example. The phrase, "he won the argument," is almost certainly literal. In, "the conservative debater defeated her liberal opponent," there is somewhat less certainty. The terms "defeated" and "opponent" allude rather more strongly to physical combat than does the term "win," but the former are still within the realm of what might be considered customary or literal usage. The phrase, "with her final point, she delivered a crushing blow and destroyed her opponent's argument," would almost surely be considered more metaphorical than literal; the first two examples exhibit language that hardly seems contextually anomalous, cf. (Ortony 1980), whereas the third seems highly contextually anomalous. However, each of these three phrases represents an instantiation the metaphor argument is war, and focusing on whether each should be considered literal or figurative belies this larger conceptual structure. Making such distinctions, then, is inherently at odds with the goals of the system presented here.

Another line of research has taken a slightly different approach by atteptmting to identify analogical mappings in written text, drawing largely on structure-mapping theory (Gentner 1983) and the structure mapping engine (Falkenhainer, Forbus, and Gentner 1989). The general approach involves using a textual corpus as input and producing set of analogical mappings as output. Kuehne and Forbus (2004) developed a technique for extracting structured representations from restricted subsets of English, which has been used as input for various analogical reasoning techniques, e.g., moral decision-making (Dehghani et al. 2008). Other approaches have combined natural language with annotated sketches as input (Lockwood and Forbus 2009). While effective at certain analogical reasoning tasks, because of their reliance on restricted forms of English, such techniques are not amenable to the goals in this article.

Following a slightly different approach, Turney (2008) presents the latern relation mapping engine (LRME) for identifying analogies in written text. This system uses latent relation analysis (Turney 2006), based on the premise that co-occurrence implies semantic relatedness, to identify semantic relations, then finds mappings based on co-occurrences matrices. such a technique could be useful for the goals here; indeed, Turney (2008) suggests the possibility of combining his technique with those similar to the one presented in this article. however, LRME requires a massive amount of data (corpora on the order of 10^{10} words) and does not incorporate linguistic relations, i.e., it identifies mappings based on relatedness of different words (as derived from co-occurrence) but not based on different types of relations.

One previous system does align well with the goals in this article. CorMet (Mason 2004) is designed to extract known conventional metaphors from domain-specific textual corpora. These corpora consist of documents returned by Google queries with a conjunction of domain-specific keywords and a particular verb. CorMet then calculates selectional preferences and associations (Resnik 1993) for each corpus's characteristic verbs, i.e., those verbs at least twice as frequent in the corpus as in general English. Based on these selectional associations, CorMet clusters the nouns for which the characteristic verbs select. To identify metaphors, mappings are sought from clusters in the source corpus to clusters in the target corpus, based on the degree to which the same verbs select for members of both clusters. For example, CorMet was used to identify a mapping corresponding to the metaphor MONEY IS A LIQUID by mapping from a cluster for the concept *liquid* in a corpus for the domain LABORATORY to a cluster for the concept *money* in a corpus for the domain FINANCE , based on the selectional associations of verbs such as "pour," "flow," "freeze," and "evaporate" for components of each cluster in their respective domains.

The computational metaphor identification system presented here draws largely on the techniques used in CorMet, and Section 3 describes how CMI extends and refines those techniques in a number of ways. However, CMI is not simply an improved version of CorMet; there are three important differences between CorMet and CMI. First, CorMet is designed to extract conventional metaphors, i.e., metaphors of whose presence we are already aware. In contrast, CMI seeks to identify potential metaphors of whose presence we may or may not be aware, i.e., CMI does not require *a priori* decisions about what metaphors one expects to find. Second, CMI includes a number of features not present in, and improvements upon, CorMet, as noted throughout this article. Third, by all appearances, CorMet is designed purely as an analytic tool. In contrast, CMI is designed largely for the goal of fostering critical and creative thinking about metaphors, i.e., it has a specific intended use beyond abstract analysis. Previous studies have demonstrated that CMI can be used to foster critical and creative thinking about conceptual metaphors (Baumer et al. 2009; Baumer, Sinclair, and Tomlinson 2010; Baumer et al. 2009). The primary contribution of this article is in provide the full implementation details of CMI (see Section 3), as well as an evaluation thereof (see Section 5).

3. Computational Metaphor Identification

The computational metaphor identification system presented in this article consists of four main phases: corpus preparation, source suggestion, finding mappings, and improving metaphor legibility. Corpus preparation is done on any potential source or target corpus; the preparation process is the same for any corpus, regardless of whether it is a source or target. Source suggestion, an optional intermediate step, compares a prepared target corpus with several potential prepared source corpora and suggests which corpora might lead to informative metaphorical mappings. Generally speaking, source suggestion can provide a rough indication of the types of metaphor that may be employed in a target corpus, i.e., on what source domains metaphors might draw. Potential metaphors are identified by mapping specific linguistic patterns from a prepared source corpus to a prepared target corpus. Identified mappings are then further processed to make them more legible as metaphors to human readers. This section describes each of these phases in turn.

3.1 Corpus Preparation

Since conceptual metaphors map from concepts in a source domain to those in a target domain, CMI works by mapping from linguistic patterns in a source corpus to similar patterns in a target corpus. The target corpus is the text in which potential metaphors are to be identified, and the source corpus is a text pertaining to some coherent source domain for those metaphors. For example, if we wish to identify metaphors in political news that draw on military concepts–armies, generals, tactics, battles, tanks, invasions, etc.–the target corpus would be a collection of articles covering politics, and the source corpus would be some collection of documents about military people and events.

As mentioned above, the preparation process for any source or target corpus is the same. Preparation includes five main steps: acquiring corpora, parsing, finding characteristic nouns, selectional preference learning, and synset clustering, each of which are explained in detail below.

3.1.1 Acquiring Corpora. Acquiring a target corpus is relatively straightforward; simply collect the documents in which metaphors are to be identified. Acquiring source corpora, on the other hand, is somewhat more complex. A source corpus should consist of a large number of documents about a single conceptual domain, where a domain consists of the concepts and inferences associated with a set of related experiences, cf. (Gentner 1983). While the techniques described here are applicable to any source corpus, the system presented in this article uses articles from an April 2007 snapshot of the English Wikipedia (http://en.wikipedia.org), as it provides a large, readily available, hierarchically categorized body of content from a wide variety of domains.

Domain-specific corpora are built using Wikipedia's categorization scheme. Each article in Wikipedia belongs to at least one category. Furthermore, categories can have subcategories, forming a directed acyclic graph of categories and subcategories. To build the corpus for a given domain, CMI collects all the articles in the category for that domain and all of the category's immediate subcategories. For example, the source corpus for the MILITARY domain consists of all the articles in the _Military_[2] category, such as _Armed_forces_1, _Demilitarization_, and _Military_command_, as well as articles from all of _Military_'s subcategories, such as _Military_art_, _Military_history_, and _Military_personnel_. Subcategory recursion is only one level deep. For example, the MILITARY corpus does not include articles from the subcategories of _Military_personnel_, such as _Military_leaders_ or _Military_careers_.

[2] For clarity, Wikipedia articles and categories are written using _Initial_capital_with_underscore_accents_.

It is not necessarily the case that recursing fully through subcategories would be beneficial. For example, if building a corpus for the SCIENCE domain, one might use Wikipedia's _Science_ category, which has as a subcategory _Scientists_, which has as a subcategory _Astronomers_, which has as a subcategory _Fictional_astronomers_. This last category contains two articles, one on _Comet_Man_, a character from comic books published by Marvel Comics, the other on _Trillian_, a character from Douglas Adams' *The Hitchhiker's Guide to the Galaxy* series. While it may be the case that these two characters are fictional astronomers, the content of the articles about them have little to do with science in general. Thus, fully recursing through all subcategories may end up giving noisy data that are not much better than those used in CorMet (described in the following paragraph). Given the dynamic and somewhat irregular nature of Wikipedia's categorization scheme, the approach of using a single level of subcategories used here serves as a useful heuristic. While a method of automatically determining the optimal depth at which to choose a Wikipedia category that corresponds to a given domain might be useful, such a technique is beyond the scope of this work. Indeed, the computational techniques presented in this article do not strictly depend on using Wikipedia as a source corpus; it is not the only possible method for acquiring source corpora, but it effectively serves its purpose here.

In some instances, the most appropriate category for a given domain may not be obvious. For example, for the SPORTS domain, one might use Wikipedia's _Sports_ category. However, _Sports_ is a very high level category, such that its articles, and those of its subcategories, do not often discuss specific sports; the article for _Baseball_ is in the category _Baseball_, which is a subcategory of, among others, the categories _Team_sports_, _Ball_games_, and _Olympic_sports_, each of which is a subcategory of _Sports_. If the _Sports_ category were used, these second-level subcategories, such as baseball, that contain detailed descriptions of specific sports would not be included. Therefore, this implementation instead uses, for example, the _Olympic_sports_ category for the SPORTS domain.

This process of corpus acquisition improves upon that of CorMet, which queries Google using conjunctions of randomly selected subsets of domain-specific keywords with specific verbs. On average, only about 75% of the documents returned by such a query are relevant (Mason 2004), as based on a random sample from each corpus hand-evaluated for relevance; in some cases, documents may be relevant, but not in the way intended. Mason cites an example, looking for mappings from ARCHITECTURE to THEORY, wherein most documents returned for architecture queries were about zoning laws or planning rather than the buildings themselves, and most of the documents ostensibly about theory were calls for papers or university department pages.

In contrast, the method used here yields corpora from which about 80% of documents are directly relevant. This statistic is based on random samples of 50 documents from 7 different category-domain pairs; Table 1 lists the percent relevance for each of these. Irrelevant articles took many forms: some from _Cities_ were about ancient empires based around a single city, while some from _Military_ and _Science_ discussed military or science museums, respectively, rather than the conduct of military activities or scientific practice. While this method of corpus acquisition seems to provide only a slight improvement over that of CorMet (Mason 2004), the method used here does not require pre-specified verbs to constrain the results, leading to more representative, naturally occurring corpora.

Furthermore, greater specificity may be achieved with this technique by experimenting with selecting different Wikipedia categories. For example, one of _Architecture_'s subcategories is _Urban_studies_and_planning_, which gives the ARCHITEC-

Table 1
Percentages of articles from several Wikipedia categories relevant to the intended domain.

Domain	Category	% Relevant
ARCHITECTURE	_Architecture_	74%
CITY	_Cities_	96%
FINANCE	_Finance_	72%
MEDICINE	_Medicine_	80%
MILITARY	_Military_	90%
SCIENCE	_Science_	84%
SPORTS	_Olympic_sports_	74%
WAR	_War_	70%
Average	n/a	80.0%

Table 2
Sizes of two source corpora and one target corpus.

Corpus	Documents	Sentences	Words
MILITARY	1,769	24,791	505,205
SCIENCE	3,485	89,010	1,749,925
Democratic Blogs	728	9,740	149,967

TURE corpus many articles about city planning and the affects of urban living. If instead one wanted a corpus strictly about buildings or their construction, one could use the _Buildings_and_structures_ category or the _Construction_ category, respectively. The previous approach of using an arbitrary list of keywords may, as Mason (2004) claims, allow more flexibility in defining domains, but it can also lead to irrelevant documents in the corpus. While he argues that one of CorMet's strengths is its ability to tolerate noisy data, the informal relevance statistics here suggest that better results are achieved by using documents that are designated in advance as relevant to a certain domain.

Table 2 describes the sizes of a few sample corpora. The first two lines describe source corpora derived from Wikipedia. The MILITARY corpus contains those articles in the _Military_ category, as well as articles in any of the _Military_ category's subcategories; the SCIENCE corpus comes similarly from the _Science_ category. For Wikipedia, each article is treated as a single document. These corpora are used throughout this article as example source corpora. The third line describes a sample target corpus of posts from nine democratic or left-leaning political blogs during the 2008 US presidential election (see Appendix 7 for a list of blogs and URLs), specifically, from 6:00 a.m. on November 1 to 12:00 a.m. on November 5 (the election was held on November 4). These blogs were chosen due to their high authority rating on Technorati (http://technorati.com), a service that tracks and ranks blogs. Blog posts were acquired from the blogs' RSS feeds, and HTML tags were stripped. For these data, each blog post is treated as a single document. This corpus is used throughout this article as an example target corpus.

These size data give a rough impression of the three corpora described. Wikipedia articles in these two corpora tend to be relatively long, 429 words per article, with longer

Table 3
Grammatical relation counts and ratios for some example corpora.

Corpus	Grammatical Relns	Relns pre Sentence	Relns per Word
MILITARY	423,358	17.1	0.838
SCIENCE	1,446,739	16.3	0.827
Democratic Blogs	120,640	12.4	0.804

sentences, 20 words per sentences. The blog corpus is, unsurprisingly, less wordy, with 206 words per post and 15 words per sentence.

3.1.2 Parsing. Each document in the source and target corpora is parsed using the Stanford parser (Klein and Manning 2002, 2003) as well as a typed dependency parser (de Marneffe, MacCartney, and Manning 2006). Parsing helps with finding characteristic nouns, and typed dependencies are used in selectional preference learning, both of which are described below.

Table 3 describes the number of grammatical relations in each corpus, as well as the average number of grammatical relations per sentence. Again, we see that the Wikipedia corpus has a slightly higher degree of complexity, with more relations per sentence and slightly more relations per word than the blogs.

3.1.3 Finding Characteristic Nouns. Since metaphors map between salient concepts, CMI attempts to determine salient concepts in a corpus. This is done first by finding characteristic nouns, which are those nouns in a corpus with the highest frequency relative to their frequencies in common English, as derived from the British National Corpus (BNC) (Kilgarriff 2003). Nouns are listed in the frequency dictionary under, and looked up by, their lemmatized forms. A simple look-up table is used to convert from American English spellings to British English, e.g., from "center" and "vaporize," which are not in Kilgarriff's frequency dictionary, to "centre" and "vapourise," which are. The frequency dictionary only lists those words that appear at least 800 times in the 1.0e8-word BNC. Frequencies for unlisted words are estimated at 400 / 1.0e8 = 4.0e-6. For any given source or target corpus, CMI uses only those nouns that are twice as frequent in the corpus as in general English, up to a maximum of 400 characteristic nouns per corpus. This limit was chosen to keep computation tractable.

CMI's approach here differs from that of CorMet, which finds relatively frequent verbs by counting any word that WordNet[3] (Fellbaum 1998) says might be a verb as such. CorMet's approach is subject to nominal homonyms, e.g., "attack" could be either a verb or a noun. By parsing source and target corpora, CMI mostly avoids problems with such homonyms, because parsing a sentence determines the grammatical function of each word, indicating whether it is being used as a verb or a noun. The reason CMI finds characteristic nouns rather than characteristic verbs will be explained more fully below in the contexts of selectional preference learning and finding mappings.

[3] The version of CMI presented in this article uses a customized version of WordNet that included entries for candidates from the two major political parties in the 2008 US election, as well as candidates from some of the other well-known parties: Charles Baldwin, Robert Barr, Joe Biden, John McCain, Ralph Nader, Barack Obama, and Sarah Palin.

Table 4
Top 20 characteristic nouns from three example corpora with relative frequencies.

MILITARY		SCIENCE		Democratic Blogs	
Noun	Rel Freq	Noun	Rel Freq	Noun	Rel Freq
military	268.70	years	221.15	obama	1548.67
troops	175.67	datum	159.15	mccain	1385.30
operations	170.72	placebo	135.29	john	450.10
years	170.23	links	113.86	palin	406.76
combat	156.87	fields	110.58	today	335.07
warfare	127.67	things	106.58	polls	333.41
links	121.24	laws	99.43	years	285.06
infantry	118.76	effects	86.58	voting	243.39
arms	107.88	one	75.15	america	230.05
massacre	103.92	elements	69.57	senator	185.04
battalion	84.12	mechanics	69.00	one	185.04
salute	82.14	physicist	67.29	days	176.71
artillery	82.14	times	64.15	american	165.04
veteran	79.40	today	64.00	sarah	163.37
brat	76.70	humans	62.29	virginia	160.04
no.	75.71	something	59.15	r	145.03
days	75.71	parts	56.72	iraq	138.36
services	70.27	ways	55.57	d	136.70
ammunition	61.86	no.	55.29	numbers	135.03
em	59.38	simulation	51.43	ad	130.03

Table 4 lists the top 20 characteristic nouns for each of the three example corpora in order of decreasing relative frequency. For example, "troops" is 175.67 times more frequent in the MILITARY corpus than in general English.

These characteristic nouns align with naÃŕve expectations. A few of these nouns warrant comment. In both MILITARY and SCIENCE, "no." is an abbreviation for "number," which does not appear in the BNC. In MILITARY, "brat" refers to military brat, which is slang for the child of a military service person, and "em" is likely either an abbreviation for enlisted man or the slang ′em, which is short for "them." In the blogs, "american" is not an adjective but refers to a citizen of America; the full parse indicates that "american" is being used as a noun, not as an adjective. "r" and "d" are likely abbreviations for Republican and Democrat, respectively. Also, some words, such as "years," are likely plural forms of other words, here, "year." However, the morphology "years," referring to a long period of time, exists separately in WordNet, which causes "years," when observed in the corpus, to appear as if it is already in a lemmatized form rather than a plural of "year." Kilgarriff's (Kilgarriff 2003) frequency dictionary, on the other hand, has no separate entry for "years," making the common English frequency low and, subsequently, the relative frequency high.

3.1.4 Selectional Preference Learning. CMI next calculates selectional preferences and associations (Resnik 1993) for each characteristic noun. Before proceeding, the characteristic noun list is trimmed using a custom stopword list (see Appendix 7), which includes pronouns (e.g., he, she, it), prepositions, (e.g., among, between, from), some common verbs (e.g., be, do, get, made), and other words with nominal homonyms that the parser occasionally mistakes for nouns. Selectional preferences and associations are calculated for the trimmed characteristic noun list based on grammatical relations derived from the typed dependency parser (de Marneffe, MacCartney, and Manning 2006). The use of typed dependencies is a significant improvement over Mason's admittedly rough, error-prone heuristic of using templates to extract case frames. For example, "(S (NP & OBJ) (VP (were | was | got | get) (VP WORDFORM-PASSIVE)) [was] used to extract roles for passive, agentless sentencesâĂę where WORDFORM-PASSIVE is replaced by a passive form of [a] verb" and NP & OBJ is used to identify the object of that passive verb (Mason 2004, page 25). The typed depedency parser provides both more accuracy, in terms of assigning words to the proper grammatical relation, and more detail, in terms of the 48 different grammatical relations identified (de Marneffe, MacCartney, and Manning 2006).

Selectional preference is an overall measure of how selective a given class of nouns is for the verbs that appear in conjunction with it in a given grammatical relationship, and is calculated using the relative entropy of the poster distribution of verbs conditioned on a specific noun class and grammatical relation (i.e., typed dependency) with respect to the prior distribution of verbs in general English:

$$S(c) = \sum_{v} P\left(v|c\right) \log \frac{P\left(v|c\right)}{P\left(v\right)} \tag{1}$$

where c is a class of nouns and a grammatical relation, and v ranges over all the verbs for which c appears in the given grammatical relation. Selectional preference is then used to calculate selectional associations, which quantify the degree to which a given noun class and grammatical relation are associated with a particular verb:

$$\lambda\left(\text{c,v}\right) = \frac{1}{S\left(c\right)} P\left(v|c\right) \log \frac{P\left(v|c\right)}{P\left(v\right)} \tag{2}$$

with variables as above. Selectional association gives a measure of how much of the selectional preference for a given noun class and grammatical relation pair is due to each verb for which the pair selects. Individual selectional association numbers are theoretically unbounded, but because selectional association is normalized by selectional preference strength, the sum of all selectional associations for any single grammatical relation of a given noun class is 1.

Typically, selectional preferences and associations are calculated for verbs, that is, the verb is said to select for nouns in various relations. For example, the verb "eat" selects for "food" as its direct object. However, it can just as readily be said that nouns select for verbs (Light and Greiff 2002), for example, "food" selects to be the direct object of "eat."

Herein lies part of the justification for finding characteristic nouns, as in CMI, rather than characteristic verbs, as in CorMet. Characteristic verbs are indicative of the types of actions and relations described in a corpus, but the characteristic verbs do not necessarily select for characteristic nouns. In CMI, verbs mediate metaphorical mappings, but classes of nouns are what are actually being mapped. Based on early tests, finding mappings by starting with selectional preferences of characteristic verbs does not often describe the metaphorical framings of what might be considered the central concepts in a domain. Thus, the approach used here is based on learning selectional preferences and associations of characteristic noun classes.

In order to calculate selectional associations, distributions of verbs conditioned on noun classes are needed, but what is observed in the corpus are distributions conditioned on noun tokens. Resnik's (1993) approach counts each occurrence of a word as a partial observation of every word class it might represent, where word classes are WordNet (Fellbaum 1998) synsets. For example, the word "food" belongs to three different synsets in WordNet. Furthermore, the word "food" may be considered an instance of any hypernym of any of the word's synsets. Going up the hypernym path away from the synset of interest yields synsets that are progressively more general and less germane to the concept of interest. Thus, when finding potential synsets a given word might represent, CMI uses the first three hypernyms along the hypernym path. For example, the hypernyms for {food, nutrient} are, in order, {substance}, {matter}, {physical entity}, and {entity}. An occurrence of the word "food" would be counted as a partial occurrence of {food, nutrient}, {substance}, {matter}, and {physical entity}, but not {entity}, as the latter is more that three nodes distant along the hypernym path. This threshold was chosen by testing thresholds between two and five hypernyms and determining which gave the most empirically satisfactory results. Using this method, an appearance of the word "food" may represent any one of ten distinct synsets. Thus, whenever the word "food" is observed in relation with a given verb, it is counted as a 1/10 observation for each of those ten synsets. This approach sidesteps sense disambiguation to some extent, but it provides a simple heuristic for deriving distributions conditioned on noun classes from observations of noun tokens.

Furthermore, this approach resonates with Lakoff's (1993) argument that metaphorical mappings occur not at the level of situational specifics, but at the superordinate level of general concepts. For example, the metaphor LOVE IS A JOURNEY includes the mapping that the relationship is a vehicle. Although specific instantiations of the metaphor may frame that vehicle variously as a boat ("smooth sailing"), a car ("long, bumpy road"), or a plane ("just getting off the ground"), "the categories mapped will tend to be at the superordinate level rather than the basic level" (Lakoff 1993, page 212). In this method of counting each observed word token as a partial observation under each of the synsets it might represent, observations at the basic level tend to cause the accumulation of observations in the superordinate levels they collectively represent.

However, Resnik's (1993) approach was designed to calculate how strongly verbs select for nouns, whereas the current system needs to calculate how strongly nouns select for verbs. Thus, each time a verb is observed, it is counted as a partial observation of that verb in that grammatical relation for each of the synsets the accompanying noun might represent. For example, when "food" is observed as the direct object of "eat," a 1/10 observation of "eat" is counted in the direct object slot for each of the 10 synsets that "food" might represent.

Tables 5 through 7 presen sample selectional associations from the three example corpora. These tables can be read left-to-right, e.g., the synset troops_1[4] selects to be the agent of "slaughter" with association strength 0.311, i.e., "slaughter" contributes 31.1% of the selectional preference strength of the agent relation for the synset troops_1. Relation abbreviations are from the typed dependency parser (de Marneffe, MacCartney, and Manning 2006).

[4] Subscript notation is used for words with multiple WordNet senses, e.g., troops_1 refers to the first sense of arms, which is the synset {military personnel, soldiery, troops}.

Table 5
Sample selectional associations from the MILITARY corpus.

Synset	Grammatical Relation	Verb	Selectional Association
troops_1	agent	slaughter	0.311
		route	0.146
		massacre	0.146
		reverse	0.122
		wear	0.095
	dobj	transport	0.148
		send	0.057
		exceed	0.042
		attack	0.037
		use	0.034
combat_1	prep_into	deploy	0.263
		engage	0.222
		charge	0.209
		introduce	0.191
		go	0.115
	dobj	see	0.155
		experience	0.123
		perform	0.118
		miss	0.114
		enter	0.109

Table 6
Sample selectional associations from the SCIENCE corpus.

Synset	Grammatical Relation	Verb	Selectional Association
datum_1	agent	refute	0.443
		confirm	0.148
		determine	0.142
		bear	0.134
		support	0.132
	prep_with	provide	0.198
		work	0.187
		interact	0.165
		test	0.117
		fit	0.114
physicist_1	agent	accept	0.229
		scrutinize	0.107
		devise	0.098
		dream	0.086
		acknowledge	0.082
	prep_by	use	0.171
		pioneer	0.106
		espouse	0.106
		conduct	0.078
		head	0.076

Table 7
Sample selectional associations for the Democratic blogs.

Synset	Grammatical Relation	Verb	Selectional Association
obama_1	nsubjpass	elect	0.384
		out	0.146
		boost	0.125
		suit	0.112
		establish	0.092
	prep_for	vote	0.765
		resolve	0.043
		look	0.038
		go	0.031
		canvass	0.027
america_1	nsubj	need	0.121
		yearn	0.109
		wane	0.109
		survive	0.077
		end	0.068
	prep_in	screw	0.273
		end	0.198
		face	0.195
		live	0.177
		work	0.157

3.1.5 Synset Clustering. In some cases, selectional associations of conceptually similar synsets are surprisingly different. For example, in the MILITARY corpus, battalion_2 selects to be the direct object of "destroy" but not of "attack," while the selectional associations for troops_1 are the opposite. To address this, synsets are clustered based on their selectional association strengths. Each synset is represented as an n-dimensional vector, where the nth element is the synset's selectional association for the nth grammatical-relation-and-verb pair. In CorMet, Mason (2004) uses two iterations of nearest-neighbor clustering with a vector dot product as the similarity measure. CMI uses a single iteration of nearest-two-neighbor clustering, but with a different similarity measure.

While analyzing data from the Wikipedia corpora, dot product proved empirically unsatisfactory. Dot product would reward similarity without penalizing dissimilarity, resulting in clusters of mostly unrelated synsets centered around synsets such as whole_1, the ancestor of many synsets for people in WordNet. In one case, using dot product yielded a cluster containing 47 quite diverse synsets, including {room}, {device}, {inactivity}, {state}, {cell}, {musical notation}, {office, power}, {quality}, {substance}, and {death}. The selectional associations for each of these synsets bore some similarity to those for {whole, unit}, i.e., whole_1, even though they were not necessarily similar to one another. To alleviate this problem, an alternative similarity calculation was used:

$$\sum_{v} \lambda\,(\text{v,a}) + \lambda\,(\text{v,b}) - |\lambda\,(\text{v,a}) - \lambda\,(\text{v,b})| \tag{3}$$

Table 8
Two sample clusters from each of three example corpora.

MILITARY	SCIENCE	Democratic Blogs
{delivery, livery, legal transfer}, {conveyance, conveyance of title, conveyancing, conveying}, {giving up, yielding, surrender}, {failure}, {renunciation, renouncement}, {resignation, **surrender**}, {despair}, {relinquishment, relinquishing}, {surrender}, {loss}, {feeling}, {capitulation, fall, surrender}	{**theory**}, {hypothesis, possibility, theory}, {thinking, thought, thought process, cerebration, intellection, mentation}, {theory}, {explanation}, {belief}	{support}, {**endorsement**, indorsement, blurb}, {aid, assist, assistance, help}, {endorsement, indorsement}, {endorsement, indorsement}, {agreement}, {sanction, countenance, endorsement, indorsement, warrant, imprimatur}, {approval, commendation}, {signature}, {second, secondment, endorsement, indorsement}, {name}
{**weapon**, arm, weapon system}, {instrument}, {device}, {weapon, artillery}, {communication, communicating}, {persuasion, suasion}	{**datum**, data point}, {information}	{**terrorism**, act of terrorism, terrorist act}, {coercion}, {terror}

where a and b are two synsets and v ranges over all the verbs in each grammatical relation for which either a or b selects. If there is some verb in a given grammatical relation, v_1, for which b selects but for which a never selects, then $\lambda(v_1, a) = 0$. This similarity function adheres to the spirit of dot product, in that synsets are more similar if they strongly select for the same verbs in the same grammatical relations, but it also penalizes synsets that select for rather different sets of verbs. For example, in the MILITARY corpus, battalion1 selects to be the direct object of "destroy" with association 0.208, while fort_2's association for the direct object of "destroy" is 0.210. However, these two synsets select for very few other common verb-relation pairs, and they share no common verbs in the direct object relation. Using dot product would focus on the single verb and grammatical relation they share in common, but the similarity calculation used here weights more heavily their differences and thus places fort_2 with other synsets similar to defense_8, the synset for a defensive structure.

Table 8 lists some sample clusters from the three example corpora. Clusters are listed using their full synset contents, rather than the abbreviated subscript notation, so as to provide a better sense of the clusters' conceptual scopes. In this table, a single bolded word has been manually chosen as a label for the cluster. Further below, Section 3.4.2 describes a method for automatic labeling of these clusters.

As mentioned above, this same corpus preparation process is applied to any source or target corpus, with the noted exception that corpus acquisition differs between source and target corpora. The resulting clusters serve as the basis for identifying potential metaphorical mappings.

3.2 Source Suggestion

When analyzing a target corpus, there may be clear expectations as to the source domain from which metaphors may map. For example, previous linguistic analyses might indicate potential source domains, such as sports metaphors or military metaphors in political discourse (Howe 1988). However, in other instances, one might not have specific expectations about the types of metaphors being found in a corpus. Additionally, even if one does have such expectations, it may be beneficial to explore alternate potential source domains. This section presents a technique for suggesting, for a given target corpus, potential source corpora from which to map metaphors.

Since CMI uses shared verbs to mediate mappings between a source and target corpus (described below in Section 3.3), the suggestion algorithm finds source corpora with verbs and grammatical relations similar to the verbs and grammatical relations in the target corpus. First, characteristic verbs are sought for a target corpus and all potential source corpora by a process analogous to finding characteristic nouns (Section 3.1.3). Occurrence maps are generated for each characteristic verb of the form:

$$\mathrm{v} \rightarrow \{\mathrm{r}|\mathrm{observed}\,(\mathrm{v,r})\} \tag{4}$$

where v is a characteristic verb, r is some grammatical relation, and observed(v, r) means that v is observed somewhere in the corpus with some noun in relation r.

To suggest source corpora, a target corpus's occurrence map is compared with those for potential sources. This comparison can use either the characteristic verbs (the keys of the mappings) or the verb-relation combinations (the keys and the values to which they map). Given a source corpus α and a potential target corpus β, the suggester scores for the two comparisons, respectively, are:

$$\begin{aligned} \mathrm{SuggesterVerbs}\,(\alpha,\beta) &= \tfrac{(\mathrm{v}_\alpha \cap \mathrm{v}_\beta)}{(\mathrm{v}_\alpha \cup \mathrm{v}_\beta)} \\ \mathrm{SuggesterRelations}\,(\alpha,\beta) &= \tfrac{(\mathrm{vr}_\alpha \cap \mathrm{vr}_\beta)}{(\mathrm{vr}_\alpha \cup \mathrm{vr}_\beta)} \end{aligned} \tag{5}$$

where v_α is the set of verbs in corpus α, vr_α is the set of verb-relation pairs in corpus α, and similarly for β. The suggester score measures either the characteristic verbs or verb-relations shared by the two corpora. A suggester score of 1 means that the two corpora have exactly the same characteristic verbs or verb-relations; a suggester score of 0 means that the two corpora have no characteristic verbs or verb-relations in common.

Verb and case-slot distributions from the Democratic Blogs were compared with 28 somewhat arbitrarily chosen potential source corpora derived from Wikipedia categories. Table 9 lists the top five source corpora, as rated by the suggester. For comparison, the table also lists the suggester scores for _Military_ and _Olympic_sports_, since these sources are suggested by previous work on political rhetoric, as well as the lowest five source corpora, along with the rank out of 28 for each potential source corpus.

Of the top five suggested sources, two resonate with previous literature on metaphors in politics: _War_ (Howe 1988) and _Parenting_ (Lakoff 2002). On the other hand, the suggestion of _War_, _Money_, _Travel_, and, perhaps to a lesser extent, _Humor_ may be due to literal similarities between these domains and politics: during this election, the US was engaged in wars in Iraq and in Afghanistan; money, in terms of both fiscal policy, campaign contributions, and many other aspects, permeates politics; candidates do a good deal of traveling during elections; many political blog posts are humorous or satirical. This analysis suggests that _War_ may be a good source corpus.

Table 9
Suggester scores for several potential source corpora. Each row lists the Wikipedia category for the corpus, its rank among the 28 potential source corpora, the score based on shared verb-relation pairs, and the score based only on shared verbs.

Source Corpus (rank of 28)	SuggesterRelations	SuggesterVerbs
Humor (1)	0.04220	0.08695
War (2)	0.03701	0.09031
Parenting (3)	0.03495	0.08580
Money (4)	0.03014	0.07907
Travel (5)	0.02725	0.08149
_Olympic_sports_ (9)	0.02128	0.07544
Military (12)	0.01838	0.05311
Anatomy (24)	0.01162	0.02458
Laboratories (25)	0.01135	0.04505
Biology (26)	0.01067	0.04938
Medicine (27)	0.01047	0.05100
Electricity (28)	0.00975	0.03974

3.3 Finding Mappings

Once a target and a source corpus are selected and prepared, the process of finding potential metaphorical mappings begins with calculating selectional associations for each cluster in both the source and target corpora. That is, we want to know the degree of association that the cluster as a whole has for all the verbs for which its component synsets select. A cluster's selectional associations are the averages of the selectional associations for all the synsets in the cluster:

$$\Lambda\left(\mathrm{C,v}\right) = \frac{\sum_{s} \lambda\left(\mathrm{s,v}\right)}{|\mathrm{C}|} \tag{6}$$

where C is the cluster whose selectional associations are being calculated, v is some verb-relation pair, and s ranges over the synsets in C. Λ(C,v) is calculated for each verb-relation v selected for by at least one synset in C.

Once each cluster's selectional associations are calculated, CMI finds the polarity of potential mappings from each cluster in the source to each cluster in the target domain. Polarity indicates the degree to which a conceptual cluster in one corpus is framed in terms of a cluster from another corpus, measured by the correspondence in the clusters' selectional associations. The polarity of a potential mapping between two clusters is defined as a weighted sum of their selectional associations for all verb-relations for which both clusters select:

$$\mathrm{pol}\left(\mathrm{X,Y}\right) = \sum_{v} \mathrm{w_{vX}} * \Lambda\left(\mathrm{X,v}\right) + \mathrm{w_{vY}} * \Lambda\left(\mathrm{Y,v}\right) \tag{7}$$

where X and Y are clusters from the source corpus α and target corpus β, respectively, and v ranges over all the verb-relations for which both X and Y select. $\mathrm{w_{vX}}$ and $\mathrm{w_{vY}}$ are the weightings for the selectional associations of X and Y for v, respectively:

$$\begin{aligned} w_{\mathrm{vX}} &= 0.75 * \ln\left(\mathrm{RelFreq}_{\alpha}\left(\mathrm{verb}\right)\right) * S_{\alpha}(\mathrm{verb}) \\ w_{\mathrm{vY}} &= 0.25 * \ln\left(\mathrm{RelFreq}_{\beta}\left(\mathrm{verb}\right)\right) \end{aligned} \tag{8}$$

where verb is the verb in the verb-relation v and $\mathrm{RelFreq}_{\alpha}(\mathrm{verb})$ is the relative frequency of verb in corpus α, and $\mathrm{S}_{\alpha}(\mathrm{verb})$ is the selectional preference strength of verb in α; note that $\mathrm{S}_{\alpha}(\mathrm{verb})$ is calculated using Resnik's method for verbs' selectional preference strength, and it is the only instance in this implementation of calculating the strength with which a verb selects for nouns.

This weighting has two important components. First, the selectional associations of the source cluster are weighted more strongly than those of the target cluster. This weighting is informed in part by the notion of salience imbalance (Ortony et al. 1985), that metaphors highly salient aspects of source concepts and less salient aspects of target concepts. For example, in the sentence, "John's face was like a beet," we understand that John's face was red, because redness is a highlight salient aspect of beets, but is not often a highly salient aspect of faces (see Wilcox (1995) for further specification and experimental exploration of this theory). Additionally, the source cluster's selectional associations are weighted by the selectional preference strength of the verb in the source corpus, further ensuring that the mediating verbs represent salient aspects of the source concept. For example, in the MILITARY corpus, the cluster for the concept *battalion* selects for the verb "form" in the nsubj relation. However, "form" is a rather general verb selected for by many clusters, and thus is not likely a highly salient aspect of a *battalion*. Indeed, the selectional preference strength of the nsubj relation for "form" is relatively low. In contrast, *battalion* also selects to be the passive subject (the nsubjpass relation) of "station," which has a relatively high selectional preference strength, making it a more salient aspect of *battalion* and thus a better candidate for mediating a metaphorical mapping.

This weighting differs from Mason's (2004) polarity calculation, which only sums selectional associations for the target cluster. CorMet only uses verbs with high selectional preference strengths. Rather than using a binary include/exclude approach, CMI instead incorporates selectional preference strength into the polarity calculation, allowing a more fine-grained approach. Empirically, using only the target's selectional associations led to mappings that seemed conceptually unfounded. For example, if a cluster in the MILITARY corpus for *money* selects weakly to be the direct object of the verb "save," but a cluster in the MEDICINE corpus for *life* selects strongly to be the direct object of "save," then the mapping LIFE IS MONEY might be identified as a result. Not only would this mapping be based on weak correspondence of linguistic patterns, but it does not immediately appear informative or well founded given the source and target domains. To prevent such situations, both the source and target's selectional associations are incorporated into the polarity calculation.

The second important component of the weightings in the polarity calculation is the natural logarithm of the relative frequency of the verb mediating the mapping in the relevant corpus (e.g., the verb's relative frequency in the source corpus when weighting selectional associations of clusters from the source corpus). As a result, verbs that are characteristic of a corpus have more strength in mediating metaphorical mappings. If a verb is just as frequent in a corpus as in general English, the relative frequency is 1, and the natural logarithm of which is 0, thus nullifying the verb's contribution to the mapping. If a verb is half as frequent in the corpus as in general English, the relative frequency is 0.5, the natural logarithm of which is -0.693, which has an overall effect of weakening the polarity of the mapping. If a verb is twice as frequent in the corpus as in general English, the relative frequency is 2, the natural logarithm of which is 0.693, which increases the overall polarity of the mapping. Thus, the natural logarithm of the verb's relative frequency is an effective means of modulating mapping polarity so that metaphors mediated by particularly characteristic verbs receive a higher polarity. This use of characteristic verbs and grammatical relations associated with them connects back to the method of source suggestion; corpora that share characteristic verbs in common are likely to generate mappings of a higher polarity, so source suggestion is based on shared characteristic verbs. Furthermore, this approach also provides a finer-grained alternative to Mason's approach of including only characteristic verbs and excluding others from the mapping process.

Two additional measures are taken to ensure that the verb-relation pairs that mediate metaphorical mappings are meaningful in the context of those mappings. First, before calculating polarity, a cluster's selectional associations are pruned, such that any verb-relation pair for which the cluster's degree of association is less than a given threshold is not used in the polarity calculation. The implementation described here uses a threshold of 0.015, which means that a verb-relation pair is pruned if it contributes on average less than 1.5% of the selectional preference strength of the synsets in the cluster selecting for it. Second, while calculating polarity, only those verb-relations making a significant contribution to polarity are included. This prevents the accumulation of many common verbs for which a cluster may select somewhat weakly, such as "do," "make," or "get," from leading to the appearance of a strong metaphorical mapping. In this implementation, a threshold of 0.01 is required for a verb-relation pair to be counted in the polarity calculation. Because of the various values involved, there is not a simple, intuitive explanation for the value of this threshold, but it can be considered in the following way: the selectional associations of the source cluster plus those of the target cluster for a given verb-relation, modulated by the logarithm of the verb's relative frequency, must account for at least 1% of the overall selectional preferences of those clusters for the verb-relation pair to contribute to a potential mapping. As Mason (2004) points out, there is little precedent for systems designed around identifying conceptual metaphors; thus, the constants used here should be considered reasonable approximations at best and may be improved in later work.

The overall polarity of a mapping is the difference between the polarity of the mapping from source to target and from target to source:

$$\mathrm{TotPol}\,(X{,}Y) = \mathrm{pol}\,(X{,}Y) - \mathrm{pol}\,(Y{,}X) \tag{9}$$

which prevents situations where, for example, "sit" is selected for by scientists in the LAB domain and bankers in the FINANCE domain. The potential range for polarity values is theoretically (-∞, ∞); polarity is based on weighted sums of clusters' selectional associations for verb-relation pairs, and there is no technical limit on the number of verb-relations that can mediate a potential mapping. However, empirically, polarity values tend to fall in the range (-3, 5).

For all mappings where the total polarity is greater than 0, CMI calculates a confidence score, based partly on the total polarity and partly on the number of verb-relations that mediate the mapping:

$$\mathrm{Conf}\ (M_{X,Y}) = 0.65 * \mathrm{TotPol}\,(X{,}Y) + 0.35 * \mathrm{NumVR}\,(M_{X,Y}) \tag{10}$$

where $M_{X,Y}$ is a mapping from cluster X to cluster Y, and NumVR($M_{X,Y}$) is the number of verb-relations that mediate $M_{X,Y}$. This confidence measure differs slightly from that used in CorMet (Mason 2004), which normalizes total polarity and number of verb-relations. That is, for any given set of mappings from a source corpus to a target corpus, CorMet divides each mapping's total polarity by the highest polarity for that source and target corpus, as well as divides the number of verb-relations by the maximum number of verb-relations mediating a mapping from that source corpus to that target corpus. While such an approach means that confidence measures for mappings between any two corpora are scaled to the range (0, 1], it makes difficult comparing the confidences of mappings between different sets of corpora. This issue is discussed further below in the context of example results. Since confidence is only calculated for mappings with positive polarity, the theoretical range for confidence values is $(0, \infty)$. In practice, confidence values typically fall in the range (0, 5) and generally follow a distribution resembling a power law, with a few high-confidence mappings and many-low confidence mappings. Because of the various factors involved, the specific numerical value for a confidence score could have a variety of meanings. For example, a mapping with a confidence of 1.0 could have a total polarity of 1.0 and be mediated by one verb-relation, or it could have a total polarity of 0.461 and be mediated by two verb-relations, etc.

The result of the mapping process is a series of mappings, each from a source cluster to a target, each with an associated confidence score. However, a few transformations are performed on the results to make them more legible to, and interpretable by, a human user. The next section describes the process of making those results more comprehensible, which is then followed by sample mapping results.

3.4 Improving Metaphor Legibility

Once confidence ratings have been calculated for all potential mappings, the mappings are pivoted around the target clusters. Thus, rather than a list of all mappings ordered by their confidence scores, CMI produces a set of target clusters with accompanying potential metaphorical mappings for each target cluster. For each cluster, potential mappings are listed in order of descending confidence. Since the purpose is to facilitate human metaphorical thinking, this format allows an individual to find conceptual clusters of interest from the target corpus quickly and identify potential metaphorical mappings for those clusters. Furthermore, since there may be a very high number of mappings from any given source corpus to target corpus (see below for examples), only the upper one percentile in terms of confidence score are presented. Two additional steps are taken to make the results more comprehensible: example sentence fragments are found for each identified metaphor, and clusters, which may contain many potentially diverse synsets, are assigned a single label.

3.4.1 Example Sentence Fragments. Example sentence fragments for each mapping are sought from the target and source corpora based on the clusters and verb-relations involved in the mapping. Conceptual metaphors occur at the level of superordinate concepts, but individual instantiations of the metaphor may be at the level of subordinate concepts (Lakoff 1993). For example, metaphor mappings involving the *weapon* cluster (see Table 8) may be linguistically instantiated with any number of words representing the concept of a weapon. Therefore, to find example phrases, CMI inverts the process of finding superordinate concepts used during selectional preference learning. Given a cluster of synsets, all words that might be instances of hyponyms (children in the WordNet ontology) of any synset in the cluster are taken as potential instances of that concept. For example, the synset {politician}, referring to a leader in civil administration, has as hyponyms {governor}, {legislator}, and {mayor, city manager}. Thus, when searching for example sentences for a mapping that involves politicians, sentences are also sought that involve the words "governor," "legislator," etc. The example sentences are then scored based on the product of the relative frequency of the noun involved, the relative frequency of the verb involved, and the strength of the verb-relation mediating the mapping. These scores are used to sort the example sentences when presented to a user, bringing examples that are likely to be more compelling to the top.

3.4.2 Cluster Labels. The source and target clusters involved in any given mapping may consist of many different synsets but represent one coherent concept. Thus, to make mappings more comprehensible, clusters are given a single label that attempts to capture the concept that the cluster represents. CorMet (Mason 2004) uses cluster labels assigned by hand. This section suggests a method for automatically labeling clusters.

The clusters of synsets involved in each mapping are labeled based on the word that appears most frequently in the cluster′s synsets. The cluster in Table 8 for *weapon* would be assigned the label "weapon." If no word appears more than others, the word from the cluster that appears most often in the example sentences is used.

CMI does not automatically determine whether the label for a cluster should receive a definite or indefinite article. It may be computationally possible to discriminate between single nouns, which would receive "a" or "an;" mass nouns, which would receive no article; proper nouns, which would receive "the;" and the various other special cases that exist. However, the development of such a technique is neither within the scope of this work and nor necessary for the goals here. The leap from, e.g., ELECTION IS LIKE BATTLE to AN ELECTION IS LIKE A BATTLE seems more computationally costly for the computer than cognitively demanding for a person.

4. Sample Metaphor Mapping Results

This section presents sample results of potential metaphors identified by CMI. The target corpus analyzed for metaphors is the set of posts from Democratic political blogs during the 2008 US election described in Section 3.1.1. The source corpora used here were derived from the Wikipedia categories for _Military_ and _Science_.

As mentioned above, presenting the entire list of potential metaphors identified by CMI would take far too much space; 771 potential mappings were identified from the MILITARY corpus to the blog posts, and 1329 from the SCIENCE corpus. Even with focusing on the upper one percentile in terms of confidence score, showing the results in full, complete detail would be excessive. Therefore, the following two subsections provide an overview of that upper one percentile, as well as close inspection of a few metaphorical mappings therein. The subsequent section then assesses and evaluates the entirety of the upper one percentile for both these sets of mappings.

4.1 Military

Previous analysis of political discourse has suggested that WAR is a common source domain for political metaphors (Howe 1988). However, as described in Section 3.1.1, only 70% of the articles from the _War_ category were relevant in the manner intended, e.g., many dealt with war museums rather than the conduct of war. In contrast, 90% of the articles from the _Military_ category were relevant in the manner intended, i.e., they dealt directly with military concepts and activities. Thus, articles from the _Military_ category are used as the source corpus here.

Table 10 presents an overview of the mappings from the MILITARY corpus to the blog posts with confidence in the upper one percentile. Target clusters are listed in the left hand column, source clusters that map to each target are listed in the center column, and the confidence score for each mapping is listed in the right hand column. All synsets in each cluster are listed, with the automatically assigned label for the cluster at the top in bold. As described above, all source clusters that map to a given target are grouped together, and the groups of mappings are listed in order of the confidence score for each mapping. The table can be read left-to-right for the identified metaphors, e.g., based on the first row a state is like a battle with a confidence of 1.578 (as stated above, confidence scores typically fall in the range (0, 5) with a roughly power law distribution).

Many of these results align with intuitive expectations based on previous linguistic analysis (Howe 1988). For example, the mapping A STATE IS A BATTLE resonates with the common notion of a "battleground state." Similarly, A CAMPAIGN IS A MASSACRE resonates, at least to some extent, with the general metaphor that AN ELECTION IS A WAR, in that it frames the election in terms of violence and conflict. Other mappings, though, may not be as readily comprehensible. For example, the mapping A CANDIDATE IS A UNIT may at first seem anomalous. However, the concept unit refers here to a military *unit*, as evidence by phrases from the example sentence fragments, such as "airborne unit" or "operational unit" (see Table 11 below). Also, the fact that there are distinct clusters for *state* and *Ohio* may seem somewhat problematic; intuitively, one might expect that, since Ohio is a state, it should be subsumed by the general cluster for state. In this interpretation, the existence of two alternative clusters might be seen as a problem with the clustering method. On the other hand, existence of two distinct clusters might be interpreted as indicating that the selectional associations of Ohio are significantly different from those for states in general. This issue is returned to further below in the context of presenting additional details for some of these mappings.

It is also important to note that the cluster for the concept *Iraq* is the target of several different metaphors. This support for metaphorical pluralism(Lakoff and Johnson 1980) is a key feature in terms of ensuring that the metaphors identified by CMI align as closely as possible with readers understandings and experiences of metaphors.

We now examine some of these mappings more closely by providing additional details. Table 11 shows details about the mappings A STATE IS A BATTLE, A CANDIDATE IS A UNIT, and A CAMPAIGN IS A MASSACRE. These particular metaphors were chosen as representative of the range of metaphors within the upper one percentile.

The details of these mappings provide an understanding of, for example, why a state might be like a battle. When taken alone, individual example sentences may not appear particularly compelling. "Serving as" and "supporting" may not seem particularly associated with the essence of a military unit. Therein lies the strength of CMI. While these examples may not be individually compelling, what is compelling is their aggregate. It is easy not to notice such phrases or expressions as these when they are encountered individually and separately. CMI finds such patterns, patterns that might otherwise be easily missed, and draws attention to them, encouraging consideration of what those patterns might imply. As described above, conceptual metaphors can be quite difficult to consider critically, let alone to notice in the first place (Lakoff and Turner 1989). The goal of CMI, then, is to draw to people's attention patterns in a target corpus and encourage them to consider how those patterns might be variously interpreted, either to bring to light current implicit framings, or to provide a new, alternative framing for familiar concepts and ideas.

Table 10
Overview of the upper one percentile in terms of confidence score for mappings from MILITARY to Democratic blog posts, including the target concept being mapped to, the source concept being mapped from, and the mapping's confidence score.

Target (label and cluster)	Source (label and cluster)	Conf
state – {state, province}, {American state}	**battle** – {event}, {struggle, battle}, {group action}, {military action, action}, {attempt, effort, endeavor, endeavour, try}, {conflict, struggle, battle}, {battle, conflict, fight, engagement}	1.578
candidate – {legislator}, {campaigner, candidate, nominee}, {senator}, {politician, politico, pol, political leader}	**unit** – {thing}, {object, physical object}, {measure, quantity, amount}, {whole, unit}, {definite quantity}, {unit}, {unit, building block}, {part, portion, component part, component, constituent}, {concept, conception, construct}, {unit}, {whole}, {unit of measurement, unit}, {relation}	1.451
iraq – {Asian country, Asian nation}, {country, state, land}, {Iraq, Republic of Iraq, Al-Iraq, Irak}	**unit** – {thing}, {object, physical object}, {measure, quantity, amount}, {whole, unit}, {definite quantity}, {unit}, {unit, building block}, {part, portion, component part, component, constituent}, {concept, conception, construct}, {unit}, {whole}, {unit of measurement, unit}, {relation}	1.334
	corps – {corps}, {corps, army corps}, {body}	1.216
	army – {crowd}, {gathering, assemblage}, {army}, {United States Army, US Army, U. S. Army, Army, USA}, {military service, armed service, service}, {agency, federal agency, government agency, bureau, office, authority}, {army, regular army, ground forces}, {administrative unit, administrative body}	1.209
	force – {organization, organisation}, {force, personnel}	1.174
ohio – {Ohio, Buckeye State, OH}, {Ohio, Ohio River}	**battle** – {event}, {struggle, battle}, {group action}, {military action, action}, {attempt, effort, endeavor, endeavour, try}, {conflict, struggle, battle}, {battle, conflict, fight, engagement}	1.271
campaign – {activity}, {venture}, {campaign, military campaign}, {journey, journeying}, {operation, military operation}, {political campaign, campaign, run}, {expedition}, {contest, competition}, {campaign, hunting expedition, safari}, {undertaking, project, task, labor}, {campaign, cause, crusade, drive, movement, effort}, {race}	**massacre** – {slaughter, massacre, mass murder, carnage, butchery}, {murder, slaying, execution}, {homicide}	1.144

Table 11
Details of three example metaphorical mappings from MILITARY to the Democratic blogs, including the target concept, the source concept, the verb-relations mediating the mapping, and example sentences from the source and target corpora.

Target	Source	Verb-Rln	Target Ex Frag	Source Ex Frag
state	battle	"win" – dobj	"**win Virginia**"	"**winning battles**"
		"defeat" – in	"**defeated in California**"	"**defeat** the Spartans **in** the **Battle**"
		"fight" – in	"**fighting** the discriminatory Proposition 8 **in California**"	"**fight in** hand-to-hand **combat**"
		"invade" – dobj	"**invade Hawaii**"	"**invaded Vietnam**"
candidate	unit	"support" – dobj	"**support Obama**"	"**supported** Navy Special Naval Landing Force airborne **units**"
		"serve" – as	"**served as** a state **senator**"	"**serves as** the military police **unit**"
		"schedule" – nsubjpass	"**candidate** is **scheduled**"	"**unit** was **scheduled**"
		"support" – agent	"**supported by** almost every single **Democrat**"	"**supported by** RHAF **units**"
campaign	massacre	"attack" – nsubj	"**campaign** has been **attacking**"	"**Massacre** || Pro-Indonesian Militia group **attack**"
		"survive" – dobj	"**survived** this **campaign**"	"**survived** the **massacre**"
		"involve" – prep_in	"**involved in** a fraud **campaign**"	"**involved in** the Centralia **Massacre**"

The example sentence fragments also adhere to the idea that metaphorical mappings occur at the level of superordinate concepts but individual instances occur with respect to situational specifics (Lakoff 1993). The example fragments listed above for a STATE IS A BATTLE involve specific words that are instances of the concept, such as "Virginia," "California," and "Hawaii." Similarly, example sentences from the target corpus for the mapping A CANDIDATE IS A UNIT involve the word "candidate," as well as other instantiations of the concept, such as "Obama" and "McCain."

These results also expose a few potential weaknesses in the implementation. For example, in the examples for A STATE IS A BATTLE, WordNet mistakes Vietnam as referring to the Vietnam war, rather than the country of Vietnam. The first example sentence fragments for A CAMPAIGN IS A MASSACRE also contains a parse error; the parser failed to interpret the "||" symbol as a sentence break and thus incorrectly identifies "massacre" as the subject of "attack."

Such irregularities notwithstanding, CMI is able to identify what we might consider intuitively reasonable potential metaphors. After presenting samples mapping from a different corpus, we move on to evaluating these results beyond simple intuitions.

4.2 Science

As described above, part of the point of CMI is to identify potential metaphors that might otherwise go unnoticed. The above examples from the MILITARY domain demonstrate how CMI behaves when using a source corpus for which there are specific expectations. This section demonstrates that CMI can also work when using a source corpus for which such expectations do not exist.

The choice of such a source domain was guided by the notion of near and far sources (Blanchette and Dunbar 2000; Gentner, Rattermann, and Forbus 1993; Holyoak and Koh 1987). Near sources are those that bear some surface similarity with the target concept or domain, whereas far sources bear only relational similarity. For example, Blanchette and Dunbar (2000) describe a study in which participants were asked to generate analogies[5] either for or against a zero-deficit policy for the Canadian government. Near sources included other political economic situations, which also dealt with governments and deficit, while far sources included natural resources or natural disasters, which did not involve governments or deficits but did involve similar supply-demand relationships.

Using politics as a target domain, MILITARY represents a relatively near source domain: both deal with nations, states, and governmental bodies; both involve opinions and ideologies; both involve conflict. Thus, to complement the above results from MILITARY, this section uses SCIENCE as a source domain that is relatively far from politic–documents about science do not usually focus on governmental bodies or political viewpoints–that also has a moderate score given by the suggester (17th out of 28 based on shared verb-relations, see Section 3.2).

Table 12 presents an overview of the mappings from SCIENCE to the Democratic blogs, showing the upper one percentile in terms of confidence. This table follows a format similar to that of Table 10.

Since, to our knowledge, no previous linguistics or political science work has investigated SCIENCE metaphors in politics, it is difficult to say whether or not these mappings align with expectations. However, they may help provide alternative reframings of familiar concepts. For example, A CAMPAIGN IS AN EXPERIMENT offers a different way of thinking about politics, not as a combative process between two adversaries, but rather as experimentation with different policies or political view points. The mappings A CANDIDATE IS AN IDEA and A CANDIDATE IS A METHODOLOGY arguably provide similar reframings, discussed further below when examining details of these mappings.

[5] While beyond the scope of the current article, see Gentner et al. (2001) for a description of different perspectives on the relationship between metaphor and analogy.

Table 12: Overview of the upper one percentile in terms of confidence score for mappings from SCIENCE to Democratic blog posts, including the target concept being mapped to, the source concept being mapped from, and the mapping's confidence score.

Target (label and cluster)	Source (label and cluster)	Conf
campaign – {activity}, {venture}, {campaign, military campaign}, {journey, journeying}, {operation, military operation}, {political campaign, campaign, run}, {expedition}, {contest, competition}, {campaign, hunting expedition, safari}, {undertaking, project, task, labor}, {campaign, cause, crusade, drive, movement, effort}, {race}	**experiment** – {research}, {problem solving}, {experiment, experimentation}, {experiment, experimentation}, {experiment}, {venture}, {inquiry, enquiry, research}, {undertaking, project, task, labor}, {scientific research, research project}	1.841
	theory – {theory}, {hypothesis, possibility, theory}, {thinking, thought, thought process, cerebration, intellection, mentation}, {theory}, {explanation}, {belief}	1.417
	behavior – {behavior, behaviour, conduct, doings}, {behavior, behaviour}, {behavior, behaviour}, {action, activity, activeness}, {state}, {demeanor, demeanour, behavior, behaviour, conduct, deportment}	1.256
race – {raceway, race}, {flow}, {race}, {group, grouping}, {subspecies, race}, {race}, {abstraction, abstract entity}, {biological group}, {social event}, {natural process, natural action, action, activity}, {watercourse, waterway}, {canal}, {slipstream, airstream, race, backwash, wash}, {taxonomic group, taxonomic category, taxon}	**physics** – {physics, natural philosophy}, {natural science}, {physics, physical science}	1.423
	individual – {person, individual, someone, somebody, mortal, soul}, {scientist}, {organism, being}, {causal agent, cause, causal agency}	1.328
	event – {discovery}, {event}, {discovery, breakthrough, find}, {insight, brainstorm, brainwave}, {discovery}, {disclosure, revelation, revealing}, {speech act}, {understanding, apprehension, discernment, savvy}, {discovery, find, uncovering}	1.215
	study – {cognition, knowledge, noesis}, {science, scientific discipline}, {knowledge domain, knowledge base, domain}, {skill, science}, {discipline, subject, subject area, subject field, field, field of study, study, bailiwick}, {ability, power}	1.815
	chemistry – {chemistry, chemical science}, {chemistry, interpersonal chemistry, alchemy}, {chemistry}, {relation}, {social relation}	1.180
candidate – {legislator}, {campaigner, candidate, nominee}, {senator}, {politician, politico, pol, political leader}	**idea** – {concept, conception, construct}, {idea, thought}, {content, cognitive content, mental object}	1.417
	methodology – {methodology}, {know-how}, {epistemology}, {methodology, methodological analysis}	1.377

Table 12: Mappings from SCIENCE to the Democratic blogs.

Target (label and cluster)	Source (label and cluster)	Conf
	program – {writing, written material, piece of writing}, {program, programme}, {social event}, {program, programme, computer program, computer programme}, {publication}, {shaft}, {announcement, promulgation}, {axle}, {daybook, journal}, {diary, journal}, {record}, {presentation, presentment, demonstration}, {journal}, {code, computer code}, {message, content, subject matter, substance}, {book, volume}, {ledger, leger, account book, book of account, book}, {performance}, {periodical}, {written communication, written language, black and white}, {platform, political platform, political program, program}, {broadcast, program, programme}, {program, programme}, {document, written document, papers}, {journal}, {plan, program, programme}, {software, software program, computer software, software system, software package, package}, {information, info}, {program, programme}, {course of study, program, programme, curriculum, syllabus}, {show}, {journal}, {product, production}	1.221
individual – {person, individual, someone, somebody, mortal, soul}	**principle** – {logic}, {law, natural law}, {logic, logical system, system of logic}, {logic}, {logic}, {common sense, good sense, gumption, horse sense, sense, mother wit}, {sagacity, sagaciousness, judgment, judgement, discernment}, {principle}, {logic}	1.212
leader – {leader}	**photon** – {boson}, {gauge boson}, {photon}	1.171

These mapping results also expose some of the strengths, as well as potential limitations, of the clustering and labeling techniques used here. For example, in A RACE IS AN INDIVIDUAL, the cluster for the concept *individual* might better be labeled "scientist." However, there were more example sentences found for this mapping that use the word "individual" than use the word "scientist," which results in a potentially unintuitive label. Likewise, the cluster *event* in A RACE IS AN EVENT might better be labeled "discovery," but is instead labeled "event" for similar reasons.

Table 13 shows further details for A CAMPAIGN IS AN EXPERIMENT, A CANDIDATE IS AN IDEA, and A LEADER IS A PHOTON, using the same format as that of Table 11.

The third mapping here, a leader is a photon, demonstrates how some of the mappings with high confidence scores might be somewhat spurious. For example, leaders are certainly "tested," as are photos, but the tests to which a photon is subjected are slightly different than those to which a political leader is subjected. While there is some measure of linguistic correspondence, in this case, that linguistic correspondence does not seem to indicate a strong conceptual mapping.

Table 13
Details of three example metaphorical mappings from SCIENCE to the Democratic blogs, including the target concept, the source concept, the verb-relations mediating the mapping, and example sentences from the source and target corpora.

Target	Source	Verb-Rln	Target Ex Fragment	Source Ex Fragment
campaign	experiment	“violate” – nsubj	“**Campaign violated**”	“**experiment** would **violate**”
		“conduct” – dobj	“**conducted** an extremely effective **campaign**”	“**conducting** animal **testing**”
		“refer” – prep_to	“**refers to** the vicious NRA **campaign**”	“**refers to** an **experiment**”
		“focus” – prep_on	“**focus on** the Miss 1 **race**”	“**focused on experimentation**”
candidate	idea	“disagree” – prep_with	“**disagree with Obama**”	“**disagreed with** the **concept**”
		“write” – prep_about	“**written** much **about** John **McCain**”	“**wrote about** the **philosophy**”
		“support” – dobj	“**support Obama**”	“**support** the Big Bang **Theory**”
		“focus” – prep_on	“**focused on Palin**”	“**focused on** the **philosophy**”
leader	photon	“test” – nsubjpass	“**President** was **tested**”	“**photon** in region A is **tested**”
		“apply” – prep_to	“**applied to** Hillary **Clinton**”	“**apply** a test for vertical polarization **to** the **photon**”

However, the details of the first two mapping in Table 13 offer a compelling alternative conceptualization of politics and elections. Most political metaphors are combative and conflict-oriented (Howe 1988). In contrast, these metaphors present an electio as "conducting" a process of scientific experimentation, that candidates represent different political ideas, methodologies, or theories, where some voters may "support" one candidate while "disagreeing with" another. Ultimately, one candidate will be proven in an election, just as an experiment may eventually prove a given theory. To reiterate, part of the goal of CMI is to encourage examination of current metaphors, as well as consideration of novel alternatives. "In getting us to understand how it could be true, [the metaphor] makes possible a new understanding" (Lakoff and Johnson 1980, page 175).

5. Evaluation

CMI is designed as a tool that identifies potential conceptual metaphors in order to foster critical thinking and creativity. Previous studies in the context of science education (Baumer et al. 2009) and political blogs (Baumer et al. 2009; Baumer, Sinclair, and Tomlinson 2010) have demonstrated that CMI can effectively be used to foster critical and creative thinking about conceptual metaphors. The evaluation presented here focuses on the question of whether CMI effectively identifies potential conceptual metaphors.

As Mason (2004) notes, the intersubjective nature of metaphor makes evaluating any system based on finding conceptual metaphors quite challenging. No formal, standardized metrics exist for the task. Furthermore, we argue that such quantitative metrics might only capture a small portion of the potential value of techniques such as CMI. For example, one might take an approach inspired by Mason (2004) and compare computationally identified metaphors to analyses by expert linguists, measuring both recall (i.e., of the list of established metaphors, how many did the system identify?) and precision (i.e., of the identified metaphors, what portion fit with established metaphors?). Such an approach might be viable in evaluating results about which we have established expected metaphors, for example, the above results of mapping from the MILITARY corpus to the Democratic political blogs. However, this approach would not do as well in evaluating the ability to discover unexpected, yet informative or insightful metaphors, nor would it deal well with cases where there are little to no established expectations, for example, the mappings presented above from the SCIENCE corpus to the Democratic blogs. While one might argue that, if there are no reasonable expectations, the system should not return much in the way of results. On the other hand, if the system can return results that, while unexpected, might be interesting, informative, or insightful, we may prefer an evaluation method that accounts for such possibilities.

To that end, the evaluation present here consists of two portions. The first portion draws on the techniques pioneered by Mason (2004) of comparing his results with the Master Metaphor List (Lakoff, Espenson, and Schwartz 1991). As described above, CMI is fundamentally different from CorMet (Mason 2004) in its goals and methods, and therefore a direct comparison with Mason's results would neither be informative nor demonstrate fully CMI's strengths. Instead, the first portion of the evaluation compares computationally identified metaphors to conceptual metaphors identified in previous expert linguistic analysis. The results from this first portion of the evaluation demonstrate that computationally identified metaphors bear conceptual similarity to metaphors identified in previous expert linguistic analysis.

The second portion of the evaluation explores CMI's capacity to discover potentially unexpected but informative metaphors. In evaluating different WordNet-based measures of similarity, Budanitsky and Hirst (2006) note some of the difficulties in comparing quantitative results from computational linguistic analyses with the often rich and qualitative subjective understandings that humans have of the same phenomenon. Ultimately, they conclude that human subjective judgments are the ultimate determiner of semantic similarity, and therefore evaluate several quantitative measures via comparison with data collected from human subjects in psychological studies. We take a similar view, that human judgment is the ultimate determiner of whether a potential metaphor is sensible, unexpected, insightful, etc., and thus the second portion of this evaluation assess computationally identified metaphors based on ratings given by human subjects in controlled studies. The results from this second portion of the evaluation demonstrate that assessments by non-expert human subjects correlate significantly with the confidence scores assigned by CMI.

5.1 Comparison with Expert Linguist Analysis

In analyzing discourse from news media coverage, Howe (1988) describes a number of war metaphors. Rather than listing overarching metaphors, Howe focuses on specific instantiations, for example, the various implications of describing a political conflict in terms of *trench warfare* as opposed to *guerilla warfare*. However, he does suggest a number of general metaphorical framings, such as that different parties are battling armies or opposing sides in a conflict, that an election is a war between these armies, and that individual candidates represent either soldiers in that army or generals leading that army. This first portion of the evaluation here compares the results of CMI to Howe's results, specifically, the general metaphorical framings of ELECTION IS A WAR, PARTIES ARE ARMIES, and POLITICIANS ARE SOLDIERS. This section describes the method used for comparison as well as the results and implications of that comparison.

5.1.1 Methods. In evaluating CorMet, Mason (2004) uses a subjective, intuitive comparison of whether or not the metaphors his system finds align with those from the Master Metaphor List (Lakoff, Espenson, and Schwartz 1991). Here, a quantitative metric is suggested for comparing metaphors based on WordNet similarity metrics.

The metaphors identified by Howe (1988) relate two words, each representing a single concept. The metaphors identified by CMI relate two clusters of WordNet synsets, where each cluster represents a single concept. Therefore, one method of comparing results is to compare the similarity of the concepts represented by the synset clusters to the canonical words used by Howe to represent those concepts, which can be calculated via a conceptual similarity metric that uses WordNet. The approach used here is to calculate the average similarity between the synsets in a cluster and the synset that most closely represents the word Howe uses for the metaphor; this latter synset is chosen based on a subject reading of synset glosses.

This comparison could theoretically be made using any WordNet-based similarity metric. Here, two such metrics are used: Leacock and Chodorow (1998) normalized path length, and Lin's (1998) universal similarity measure with information content derived from the Brown corpus (Francis and Kučera 1982). These similarity metrics were chosen based largely on their consistently high correlation with human subjects' ratings of similarity (over 0.8 for both metrics) (Budanitsky and Hirst 2006). We include both metrics here because Lin's takes into account corpus-based information content, whereas Leacock's and Chodorow's only uses WordNet's ontological hierarchy, and Budanitsky and Hirst (2006) comparisons show marked differences between these two classes of similarity metrics.

Intuitively, one might like to have a simple test that confirms or rejects individual results of CMI; does a given computationally identified metaphor align with previously-identified linguistic metaphors or not? However, we posit that the approach used here incorporating WordNet-based similarity measures provides a more fine-grained, but still quantitative, assessment of CMI's results. Furthermore, it provides an initial performance assessment against which future improvements to the techniques presented here could be compared.

5.1.2 Results. Of the upper one percentile of potential metaphors identified from the MILITARY corpus to the Democratic blogs (Table 10), some fit conceptually with the metaphors Howe (1988) identifies. A STATE IS A BATTLE and, similarly, OHIO IS A BATTLE fit subjectively as entailments of an ELECTION IS A WAR. Such results help confirm that CMI behaves as one might expect. However, because these entailments do not align directly with Howe's metaphors, we do not here perform a direct comparison. Similarly, while the metaphors involving the concept *iraq* are potentially reasonable mappings, they have no obvious parallel in the metaphors Howe describes and are thus not involved in the comparisons here. Also, no computationally identified metaphor seems an obvious match for PARTIES ARE ARMIES. Thus, we focus our quantitative comparison on the metaphors A CANDIDATE IS A UNIT and A CAMPAIGN IS A MASSACRE.

The mapping A CANDIDATE IS A UNIT (again, unit here refers to a military unit) resembles most closely the metaphor POLITICIANS ARE SOLDIERS. For the concept *politician*, we use the synset politician$_3$, referring to a person engaged in party politics, and for the concept *soldier* we use soldier$_1$, referring to a person who serves in an army (to reiterate, soldier$_1$ refers to the first sense of the word "soldier" in WordNet). The mapping A CAMPAIGN IS A MASSACRE resembles most closely the metaphor AN ELECTION IS A WAR. For the concept *election* we use election$_3$, referring to a vote to select a political official, and for the concept *war* we use war$_3$, referring to the waging of armed conflict. The choice of mappings and metaphors to compare, and the selection of synsets to represent individual concepts, are based on the authors' subjective judgments.

Table 14 compares the semantic similarity of the target and source from the computationally identified metaphor to those from the metaphor identified by expert linguistic analysis. This table lists synsets from the source and target clusters identified by CMI, along with their automatically assigned labels, the WordNet synset most like the source or target concept in Howe's metaphor, the Leacock-Chodorow similarity, and the Lin similarity. For the similarity metrics, the table also includes a percentile score, indicating the percentile of the cluster's similarity to the synset in question when compared to all other clusters from the corpus. These percentile scores indicate how well the cluster in the computationally identified metaphor matches the concept in the metaphor identified by Howe; a higher percentile means the cluster is a better fit when compared to other clusters. For example, the *candidate* cluster is more similar to politician$_3$ than 99.3% of the clusters from the Democratic blogs, according to Lin's similarity metric. These similarity metrics are not meant to be purely objective assessments of whether a cluster in a computationally identified metaphor resembles a concept in a metaphor from expert analysis. Rather, they help confirm intuitive assessments of similarity by determining whether a different cluster might better fit the concept in the metaphor.

These results demonstrate the strengths, as well as some weakness, of CMI. For both target concepts (*candidate* and *campaign*), CMI identified metaphorical mappings involving clusters with very high degrees of semantic similarity to the concepts in Howe's (1988) metaphors; in both cases, both similarity metrics showed that the cluster involved in the mapping was in the 95th percentile compared to other clusters from the corpus. However, CMI was not as effective at identifying appropriate source clusters for these mappings. In both cases, one similarity metric indicated that the cluster involved in the mapping was only in the 50th percentile in terms of similarity to the desired source concept. Thus, these results show that CMI can effectively identify mappings that resemble metaphors identified in previous expert linguistic analysis, and they help provide a baseline against which future work can be compared.

It should be noted that this evaluation method only compares source concept to source concept and target concept to target concept. That is, it does not compare the relationship in the computationally identified metaphors to that in the metaphor from expert linguistic analysis, because it is not immediately obvious how such a comparison would be made. One might consider using the similarity metrics to gauge the semantic similarity between source and target concepts, but such a comparison would not truly assess the nature of the relationship involved in the metaphor. Alternatively, one might use identified metaphors as input to some analogical reasoning system (Falkenhainer, Forbus, and Gentner 1989; Turney 2008), comparing the mappings derived from computationally identified metaphors to those from metaphors identified in expert analysis, but such an approach is beyond the scope of the current article. Thus, though admit-

Table 14
Similarities between clusters in the computationally identified metaphors and concepts in the metaphors identified in previous expert linguistic analysis (Howe 1988). Similarity assessed using both Leacock-Chodorow (Leacock and Chodorow 1998) normalized path length and Lin's (Lin 1998) universal similarity. Similarities are given in terms of the raw score for each similarity measure, as well as the percentile score, e.g., the synsets in the candidate cluster are more similar to {politician, politico, pol, political leader} than 99.3% of all clusters from the blog corpus.

CMI	Howe (expert analysis)	LC	Lin
candidate – {legislator}, {campaigner, candidate, nominee}, {senator}, {politician, politico, pol, political leader}	**politician** – {politician, politico, pol, political leader}	2.715 (100%)	0.710 (99.3%)
unit – {thing}, {object, physical object}, {measure, quantity, amount}, {whole, unit}, {definite quantity}, {unit}, {unit, building block}, {part, portion, component part, component, constituent}, {concept, conception, construct}, {unit}, {whole}, {unit of measurement, unit}, {relation}	**soldier** – {soldier}	1.168 (80.5%)	0.070 (52.0%)
campaign – {activity}, {venture}, {campaign, military campaign}, {journey, journeying}, {operation, military operation}, {political campaign, campaign, run}, {expedition}, {contest, competition}, {campaign, hunting expedition, safari}, {undertaking, project, task, labor}, {campaign, cause, crusade, drive, movement, effort}, {race}	**election** – {election}	1.614 (95.0%)	0.284 (96.4%)
massacre – {slaughter, massacre, mass murder, carnage, butchery}, {murder, slaying, execution}, {homicide}	**war** – {war, warfare}	1.242 (50.2%)	0.267 (91.4%)

tedly imperfect, the above suggests one potential quantitative evaluation method for computationally identified metaphors.

The only metaphors dealt with in this portion of the evaluation are those from the MILITARY source domain. Unfortunately, metaphors from the SCIENCE source domain cannot be evaluated in the same manner, as there exists no expert linguistic analysis with which to compare them. Instead, these metaphors are evaluated via subjective assessment by non-expert human subjects, as described in the following section.

5.2 Subjective Assessment by Non-expert Human Subjects

To reiterate, the primary goal of CMI is to identify and draw human readers' attention to potential conceptual metaphors. In some cases, there may be specific expectations about the metaphors one expects to find in a corpus, such as described above in the first portion of the evaluation. One of CMI's key strengths, though, is its ability to identify potential metaphors when there is little expectation about the specific metaphors to be identified. This section attempts to evaluate this ability through subjective assessment of computationally identified metaphors by non-expert human subjects.

The methods used here are not presented as the best or the only possible technique for evaluating CMI. Rather, they represent a first step in the development of subjective evaluation methods, and they provide a baseline against which to compare future work.

5.2.1 Methods. Subjects were given a brief introduction to conceptual metaphor, using the example of money is a liquid. Subjects were then presented a single computationally identified metaphor and three supporting example sentence fragments, similar to Tables 11 and 13, except without the explication of the verb-relation mediating the mapping. Metaphors were presented in the format "<target> is like <source>," for example, "a state is like a battle," since the computationally identified metaphors may be novel to participants and the wording "is like" rather than "is" aids in novel metaphor comprehension (Bowdle and Gentner 2005; Gentner et al. 2001).

Subjects were asked to rate computationally identified metaphors along four criteria. Each of these criteria was chosen to examine a different facet of the metaphors vis-à-vis the goals of CMI, and subjects' ratings along each criterion should correlate with the confidence score assigned by CMI. First, does the metaphor make sense? If computationally identified metaphors are non-sensical, they will not likely be effective at fostering critical thinking or creativity. Second, is the metaphor expected? This criterion was included for comparison with other measures, e.g., can a metaphor be unexpected but still make sense? Third, does the metaphor provide some new insight or novel understanding of the situation? Metaphors that are seen as insightful may be more effective at fostering critical or creative thinking. Fourth, is the metaphor confusing? Metaphors that subjects see as confusing may help give insight into where CMI breaks down and how it might be improved. Additionally, this criterion should be opposite the others; a metaphor that makes sense should not be confusing. Thus, if a subject says that a metaphor both makes perfect sense and is highly confusing, it provides a means of weeding out responses that are likely invalid. Ratings for each criterion were given using a Likert scale from 1 to 7. For example, for the question "Does the metaphor make sense?" subjects chose a response ranging from "1 – the metaphor makes no sense" to "7 – the metaphor makes perfect sense." See Appendix 7 for the full text of the survey.

Subjects were recruited via Amazon's Mechanical Turk service (http://www.mturk.com). On Mechanical Turk, requesters put up small, relatively simple tasks, often ones which cannot readily be done by current computers, and workers complete those tasks for a small compensation. For example, a worker might be paid $0.05 for looking at an image and answering the question, "Is there a fish in this picture?" thereby helping to tag the image. Recently, Mechanical Turk has been used to replicate classic social science studies (Paolacci and Warglien 2009) and to recruit subjects for usability experiments (Kittur, Chi, and Suh 2008). For this evaluation, the above described survey was posted on Mechanical Turk, and workers were offered $0.20 to complete the survey, i.e., to rate a single metaphor. The task was submitted to Mechanical Turk such that each of the 21 metaphors (8 from MILITARY, 13 from SCIENCE) was assessed by 20 subjects.

Some potential issues exist with using Mechanical Turk to recruit participants for academic studies. Since subjects are paid, they may be completing the study only for the compensation and thus will likely rush through the survey without focusing on properly completing the task. This concern, however, also applies to any human subjects study where participants are compensated in some way. Additionally, Mechanical Turk provides the total time a worker took to complete a task, and the average time to completion gives an indication of whether participants completed the task as directed. Furthermore, as described above, the survey included a cross-check measure enabling the identification and removal of likely invalid responses.

These data are analyzed with two goals in mind. The primary purpose is to determine if the confidence scores assigned by CMI correlate with the assessments of non-expert human subjects; specifically, confidence should correlate positively with the Sensible, Expected, and Insightful criteria, and negatively with the Confusing criterion. The secondary goal here is to examine relationships within the data–specifically, correlations between the different criteria, as well as differences between metaphors from the MILITARY and SCIENCE domains–in order to understand better how human subjects perceive and react to computationally identified metaphors.

5.2.2 Results. As described above, 20 subject ratings were acquired for each of the 21 metaphors. After checking for validity (as described above), 46 (11.0%) of the responses were judged invalid and thrown out. No metaphor had valid data for all 20 responses, and no metaphor had valid data for fewer than 16 responses. Mean time spent on task was 243.98 seconds (median 168, std. dev. 284.20), suggesting that most subjects completed the task as required.

Figure 1 shows subjects' ratings of each metaphor in each of the four criteria. All 21 metaphors, the 8 from the MILITARY domain and the 11 from the SCIENCE domain, are listed in order of confidence, with the highest at the top. Metaphorical mappings are listed here in the same format in which they were shown to participants, i.e., using "is like" rather than "is" to compare the target to the source. Within each domain, metaphors are ordered from top to bottom by the confidence score assigned by CMI. We see that, by and large, these ratings agree with what we might expect; mappings that seem intuitively reasonable, such as OHIO IS LIKE A BATTLE, are judged to be more sensible and expected and less confusing. Conversely, seemingly less reasonable mappings, such as AN INDIVIDUAL IS LIKE A PRINCIPLE, are less sensible and expected and more confusing. Furthermore, there is a general trend apparent; metaphors with higher confidence scores are more sensible, expected, and insightful, and less confusing, while the opposite is true for those with low confidence scores. This relationship is examined in more detail below (see Table 15).

Cronbach's alpha was used to test for internal consistency, with scores for the Confusing criterion inverted since we expect scores for this criterion to follow the opposite trend as the other three. These survey responses have α = 0.895. This value indicates that the survey is internally consistent, i.e., that raters generally agree with one another, without being redundant, i.e., no two criteria measure the exact same property.

The primary purpose of this portion of the evaluation is to determine how well the confidence scores assigned by CMI align with subjective assessment by non-expert human subjects. We tested the Pearson correlation coefficient between ratings assigned by subjects for each criterion and the confidence score assigned by CMI. Additionally, subjects' ratings were tested against each other using intraclass correlations (ICC) (Shrout and Fleiss 1979); subjects' ratings could not be correlated with one another using a standard Pearson correlation because, due to the nature of task deployment on Mechanical Turk, each subject rated only a single metaphor.

The intraclass correlations serve two purposes. First, since ICC values are similar to Pearson r values, the ICCs provide rough theoretical upper bounds for the correlations between confidence and non-expert human subjects' ratings. Second, since ICC values are used to assess rater reliability, these scores give an additional indication beyond the Cronbach's alpha calculation above, as to whether subjects' ratings were consistent for

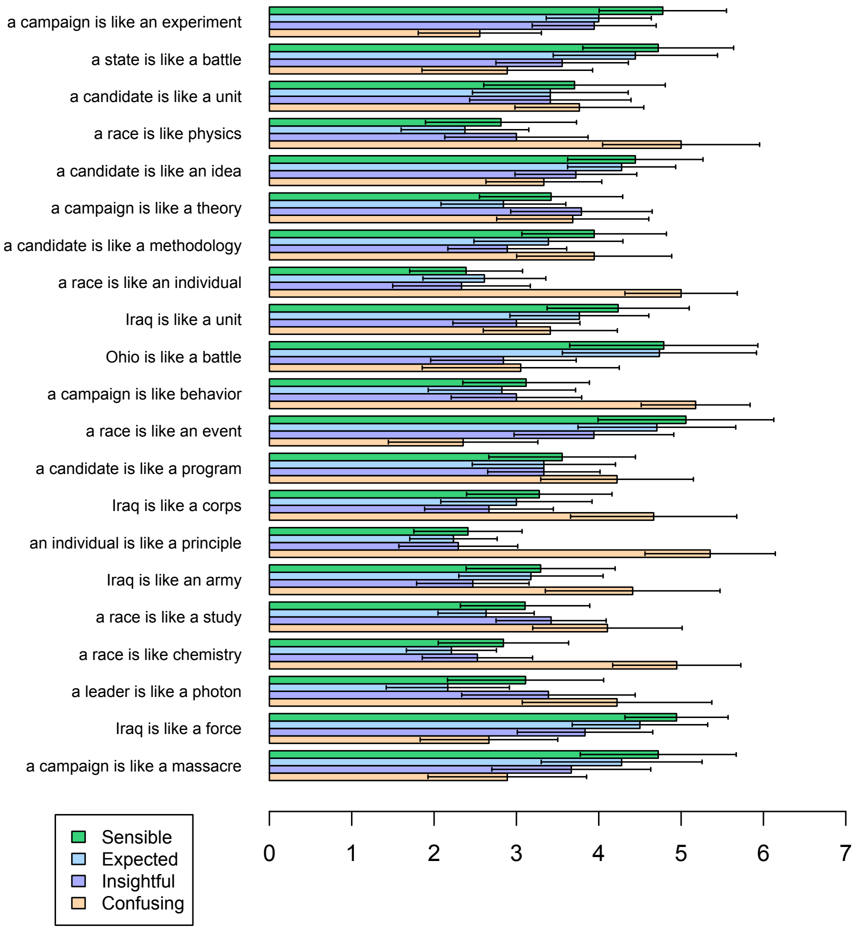

Figure 1
Human subjects' ratings of identified metaphors in political blogs from the MILITARY and SCIENCE domains, with 95% confidence intervals. Metaphors are arranged in order of confidence score, with highest at the top and lowest at the bottom.

each criterion individually. As described above, some of the survey data were deemed invalid and thrown out, leaving each metaphor with at least 16 ratings. Therefore, intraclass correlations were only calculated using 16 ratings for each metaphor; valid ratings for any metaphor beyond 16 were randomly excluded from the ICC calculations.

Table 15 lists correlations between each criterion and confidence scores assigned by CMI, as well as subjects' ICCs for each criterion. All four correlations are statistically significant at the p = 0.05 level and in the expected directions: computationally identified metaphors with higher confidence scores are more sensible, more expected, more insightful, and less confusing. However, the ICC for Insightful is low and its p-value is high relative to the other criteria's ICCs, suggesting that subjects were not in agreement with one another about which metaphors were insightful. Thus, the results for the Insightful criterion are likely not as reliable as those for the other criteria.

Table 15
Correlations between each of four criteria by which subjects rated metaphors and confidence scores assigned by CMI, as well as intraclass correlations among subjects to provide theoretical upper bounds. All correlations are statistically significant and in the expected direction.

Criterion	Correl w/ Conf	p-value	ICC	p-value
Sensible (does the metaphor make sense?)	0.127	0.0140	0.187	<0.001
Expected (would you expect this metaphor?)	0.111	0.0313	0.220	<0.001
Insightful (does the metaphor provide some insight?)	0.104	0.0438	0.057	0.0175
Confusing (is the metaphor confusing?)	-0.150	0.0036	0.212	<0.001

Table 16
Cross-correlations between different criteria for subjects' assessments of computationally identified metaphors. All correlations are significant at p<0.001.

	Sensible	Expected	Insightful
Expected	0.811		
Insightful	0.634	0.481	
Confusing	-0.822	-0.743	-0.570

Examining cross correlations within the data, shown in Table 16, reveal what the results in Table 15 would lead one to expect. All criteria were positively correlated with one another, except Confusing, which was negatively correlated with all other criteria. These correlations call into question the hypothesis that the most insightful metaphors might be the most unexpected.

These data were also used to conduct an overall comparison of metaphors from the MILITARY domain with those from the SCIENCE domain. Table 17 compares subjects' assessments of metaphors from MILITARY and SCIENCE. MILITARY metaphors were significantly more expected, more sensible, and less confusing. Since SCIENCE metaphors were less expected, they might thereby also be more insightful. However, there was no significant difference in terms of how insightful subjects found metaphors from the two domains. Since the relatively low ICC value for subjects' Insightful ratings (see Table 15) calls into question the reliability of this criterion, this specific difference should not be seen as a strongly conclusive result. Nonetheless, these results as a whole align with expectations and reaffirm that CMI functions as intended.

5.3 Discussion

The evaluation presented here demonstrates that CMI effectively performs its intended function in terms of identifying potential conceptual metaphors. The first portion of the evaluation shows that CMI can identify metaphors that align with those identified in previous expert linguistic analysis. The second portion of the evaluation shows that the confidence score assigned by CMI to potential metaphors correlates with assessments by non-expert human subjects.

Table 17
Differences in the means of the four criteria between metaphors from the two source domains used here. MILITARY metaphors were more sensible, more expected, and less confusing; there was no significant difference in how insightful the metaphors were.

	MILITARY	SCIENCE	t-test p-value
Sensible	4.23	3.46	<0.001
Expected	3.93	3.04	<0.001
Insightful	3.18	3.20	0.914
Confusing	3.46	4.14	0.00148

There are important justifications for the choice of evaluation methods used here and why potential alternatives were not used. For example, one might argue for using methods akin to precision vs. recall assessments: what percentage of the computationally identified metaphors are valid, and what percentage of valid metaphors can CMI identify? While potentially desirable, such an approach goes against the underlying stance on conceptual metaphor held by Lakoff and Johnson (1980).

Conceptual metaphor is fundamentally an intersubjective phenomenon. Therefore, any attempt to identify conceptual metaphors cannot be evaluated in a purely objective fashion but rather must be done in a manner that resonates with that intersubjectivity while still providing a quantitative assessment. For example, conducting a precision vs. recall evaluation would require a definitive, objective list of valid metaphors in a text, and the concomitant implication that no other metaphors are present. However, constructing such a list would, in practice, prove prohibitively difficult, as there would be no practical way to consider metaphors from every conceivable source domain. Furthermore, the question of which metaphors are valid cannot be settled purely objectively. Rather, determinations of validity must take into account the intersubjective nature of metaphor and the stance that what counts as a valid metaphor arises from interactions between a text and a reader. The methods used here were chosen in an attempt to adhere to this perspective while simultaneously providing a quantitative baseline against which to compare future work on improving CMI. However, the authors certainly do not rule out the possibility of other alternative methods of evaluation; some possibilities are described in the subsequent section on future work.

6. Future Work

Since computational metaphor identification is a relatively novel technique, there are numerous and varied opportunities for future work. This section outlines three such areas: technical improvements to CMI, methods for evaluation computationally identified metaphors, and potential applications of CMI.

6.1 Improvements to CMI

The technique presented here draws on and incorporates a wide variety of previous research in computational linguistics and conceptual metaphor. As such, there are many aspects of the implementation that could be altered in a number of ways in order to improve CMI, such as the inclusion of named entity recognition (Etzioni et al. 2005) or semantic role labeling (Gildea and Jurafsky 2002; Toutanova, Haghighi, and Manning

2008). This subsection describes two areas for such potential improvements: selectional preference learning and synset clustering.

6.1.1 Selectional Preference Learning. Computational metaphor identification hinges on mapping selectional preferences from a source to a target corpus. The technique used here is based on Resnik's (1993), since his was used in CorMet (Mason 2004), by which many aspects of CMI are informed. However, other selectional preference learning techniques have also been developed, e.g., (Brockmann and Lapata 2003; Light and Greiff 2002). Mason (2004) argues against the use of approaches such as Li and Abe's (1998) that use a tree cut across the WordNet hierarchy, because "it is difficult to find clusters of (possibly hypernymically related) nodes representing a selectional preference... because the tree cut includes exactly one node one each path from each leaf node to the root" (Mason 2004, page 28). Similar arguments could be applied against other approaches, such as Clark and Weir's (2002) class-based probability technique. While synset clustering may still be feasible using such methods, the adaptations that would be necessary are beyond the scope of this article.

One technique reviewed by neither Light and Greiff (2002) nor Brockmann and Lapata (2003) is that of Agirre and Martinez (2001), which calculates selectional preferences of classes of verbs rather than of specific, individual verbs. However, this algorithm is trained on the Semcor corpus (Miller et al. 1993). Since most corpora are not disambiguated in terms of WordNet senses, and since word sense disambiguation is still an active area of research with no overwhelmingly successful approach (Navigli 2009), it is unlikely that Agirre and Martinez's or similar approaches would be effective here.

A further complication is the relationship between selectional preference learning and finding sentence fragments from the corpora that exemplify the computationally identified metaphors (see Section 3.4.1). The technique of finding example sentences used here essentially inverts the process of selectional preference learning. Any change in the selectional preference learning technique should be accompanied by a change in the technique for finding example sentences fragments.

6.1.2 Synset Clustering. CMI uses vectors of selectional associations to create conceptually coherent clusters of synsets based on the verbs for which they select. The implementation described here uses nearest two neighbors clustering (Jain, Murty, and Flynn 1999), as it bears the most similarity to Mason's (2004) approach. However, a number of other clustering techniques might be used. Nearest neighbor, and similar variants, are agglomerative clustering techniques, meaning that each data point starts as a singleton cluster containing only that data point and that similar enough clusters are progressively merged to form larger clusters. As an alternative, divisive clustering could be used, where the entire data set begins as one cluster, which is then progressively divided into smaller clusters. A third option would be k-means clustering or one of its variants (Jain, Murty, and Flynn 1999), which could provide another method of varying cluster composition through adjusting k (i.e., the number of clusters formed). Furthermore, recent work suggests that spectral clustering may be highly effective for clustering based on selectional preference calculations (Sun and Korhonen 2009).

Alternatively, a non-hierarchical approach could be used that generated non-mutually exclusive clusters. For example, some methods for topic modeling (Steyvers et al. 2004) view topics as probabilistic distributions over word occurrences, wherein the occurrence of a single word within a document implies that the document may belong to one of many topics with varying probabilities. Similarly, it may be possible to represent concepts in a corpus not as mutually exclusive clusters of synsets but as probabilistic distributions over synsets, such that the occurrence of a synset would imply the occurrence of multiple different concepts with varying probabilities. Care would ned to be taken in how such a clustering approach would impact other aspects of the CMI process (mapping source concepts to target concepts, finding example sentences, cluster labeling, etc.), but such an approach may lead to more flexibly, more fluidly defined source and target concepts.

Not only could different clustering algorithms be used, but the similarity metric for comparing synsets could also be varied. CorMet (Mason 2004) uses dot product of selectional associations; CMI uses an alternative metric that rewards similarity and penalizes difference commensurately (see Section 3.1.5). Leach, Hunter, and Landsman (1999) describe a number of different clustering similarity metrics, as well as potential means of comparing them. In this context, though, comparing the clustering results themselves may not be as important as comparing the ultimate results of CMI, i.e., the computationally identified metaphors. The next subsection discusses some of the issues involved in evaluating CMI and potential future improvements on the techniques used here.

6.2 Evaluation of CMI

Because the development of computational techniques to identify conceptual metaphors is a novel and under-explored research area, there do not exist standard, accepted evaluation metrics. This article has suggested two potential means of evaluating CMI: via comparison with previous expert linguistic analysis using semantic similarity metrics, and via subject assessment by human subjects along several criteria. While the authors believe these are an important first step, they also assert that more could be done in terms of developing evaluation metrics.

In particular, the means of comparing computationally identified metaphors to previous linguistic analysis could be expanded upo. The evaluation here relies on comparing computationally identified metaphors with previous linguistic analysis of a similar corpus. It might be preferable instead to have CMI and an expert linguist analyze the exact same corpus. Furthermore, in addition to using quantitative similarity metrics, one could also ask the linguist to assess CMI's output, both in a quantitative format similar to that used with the non-expert human subjects, and in a qualitative format.

Indeed, qualitative evaluations of CMI may be almost as important as quantitative. Knowing *how much* a computationally identified metaphor differs from an expected or established metaphor might not be as informative as knowing *in what ways* they differ. For example, do the identified and expected metaphors differ in their levels of superordination (Lakoff 1993), such as the difference between A STATE IS A BATTLE and OHIO IS A BATTLE? Do they differ in terms of surface aspects, structural aspects, or both (Blanchette and Dunbar 2000; Gentner, Rattermann, and Forbus 1993; Holyoak and Koh 1987)? Such qualitative differences, while difficult to compare directly and precisely, may lead to further insights about alternative computational techniques to incorporate into CMI.

It is important to emphasize that any evaluation technique should adhere to the philosophical and epistemic stance of metaphor as intersubjective and experiential (Lakoff and Johnson 1980). The evaluation conducted here attempts to do so by emphasizing subjective assessment of potential conceptual metaphors. Future improvements to CMI should be evaluated with a similar sensibility.

6.3 Empiricism in Cognitive Linguistics

It has been argued that "questions about [metaphor] are questions requiring empirical study; they cannot be answered adequately by mere a priori theorizing" (Lakoff and Johnson 1980, page 246). However, most cognitive linguistics research is not particularly explicit about its empirical grounding or its methods of inquiry (Gibbs 2007). This is not to say that the field does not have rigorous, scientific methods, but, according to Gibbs, these methods should be more clearly explicated.

CMI could provide one means of addressing this issue, simultaneously helping make the field more empirically focused and more explicit about their empirical methods. Some work has already been done using corpus-based methods in analysis of conceptual metaphors (Deignan 2005), applying traditional corpus analysis methods, such as word counts and co-occurrence frequencies, as a resource for identifying potential metaphors. While such work represents an important advance in cognitive linguistics, the research presented here improves upon those methods by using a technique specifically designed to identify potential metaphors. This nascent area of investigation provides numerous opportunities for research that can expand our understanding of metaphor in language and thought.

However, any such computational analysis should not be seen as a definitive, objective statement of the metaphors present in a corpus. When conducting any such analysis, it is important to bear in mind the role of the analyst in interpreting computationally identified metaphors. As described above, CMI does not state definitively *the* metaphors in a corpus, but rather suggests *potential* conceptual metaphors. Thus, CMI should be seen as a tool to support such analysts, not a replacement for them. This perspective maintains the underlying stance taken in this work: the role of the computer is to identify potential patterns in data, and the role of the human user is to interpret and make meaning from those patterns.

7. Conclusion

This article has presented computational metaphor identification (CMI), a technique that identifies potential conceptual metaphors in written text. The evaluation performed here demonstrates that the CMI can identify metaphors that bear strong similarity to those identified in previous linguistic analysis, and that the confidence scores assigned by CMI to potential metaphors correlate significantly with assessments of those metaphors by non-expert human subjects. Computational metaphor identification suggests two promising directions for future research.

First, it provides an opportunity to apply more empirical, data-driven methods in cognitive linguistics. While not without rigorous, scientific methods of inquiry, cognitive linguistics, it has been argued, has not been incredibly explicit about these methods or their grounding in empirical data (Gibbs 2007). CMI is not a perfect solution; the results presented here show that, while effective at identifying conceptual metaphors, there is room for improvement. However, this work provides an important first step in the development of such techniques.

Second, as described above, CMI represents a break from previous computational linguistics work on metaphor and, indeed, from the general approach adopted in much AI research. The purpose her is not to enable *computers* to think *with* metaphors, but rather to encourage *people* to think *about* metaphors. While metaphor plays a pivotal role in human cognition, CMI is but one example of how this sensibility might be applied.

The development, deployment, and evaluation of other, similar systems can allow for exploration of alternative configurations between computation and human thought.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0757646, by the Donald Bren School of Information and Computer Sciences, and by the California Institute for Telecommunication and Information Technology (Calit2).

Appendix A: List of blogs with URLs from which analyzed posts were collected.

AMERICAblog – http://www.americablog.com/

The Daily Dish by Andrew Sullivan –
http://andrewsullivan.theatlantic.com/the_daily_dish/

Daily Kos – http://www.dailykos.com

Firedoglake – http://firedoglake.com

The Huffington Post – http://www.huffingtonpost.com/theblog/

The Liberal OC – http://www.theliberaloc.com

Pandagon – http://pandagon.net/index.php/site/index/

Paul Krugman – http://krugman.blogs.nytimes.com

Talking Points Memo – http://talkingpointsmemo.com

Think Progress – http://thinkprogress.org

Appendix B: Stopword List

a A about abr across after afterwards again against al all almost alone along already also although always am among amongst amoungst an and another any anybody anyhow anyone anything anyway anywhere are around as at b B be became because become becomes becoming been before beforehand being below beside besides between beyond both br but by c C cannot co could couldnt d D de describe did didnt do does doesnt done dont during e E each eg eight eighth either else elsewhere etc ever every everyone everything everywhere except f F fifth first five for formerly four fourth friday Friday from further g G get got h H had has hasnt havent he hence her here hereby herein hereupon hers herself hes him himself his how however i I Im im if in inc indeed into is isnt it its itself ive j J just k K l L latter latterly ‘less lets ltd m M made many me monday Monday more moreover most mostly my myself n N namely neither never nevertheless next nine ninth no nobody noone none nor not nothing o O of off often on once one only onto or other others otherwise our ours ourselves out own p P per perhaps please pm put q Q r R rather re really s S same saturday Saturday second see seem seemed seeming seems serious seven seventh several she shes should since sincere six sixth so some somebody somehow someone something sometime sometimes such sunday Sunday t T ten tenth than that thats the their them themselves thence there thereafter thereby therefore therein thereupon these they theyve thin thing third this those though three through throughout thru thursday Thursday thus to today together too tomorrow tonight toward towards two tuesday Tuesday under until u U up upon v V very via w W was wasnt we wednesday Wednesday were what whatever when whence whenever where whereafter whereas whereby wherein whereupon wherever whether which while whither who whoever whom whose why with within without would x X y Y year years yesterday yet you youd your youre yours yourself yourselves youve z Z zero

Appendix C: Full Text of Survey

Text in <Angle_brackets> was replaced with metaphor-specific content, e.g., <Domain> would be replaced with either Military or Science.

Help Evaluate Our System for Identifying Metaphors

The following survey is part of a research study being conducted at the University of California, Irvine. The survey involves answering questions about posts on political blogs. There is minimal risk, participation is entirely voluntary, and you may quit at any time. You will receive $0.20 for completion of this survey. The research team, authorized UCI personnel, the study sponsor (if applicable), and regulatory entities such as the FDA, may have access to your study records to protect your safety and welfare. Any information derived from this research project that personally identifies you will not be voluntarily released or disclosed by these entities without your separate consent, except as specifically required by law. If you have any comments, concerns, or questions regarding the conduct of this research, please contact the Mechanical Turk requester for this HIT. If you are unable to reach the researchers listed at the top of the form and have general questions, or you have concerns or complaints about the research, or questions about your rights as a research subject, please contact UCI's Office of Research Administration by phone, (949) 824-6662, by e-mail at IRB@rgs.uci.edu or at University Tower - 4199 Campus Drive, Suite 300, Irvine, CA 92697-7600.

We'll start with a few quick questions.

How often do you read political blogs? Please choose one of the following:

Never

Once or twice per month

Once a week

Once every two to three days

About once a day

Several times per day

About how many political blogs do you read regularly? Please choose one of the following:

1 to 2

2 to 5

5 to 10

10 to 20

20 to 50

50 to 100

More than 100

Please list examples of some of the political blogs you read regularly:

Now on to the metaphors. People often use conceptual metaphors to frame how they talk about and think about abstract ideas. For example, when talking about money, you might say, "He poured all his savings into bonds," "They froze my assets," or, "Capital freely flowed between investors." These words describe money as a liquid, framing our understanding of money in terms of our experiences with physical liquids.

We have developed a computational system to identify patterns of words that might indicate potential metaphors. We analyzed political blogs from around the 2008 election and identified several potential metaphors. Here is one <Domain> metaphor our system found (that is, a metaphor where a political concept is understood in terms of a <Domain> concept), along with example sentences from the blogs and from documents about <Domain> concepts.

<Target>	is like	<Source>
<Target_Example$_1$>		<Source_Example$_1$>
<Target_Example$_2$>		<Source_Example$_2$>
<Target_Example$_3$>		<Source_Example$_3$>

Please rate this metaphor along the following criteria (1 being the lowest, 7 being the highest):

Does the metaphor make sense? Please choose one of the following:

1 - the metaphor makes no sense

2

3

4

5

6

7 - the metaphor makes perfect sense

Is the metaphor one that you would expect to see? Please choose one of the following:

1 - the metaphor is completely unexpected

2

3

4

5

6

7 - the metaphor is exactly what I would expect

Does the metaphor provide some new or interesting insight? Please choose one of the following:

1 - the metaphor is not insightful at all

2

3

4

5

6

7 - the metaphor is highly insightful

Do you find the metaphor confusing? Please choose one of the following:

1 - the metaphor is not confusing at all

2

3

4

5

6

7 - the metaphor is highly confusing

Overall, do you think this is a good metaphor? Why or why not?

Thank you for your time.

References

Agirre, E. and D. Martinez. 2001. Learning class-to-class selectional preferences. In *Workshop on Computational Natural Language Learning*, volume 7, pages 1–8, Toulouse, France.

Baumer, Eric P. S., Jordan Sinclair, David Hubin, and Bill Tomlinson. 2009. metaViz: visualizing computationally identified metaphors in political blogs. In *Computational Science and Engineering, IEEE International Conference on*, volume 4, pages 389–394, Los Alamitos, CA, USA. IEEE Computer Society.

Baumer, Eric P. S., Jordan Sinclair, and Bill Tomlinson. 2010. "America is like metamucil:" critical and creative thinking about metaphor in political blogs. In *ACM Conf on Human Factors in Computing Systems (CHI)*, Atlanta, GA. ACM Press.

Baumer, Eric P. S., Bill Tomlinson, Janice Hansen, and Lindsey E. Richland. 2009. Fostering metaphorical creativity using computational metaphor identification. In *ACM Conf on Creativity and Cognition (C&C)*, Berkeley, CA, October. ACM.

Black, M. 1962. Metaphor. In M. Black, editor, *Models and Metaphors*. Cornell University Press, Ithaca.

Blanchette, I. and Kevin Dunbar. 2000. How analogies are created: The roles of structural and superficial similarity. *Memory and Cognition*, 29:730–735.

Bowdle, Brian F. and Dedre Gentner. 2005. The career of metaphor. *Psychological Review*, 112(1):193–216.

Brockmann, Carsten and Mirella Lapata. 2003. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, pages 27–34, Budapest, Hungary. Association for Computational Linguistics.

Budanitsky, Alexander and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, March.

Clark, Stephen and David Weir. 2002. Class-Based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206, June.

de Marneffe, M.-C., B. MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Lang Res and Eval (LREC)*, Genoa, Italy.

Dehghani, Morteza, Emmett Tomai, Kenneth D. Forbus, and Matthew Klenk. 2008. An integrated reasoning approach to moral Decision-Making. In *Proceedings of the AAAI National Conference on Artificial Intelligence*.

Deignan, A. 2005. *Metaphor and Corpus Linguistics*. John Benjamins, Amsterdam and Philadelphia.

Etzioni, O., M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates. 2005. Unsupervised Named-Entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134.

Falkenhainer, B., Kenneth D. Forbus, and Dedre Gentner. 1989. The Structure-Mapping engine: Algorithm and examples. *Artificial Intelligence*, 41:1–63.

Fass, D. 1991. Met*: A method for discriminating metonymy and metaphor by computer. *Comp Ling*, 17(1):49–90.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Francis, Winthrop Nelson and Henry Kučera. 1982. *Frenquency Analysis of English Use: Lexicon and Grammar*. Houghton Mifflin, Boston.

Gedigian, M., J. Bryant, S. Narayanan, and B. Ciric. 2006. Catching metaphors. In *3rd Workshop on Scalable Natural Language Understanding*, New York City. Association for Computational Linguistics.

Gentner, Dedre. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7:155–170.

Gentner, Dedre, Brian F. Bowdle, Phillip Wolff, and C. Boronat. 2001. Metaphor is like analogy. In Dedre Gentner, Keith J. Holyoak, and Boicho Kokinov, editors, *The Analogical Mind*. MIT Press, Cambridge, MA, pages 199–253.

Gentner, Dedre, M.J. Rattermann, and Kenneth D. Forbus. 1993. The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25:524–575.

Gibbs, Raymond W. 1984. Literal meaning and psychological theory. *Cognitive Science*, 8:275–304.

Gibbs, Raymond W. 2007. Why cognitive linguists should care more about empirical methods. In Monica Gonzalez-Marquez, Irene Mittelberg, Seana Coulson, and Michael J. Spivey, editors, *Methods in Cognitive Linguistics*. John Benjamins, Amsterdam, pages 2–18.

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Holyoak, Keith J. and Kyunghee Koh. 1987. Surface and structural similarity in analogical transfer. *Memory and Cognition*, 15(4):332–340.

Howe, Nicholas. 1988. Metaphor in contemporary american political discourse. *Metaphor and Symbol*, 3(2):87–104.

Jain, A.K., M.N. Murty, and P.J. Flynn. 1999. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323.

Kilgarriff, A. 2003. BNC word frequency list. http://www.kilgarriff.co.uk/bnc-readme.html, April.

Kittur, Aniket, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *ACM SIGCHI Conf*, pages 453–456, Florence, Italy. ACM.

Klein, D. and Christopher D. Manning. 2002. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, Vancouver, BC, Canada.

Klein, D. and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Mtg of the Assoc for Comp Ling*, Sapporo, Japan.

Krishmakumaran, S. and X. Zhu. 2007. Hunting elusive metaphors using lexical resources. In X. Lu and A. Feldman, editors, *Computational Approaches to Figurative Language, Workshop at HLT/NAACL 2007*, Rochester, NY.

Kuehne, S.E. and Kenneth D. Forbus. 2004. On the representation of physical quantities in natural language text. In *Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society*, Chicago, IL, August.

Kurzweil, Ray. 1990. *The Age of Intelligent Machines*. MIT Press, Cambridge, MA.

Lakoff, George. 1993. The contemporary theory of metaphor. In A. Ortony, editor, *Metaphor and thought, 2nd. ed.* Cambridge Univ Press, New York, pages 202–251.

Lakoff, George. 2002. *Moral Politics: How Liberals and Conservatives Think*. University of Chicago Press, Chicago.

Lakoff, George, J. Espenson, and A. Schwartz. 1991. The master metaphor list. Technical report, Cognitive Linguistics Group, University of California, Berkeley, Berkeley.

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago, IL, 2003 edition.

Lakoff, George and Mark Turner. 1989. *More Than Cool Reason: A Field Guide to Poetic Metaphor*. University of Chicago Press, Chicago and London.

Leach, Sonia, Lawrence Hunter, and David Landsman. 1999. Comparison of clustering metrics and unsupervised learning algorithms on genome-wide gene expression level data. In *Proceedings of the sixteenth national conference on Artificial intelligence (AAAI)*, page 966, Orlando, Florida, United States. American Association for Artificial Intelligence.

Leacock, C. and M. Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Databes*. MIT Press, Cambridge, MA, pages 265–283.

Li, H. and N. Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244, June.

Light, M. and W. Greiff. 2002. Statistical models for the induction and use of selectional preferences. *Cognitive Science*, 26(3):269–281.

Lin, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, Madison, WI.

Lockwood, Kate and Kenneth D. Forbus. 2009. Multimodal knowledge capture from text and diagrams. In *Proceedings of the fifth international conference on Knowledge capture*, pages 65–72, Redondo Beach, California, USA. ACM.

Martin, J.H. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press, San Diego, CA.

Martin, J.H. 1994. Metabank: A knowledge base of metaphoric language conventions. *Computational Intelligence*, 10(2):134–149.

Mason, Z.J. 2004. CorMet: a computational, corpus-based conventional metaphor extraction system. *Comp Ling*, 30(1):23–44, March.

Miller, G.A., C. Leacock, R. Tengi, and R.T. Bunker. 1993. A semantic concordance. In *Workshop on Human Language Technology*, Princeton, NJ, March.

Minsky, M. 1968. *Semantic information processing*. MIT Press, Cambridge, MA.

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Ortony, A. 1980. Some psycholinguistic aspects of metaphor. In R.P. Honeck and H.R. Robert, editors, *Cognition and Figurative Language*. Erlbaum Associates, Hillsdale, NJ, pages 69–83.

Ortony, A., R.J. Vondruska, M.A. Foss, and L.E. Jones. 1985. Salience, similes, and the asymmetry of similarity. *Journal of Memory and Language*, 24:569–594.

Paolacci, Gabriele and Massimo Warglien. 2009. http://experimentalturk.wordpress.com/.

Reddy, M.J. 1969. A semantic approach to metaphor. In *Chicago Linguistic Society Collected Papers*. Chicago University Press, Chicago, pages 240–251.

Reddy, M.J. and A. Ortony. 1979. The conduit metaphor: A case of frame conflict in our language about language. In *Metaphor and Thought*. Cambridge University Press, Cambridge, pages 284–297.

Resnik, P. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Dissertation, University of Pennsylvania, Department of Computer and Information Science.

Rich, E. and K. Knight. 1991. *Artificial Intelligence*. McGraw-Hill, New York, second edition edition.

Searle, J.R. and A. Ortony. 1979. Metaphor. In *Metaphor and Thought*. Cambridge University Press, London.

Shrout, P.E. and J.L. Fleiss. 1979. Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, 86:420–428.

Steyvers, Mark, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315, Seattle, WA, USA. ACM.

Sun, Lin and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 638–647, Singapore, August. ACL.

Toutanova, Kristina, Aria Haghighi, and Christopher D. Manning. 2008. A global joint model for semantic role labeling. *Comput. Linguist.*, 34(2):161–191.

Turney, Peter D. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Turney, Peter D. 2008. The latent relation mapping enginge: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33:615–655.

Wilcox, Harold. 1995. *Salience Imbalance and Metaphor*. Dissertation, University of Colorado at Boulder, Department of Linguistics.