Statistical Metaphor Processing

Ekaterina Shutova*
University of Cambridge

Simone Teufel*
University of Cambridge

Anna Korhonen*
University of Cambridge

Metaphor is highly frequent in language, which makes its computational processing indispensable for real-world NLP applications addressing semantic tasks. Previous approaches to metaphor modeling rely on task-specific hand-coded knowledge and operate on a limited domain or a subset of phenomena. We present the first integrated open-domain statistical model of metaphor processing in unrestricted text. Our method first identifies metaphorical expressions in running text and then paraphrases them with their literal paraphrases. Such a text-to-text model of metaphor interpretation is compatible with other NLP applications that can benefit from metaphor resolution. Our approach is minimally supervised, relies on the state-of-the-art parsing and lexical acquisition technologies (distributional clustering and selectional preference induction), and operates with a high accuracy.

1. Introduction

Our production and comprehension of language is a multi-layered computational process. Humans carry out high-level semantic tasks effortlessly by subconsciously using a vast inventory of complex linguistic devices, while simultaneously integrating their background knowledge, to reason about reality. An ideal computational model of language understanding would also be capable of performing such high-level semantic tasks. With the rapid advances in statistical natural language processing (NLP) and computational lexical semantics, increasingly complex semantic tasks can now be addressed. Tasks that have received much attention so far include, for example, word sense disambiguation (WSD), supervised and unsupervised lexical classification, selectional preference induction, and semantic role labeling. In this article, we take a step further and show that state-of-the-art statistical NLP and computational lexical semantic techniques can be used to successfully model complex meaning transfers, such as metaphor.

* Computer Laboratory, William Gates Building, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK.
E-mail: {Ekaterina.Shutova, Simone.Teufel, Anna.Korhonen}@cl.cam.ac.uk.

Submission received: 28 July 2011; revised submission received: 21 April 2012; accepted for publication: 31 May 2012.

doi:10.1162/COLI_a_00124

Metaphors arise when one concept is viewed in terms of the properties of another. Humans often use metaphor to describe abstract concepts through reference to more concrete or physical experiences. Some examples of metaphor include the following.

(1) How can I *kill* a process? (Martin 1988)

(2) Hillary *brushed aside* the accusations.

(3) I *invested* myself fully in this research.

(4) And then my heart with pleasure *fills*,
And *dances* with the daffodils.
("I wandered lonely as a cloud," William Wordsworth, 1804)

Metaphorical expressions may take a great variety of forms, ranging from conventional metaphors, which we produce and comprehend every day, for example, those in Examples (1)–(3), to poetic and novel ones, such as Example (4). In metaphorical expressions, seemingly unrelated features of one concept are attributed to another concept. In Example (1), a *computational process* is viewed as something *alive* and, therefore, its forced termination is associated with the act of killing. In Example (2) Hillary is not literally cleaning the space by sweeping accusations. Instead, the accusations lose their validity in that situation, in other words Hillary *rejects* them. The verbs *brush aside* and *reject* both entail the resulting disappearance of their object, which is the shared salient property that makes it possible for this analogy to be lexically expressed as a metaphor.

Characteristic of all areas of human activity (from poetic to ordinary to scientific) and thus of all types of discourse, metaphor becomes an important problem for NLP. As Shutova and Teufel (2010) have shown in an empirical study, the use of conventional metaphor is ubiquitous in natural language text (according to their data, on average every third sentence in general-domain text contains a metaphorical expression). This makes metaphor processing essential for automatic text understanding. For example, an NLP application which is unaware that a "*leaked* report" is a "disclosed report" and not, for example, a "wet report," would fail further semantic processing of the piece of discourse in which this phrase appears. A system capable of recognizing and interpreting metaphorical expressions in unrestricted text would become an invaluable component of any real-world NLP application that needs to access semantics (e.g., information retrieval [IR], machine translation [MT], question answering [QA], information extraction [IE], and opinion mining).

So far, these applications have not used any metaphor processing techniques and thus often fail to interpret metaphorical data correctly. Consider an example from MT. Figure 1 shows metaphor translation from English into Russian by a state-of-the-art statistical MT system (Google Translate[1]). For both sentences the MT system produces literal translations of metaphorical terms in English, rather than their literal interpretations. This results in otherwise grammatical sentences being semantically infelicitous, poorly formed, and barely understandable to a native speaker of Russian. The meaning of *stir* in Figure 1 (1) and *spill* in Figure 1 (2) would normally be realized in Russian only via their literal interpretation in the given context (*provoke* and *tell*), as shown under CORRECT TRANSLATION in Figure 1. A metaphor processing component could help to

1 http://translate.google.com/.

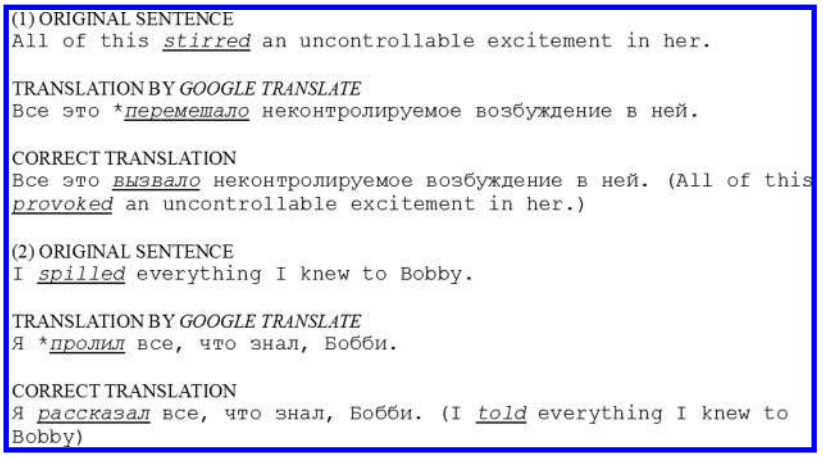

Figure 1
Examples of metaphor translation.

avoid such errors. We conducted a pilot study of the importance of metaphor for MT, by running an English-to-Russian MT system (Google Translate) on the sentences from the data set of Shutova (2010) containing single-word verb metaphors. We found that 27 out of 62 sentences (44%) were translated incorrectly due to metaphoricity. Due to the high frequency of metaphor in text according to corpus studies, such a high level of error becomes important for MT.

Examples where metaphor understanding is crucial can also be found in opinion mining, that is, detection of the speaker's attitude to what is said and to the topic. Consider the following sentences.

(5) a. Government *loosened its strangle-hold* on business. (Narayanan 1999)

b. Government deregulated business. (Narayanan 1999)

Both sentences describe the same fact. The use of the metaphor *loosened strangle-hold* in Example (5a) suggests that the speaker opposes government control of economy, however, whereas Example (5b) does not imply this. One can infer the speaker's negative attitude via the presence of a negative word *strangle-hold*. A metaphor processing system would establish the correct meaning of Example (5a) and thus discover the actual fact towards which the speaker has a negative attitude.

Because metaphor understanding requires resolving non-literal meanings via analogical comparisons, the development of a complete and computationally practical account of this phenomenon is a challenging and complex task. Despite the importance of metaphor for NLP systems dealing with semantic interpretation, its automatic processing has received little attention in contemporary NLP, and is far from being a solved problem. The majority of computational approaches to metaphor still exploit ideas articulated two or three decades ago (Wilks 1978; Lakoff and Johnson 1980). They often rely on task-specific hand-coded knowledge (Martin 1990; Fass 1991; Narayanan 1997, 1999; Barnden and Lee 2002; Feldman and Narayanan 2004; Agerri et al. 2007) and reduce the task to reasoning about a limited domain or a subset of phenomena (Gedigian et al. 2006; Krishnakumaran and Zhu 2007). So far there has been no robust statistical system operating on unrestricted text. State-of-the-art accurate parsing (Klein and Manning 2003; Briscoe, Carroll, and Watson 2006; Clark and Curran 2007), however, as well as recent work on computational lexical semantics (Schulte im Walde 2006;

Mitchell and Lapata 2008; Davidov, Reichart, and Rappoport 2009; Erk and McCarthy 2009; Sun and Korhonen 2009; Abend and Rappoport 2010; Ó Séaghdha 2010) open up many avenues for the creation of such a system. This is the niche the presented work is intending to fill.

1.1 What Is Metaphor?

Metaphor has traditionally been viewed as an artistic device that lends vividness and distinction to an author's style. This view was first challenged by Lakoff and Johnson (1980), who claimed that it is a productive phenomenon that operates at the level of mental processes. According to Lakoff and Johnson, metaphor is thus not merely a property of language (i.e., a linguistic phenomenon), but rather a property of thought (i.e., a cognitive phenomenon). This view was subsequently adopted and extended by a multitude of approaches (Grady 1997; Narayanan 1997; Fauconnier and Turner 2002; Feldman 2006; Pinker 2007) and the term **conceptual metaphor** was coined to describe it.

The view postulates that metaphor is not limited to similarity-based meaning extensions of individual words, but rather involves reconceptualization of a whole area of experience in terms of another. Thus metaphor always involves two concepts or conceptual domains: the **target** (also called the *topic* or *tenor* in the linguistics literature) and the **source** (also called the *vehicle*). Consider Examples (6) and (7).

(6) He *shot down* all of my arguments. (Lakoff and Johnson 1980)

(7) He *attacked* every weak point in my argument. (Lakoff and Johnson 1980)

According to Lakoff and Johnson, a mapping of the concept of *argument* to that of *war* is used in both Examples (6) and (7). The *argument*, which is the target concept, is viewed in terms of a *battle* (or a *war*), the source concept. The existence of such a link allows us to talk about *arguments* using *war* terminology, thus giving rise to a number of metaphors. Conceptual metaphor, or **source–target domain mapping**, is thus a generalization over a set of individual metaphorical expressions that covers multiple cases in which ways of reasoning about the source domain systematically correspond to ways of reasoning about the target.

Conceptual metaphor manifests itself in natural language in the form of **linguistic metaphor** (or metaphorical expressions) in a variety of ways. The most common types of linguistic metaphor are **lexical** metaphor (i.e., metaphor at the level of a single word sense, as in the Examples (1)–(4)), **multi-word** metaphorical expressions (e.g., "whether we *go on pilgrimage* with Raleigh or *put out to sea* with Tennyson"), or **extended** metaphor, that spans over longer discourse fragments.

Lexical metaphor is by far the most frequent type. In the presence of a certain conceptual metaphor individual words can be used in entirely novel contexts, which results in the formation of new meanings. Consider the following example.

(8) How can we build a 'Knowledge economy' if research is *handcuffed*? (Barque and Chaumartin 2009)

In this sentence the physical verb *handcuff* is used with an abstract object *research* and its meaning adapts accordingly. Metaphor is a productive phenomenon (i.e., its

novel examples continue to emerge in language). A large number of metaphorical expressions, however, become conventionalized (e.g., "I cannot *grasp* his way of thinking"). Although metaphorical in nature, their meanings are deeply entrenched in everyday use, and are thus cognitively treated as literal terms. Both novel and conventional metaphors are important for text processing, hence our work is concerned with both types. Fixed non-compositional idiomatic expressions (e.g., *kick the bucket*, *rock the boat*, *put a damper on*), however, are left aside, because the mechanisms of their formation are no longer productive in modern language and, as such, they are of little interest for the design of a generalizable computational model of metaphor.

Extended metaphor refers to the use of metaphor at the discourse level. A famous example of extended metaphor can be found in William Shakespeare's play *As You Like It*, where he first compares the world to a stage and then in the following discourse describes its inhabitants as players. Extended metaphor often appears in literature in the form of an *allegory* or a *parable*, whereby a whole story from one domain is metaphorically transferred onto another in order to highlight certain attributes of the subject or teach a moral lesson.

1.2 Computational Modeling of Metaphor

In this article we focus on lexical metaphor and the computational modeling thereof. From an NLP viewpoint, not all metaphorical expressions are equally important. A metaphorical expression is interesting for computational modeling if its metaphorical sense is significantly distinct from its original literal sense and cannot be interpreted directly (e.g., by existing word sense disambiguation techniques using a predefined sense inventory). The identification of highly conventionalized metaphors (e.g., the verb *impress*, whose meaning originally stems from printing) are not of interest for NLP tasks, because their metaphorical senses have long been dominant in language and their original literal senses may no longer be used. A number of conventionalized metaphors, however, require explicit interpretation in order to be understood by computer (e.g., "*cast* doubt," "*polish* the thesis," "*catch* a disease"), as do all novel metaphors. Thus we are concerned with both novel and conventional metaphors, but only consider the cases whereby the literal and metaphorical senses of the word are in clear opposition in common use in contemporary language.

Automatic processing of metaphor can be divided into two subtasks: *metaphor identification*, or *recognition* (distinguishing between literal and metaphorical language in text); and *metaphor interpretation* (identifying the intended literal meaning of a metaphorical expression). An ideal metaphor processing system should address both of these tasks and provide useful information to support semantic interpretation in real-world NLP applications. In order to be directly applicable to other NLP systems, it should satisfy the following criteria:

- **Provide a representation of metaphor interpretation that can be easily integrated with other NLP systems**: This criterion places constraints on how the metaphor processing task should be defined. The most universally applicable metaphor interpretation would be in the text-to-text form. This means that a metaphor processing system would take raw text as input and provide a more literal text as output, in which metaphors are interpreted.

- **Operate on unrestricted running text**: In order to be useful for real-world NLP the system needs to be capable of processing real-world data. Rather than only dealing with individual carefully selected clear-cut examples, the system should be fully implemented and tested on free naturally occurring text.
- **Be open-domain**: The system needs to cover all domains, genres, and topics. Thus it should not rely on any domain-specific information or focus on individual types of instances (e.g., a hand-chosen limited set of source-target domain mappings).
- **Be unsupervised or minimally supervised**: To be easily adaptable to new domains, the system needs to be unsupervised or minimally supervised. This means it should not use any task-specific (i.e., metaphor-specific) hand-coded knowledge. The only acceptable exception might be a multi-purpose general-domain lexicon that is already in existence and does not need to be created in a costly manner, although it would be an advantage if no such resource is required.
- **Cover all syntactic constructions**: To be robust, the system needs to be able to deal with metaphors represented by all word classes and syntactic constructions.

In this article, we address both the metaphor identification and interpretation tasks, resulting in the first integrated domain-independent corpus-based computational model of metaphor. The method is designed with the listed criteria in mind. It takes unrestricted text as input and produces textual output. Metaphor identification and interpretation modules, based on the algorithms of Shutova, Sun, and Korhonen (2010) and Shutova (2010), are first evaluated independently, and then combined and evaluated together as an integrated system. All components of the method are in principle applicable to all part-of-speech classes and syntactic constructions. In the current experiments, however, we tested the system only on single-word metaphors expressed by a verb. Verbs are frequent in language and central to conceptual metaphor. Cameron (2003) conducted a corpus study of the use of metaphor in educational discourse for all parts of speech. She found that verbs account for around 50% of the data, the rest being shared by nouns, adjectives, adverbs, copula constructions, and multi-word metaphors. This suggests that verb metaphors provide a reliable testbed for both linguistic and computational experiments. Restricting the scope to verbs is a methodological step aimed at testing the main principles of the proposed approach in a well-defined setting. We would, however, expect the presented methods to scale to other parts of speech and to a wide range of syntactic constructions, because they rely on techniques from computational lexical semantics that have been shown to be effective in modeling not only verb meanings, but also those of nouns and adjectives.

As opposed to previous approaches that modeled metaphorical reasoning starting from a hand-crafted description and applying it to explain the data, we aim to design a statistical model that captures regular patterns of metaphoricity in a large corpus and thus generalizes to unseen examples. Compared to labor-intensive manual efforts, this approach is more robust and, being nearly unsupervised, cost-effective. In contrast to previous statistical approaches, which addressed metaphors of a specific topic or did not consider linguistic metaphor at all (e.g., Mason 2004), the proposed method covers all metaphors in principle, can be applied to unrestricted text, and can be adapted to different domains and genres.

Our first experiment is concerned with the identification of metaphorical expressions in unrestricted text. Starting from a small set of metaphorical expressions, the system learns the analogies involved in their production in a minimally supervised way. It generalizes over the exemplified analogies by means of verb and noun clustering (i.e., the identification of groups of similar concepts). This generalization allows it to recognize previously unseen metaphorical expressions in text. Consider the following examples:

(9) All of this *stirred* an uncontrollable excitement in her.

(10) Time and time again he would stare at the ground, hand on hip, and then *swallow* his anger and play tennis.

Having once seen the metaphor "*stir* excitement" in Example (9) the metaphor identification system successfully concludes that "*swallow* anger" in Example (10) is also used metaphorically.

The identified metaphors then need to be interpreted. Ideally, a metaphor interpretation task should be aimed at producing a representation of metaphor understanding that can be directly embedded into other NLP applications that could benefit from metaphor resolution. We define metaphor interpretation as a paraphrasing task and build a system that discovers literal meanings of metaphorical expressions in text and produces their literal paraphrases. For example, for metaphors in Examples (11a) and (12a) the system produces the paraphrases in Examples (11b) and (12b), respectively.

(11) a. All of this *stirred* an uncontrollable excitement in her.

b. All of this provoked an uncontrollable excitement in her.

(12) a. a carelessly *leaked* report

b. a carelessly disclosed report

The paraphrases for metaphorical expressions are acquired in a data-driven manner from a large corpus. Literal paraphrases are then identified using a selectional preference model.

This article first surveys the relevant theoretical and computational work on metaphor, then describes the design of the identification and paraphrasing modules and their independent evaluation, and concludes with the evaluation of the integrated text-to-text metaphor processing system. The evaluations were carried out with the aid of human subjects. In the case of identification, the subjects were asked to judge whether a system-annotated phrase is a metaphor. In case of paraphrasing, they had to decide whether the system-produced paraphrase for the metaphorical expression is correct and literal in the given context. In addition, we created a metaphor paraphrasing gold standard by asking human subjects (not previously exposed to system output) to produce their own literal paraphrases for metaphorical verbs. The system paraphrasing was then also evaluated against this gold standard.

2. Theoretical and Computational Background

2.1 Metaphor and Polysemy

Theorists of metaphor distinguish between two kinds of metaphorical language: *novel* (or *poetic*) metaphors (i.e., those that are imaginative), and *conventionalized* metaphors

(i.e., those that are used as a part of an ordinary discourse). According to Nunberg (1987), all metaphors emerge as novel, but over time they become part of general usage and their rhetorical effect vanishes, resulting in conventionalized metaphors. Following Orwell (1946), Nunberg calls such metaphors "dead" and claims that they are not psychologically distinct from literally used terms. The scheme described by Nunberg demonstrates how metaphorical associations capture patterns governing polysemy, namely, the capacity of a word to have multiple meanings. Over time some of the aspects of the target domain are added to the meaning of a term in the source domain, resulting in a (metaphorical) sense extension of this term. Copestake and Briscoe (1995) discuss sense extension mainly based on metonymic examples and model the phenomenon using lexical rules encoding metonymic patterns. They also suggest that similar mechanisms can be used to account for metaphorical processes. According to Copestake and Briscoe, the conceptual mappings encoded in the sense extension rules would define the limits to the possible shifts in meaning.

General-domain lexical resources often include information about metaphorical word senses, although unsystematically and without any accompanying semantic annotation. For example, WordNet[2] (Fellbaum 1998) contains the *comprehension* sense of *grasp*, defined as "get the meaning of something," and the *reading* sense of *skim*, defined as "read superficially." A great deal of metaphorical senses are absent from the current version of WordNet, however. A number of researchers have advocated the necessity of systematic inclusion and mark-up of metaphorical senses in such general-domain lexical resources (Alonge and Castelli 2003; Lönneker and Eilts 2004) and claim that this would be beneficial for the computational modeling of metaphor. Metaphor processing systems could then either use this knowledge or be evaluated against it. Lönneker (2004) mapped the senses from EuroWordNet[3] to the Hamburg Metaphor Database (Lönneker 2004; Reining and Lönneker-Rodman 2007) containing examples of metaphorical expressions in German and French. Currently no explicit information about metaphor is integrated into WordNet for English, however.

Although consistent inclusion in WordNet is in principle possible for conventional metaphorical senses, it is not viable for novel contextual sense alternations. Because metaphor is a productive phenomenon, all possible cases of contextual meaning alternations it results in cannot be described via simple sense enumeration (Pustejovsky 1995). Computational metaphor processing therefore cannot be approached using the standard word sense disambiguation paradigm, whereby the contextual use of a word is classified according to an existing sense inventory. The metaphor interpretation task is inherently more complex and requires generation of new and often uncommon meanings of the metaphorical term based on the context.

2.2 Theoretical Views on Metaphor

The following views on metaphor are prominent in linguistics and philosophy: the comparison view (e.g., the Structure-Mapping Theory of Gentner [1983]), the interaction view (Black 1962; Hesse 1966), the selectional restrictions violation view (Wilks 1975, 1978), and conceptual metaphor theory (CMT) (Lakoff and Johnson 1980). All of these

2 http://wordnet.princeton.edu/.

3 EuroWordNet is a multilingual database containing WordNets for several European languages (Dutch, Italian, Spanish, German, French, Czech, and Estonian). The WordNets are structured in the same way as the Princeton WordNet for English. URL: http://www.illc.uva.nl/EuroWordNet/.

approaches share the idea of an interconceptual mapping that underlies the production of metaphorical expressions. Gentner's Structure-Mapping Theory postulates that the ground for metaphor lies in similar properties and relations shared by the two concepts (the target and the source). Tourangeau and Sternberg (1982), however, criticize this view by noting that "everything has some feature or category that it shares with everything else, but we cannot combine just any two things in metaphor" (Tourangeau and Sternberg 1982, page 226). The interaction view focuses on the surprise and novelty that metaphor introduces. Its proponents claim that the source concept (or domain) represents a template for seeing the target concept in an entirely new way. The conceptual metaphor theory of Lakoff and Johnson (1980) takes this idea much further by stating that metaphor operates at the level of thought rather than at the level of language, and that it is based on a set of cognitive mappings between source and target domains. Thus Lakoff and Johnson put the emphasis on the structural aspect of metaphor, rather than its decorative function in language that dominated the preceding theories. The selectional restrictions violation view of Wilks (1978) concerns manifestation of metaphor in language. Wilks suggests that metaphor represents a violation of combinatory norms in the linguistic context and that metaphorical expressions can be detected via such violation.

2.2.1 Conceptual Metaphor Theory. Examples (6) and (7) provided a good illustration of CMT. Lakoff and Johnson explain them via the conceptual metaphor ARGUMENT IS WAR, which is systematically reflected in language in a variety of expressions.

(13) Your claims are *indefensible*. (Lakoff and Johnson 1980)

(14) I *demolished* his argument. (Lakoff and Johnson 1980)

(15) I've never *won* an argument with him. (Lakoff and Johnson 1980)

(16) You disagree? Okay, *shoot*! (Lakoff and Johnson 1980)

According to CMT, we conceptualize and structure arguments in terms of battle, which systematically influences the way we talk about arguments within our culture. In other words, the conceptual structure behind battle (i.e., that one can shoot, demolish, devise a strategy, win, and so on), is metaphorically transferred onto the domain of argument.

Manifestations of conceptual metaphor are ubiquitous in language and communication. Here are a few other examples of common metaphorical mappings.

- TIME IS MONEY (e.g., "That flat tire *cost* me an hour")
- IDEAS ARE PHYSICAL OBJECTS (e.g., "I cannot *grasp* his way of thinking")
- LINGUISTIC EXPRESSIONS ARE CONTAINERS (e.g., "I would not be able to *put* all my feelings *into* words")
- EMOTIONS ARE VEHICLES (e.g., "[...] she was *transported* with pleasure")
- FEELINGS ARE LIQUIDS (e.g., "[...] all of this *stirred* an unfathomable excitement in her")

- LIFE IS A JOURNEY (e.g., "He *arrived* at the end of his life with very little emotional *baggage*")

CMT produced a significant resonance in the fields of philosophy, linguistics, cognitive science, and artificial intelligence, including NLP. It inspired novel research (Martin 1990, 1994; Narayanan 1997, 1999; Barnden and Lee 2002; Feldman and Narayanan 2004; Mason 2004; Martin 2006; Agerri et al. 2007), but was also criticized for the lack of consistency and empirical verification (Murphy 1996; Shalizi 2003; Pinker 2007). The sole evidence with which Lakoff and Johnson (1980) supported their theory was a set of carefully selected examples. Such examples, albeit clearly illustrating the main tenets of the theory, are not representative. They cannot possibly capture the whole spectrum of metaphorical expressions, and thus do not provide evidence that the theory can adequately explain the majority of metaphors in real-world texts. Aiming to verify the latter, Shutova and Teufel (2010) conducted a corpus-based analysis of conceptual metaphor in the data from the British National Corpus (BNC) (Burnard 2007). In their study three independent participants annotated both linguistic metaphors and the underlying source–target domain mappings. Their results show that although the annotators reach some overall agreement on the annotation of interconceptual mappings, they experienced a number of difficulties, one of which was the problem of finding the right level of abstraction for the source and target domain categories. The difficulties in category assignment for conceptual metaphor suggest that it is hard to consistently assign explicit labels to source and target domains, even though the interconceptual associations exist in some sense and are intuitive to humans.

2.2.2 Selectional Restrictions Violation View. Lakoff and Johnson do not discuss how metaphors can be recognized in linguistic data. To date, the most influential account of this issue is that of Wilks (1975, 1978). According to Wilks, metaphors represent a violation of **selectional restrictions** (or *preferences*) in a given context. Selectional restrictions are the semantic constraints that a predicate places onto its arguments. Consider the following example.

(17) a. My aunt always *drinks* her tea on the terrace.

b. My car *drinks* gasoline. (Wilks 1978)

The verb *drink* normally requires a grammatical subject of type ANIMATE and a grammatical object of type LIQUID, as in Example (17a). Therefore, *drink* taking a *car* as a subject in (17b) is an anomaly, which, according to Wilks, indicates a metaphorical use of *drink*.

Although Wilks's idea inspired a number of computational experiments on metaphor recognition (Fass and Wilks 1983; Fass 1991; Krishnakumaran and Zhu 2007), it is important to note that in practice this approach has a number of limitations. Firstly, there are other kinds of non-literalness or anomaly in language that cause a violation of semantic norm, such as metonymies. Thus the method would overgenerate. Secondly, there are kinds of metaphor that do not represent a violation of selectional restrictions (i.e., the approach may also undergenerate). This would happen, for example, when highly conventionalized metaphorical word senses are more frequent than the original literal senses. Due to their frequency, selectional preference distributions of such words in real-world data would be skewed towards the metaphorical senses (e.g., *capture* may select for *ideas* rather than *captives* according to the data). As a result, no selectional preferences violation can be detected in the use of such verbs. Another case where the

method does not apply is copula constructions, such as "All the world's a *stage*." And finally, the method does not take into account the fact that interpretation (of metaphor as well as other linguistic phenomena) is always context-dependent. For example, the phrase "All men are *animals*" uttered by a biology professor or a feminist would have entirely different interpretations, the latter clearly metaphorical, but without any violation of selectional restrictions.

2.3 Computational Approaches to Metaphor

2.3.1 Automatic Metaphor Recognition. One of the first attempts to automatically identify and interpret metaphorical expressions in text is the approach of Fass (1991). It originates in the idea of Wilks (1978) and utilizes hand-coded knowledge. Fass developed a system called `met*`, which is capable of discriminating between literalness, metonymy, metaphor, and anomaly. It does this in three stages. First, literalness is distinguished from non-literalness using selectional preference violation as an indicator. In the case that non-literalness is detected, the respective phrase is tested for being metonymic using hand-coded patterns (such as CONTAINER-FOR-CONTENT). If the system fails to recognize metonymy, it proceeds to search the knowledge base for a relevant analogy in order to discriminate metaphorical relations from anomalous ones. For example, the sentence in Example (17b) would be represented in this framework as (*car,drink,gasoline*), which does not satisfy the preference (*animal,drink,liquid*), as *car* is not a hyponym of *animal*. `met*` then searches its knowledge base for a triple containing a hypernym of both the actual argument and the desired argument and finds (*thing,use,energy_source*), which represents the metaphorical interpretation.

Goatly (1997) identifies a set of linguistic cues, namely, lexical patterns indicating the presence of a metaphorical expression in running text, such as *metaphorically speaking, utterly, completely, so to speak*, and *literally*. This approach, however, is likely to find only a small proportion of metaphorical expressions, as the vast majority of them appear without any signaling context. We conducted a corpus study in order to investigate the effectiveness of linguistic cues as metaphor indicators. For each cue suggested by Goatly (1997), we randomly sampled 50 sentences from the BNC containing it and manually annotated them for metaphoricity. The results are presented in Table 1. The average precision (i.e., the proportion of identified expressions that were metaphorical) of the linguistic cue method according to these data is 0.40, which suggests that the set of metaphors that this method generates contains a great deal of noise. Thus the cues are unlikely to be sufficient for metaphor extraction on their own, but together with some additional filters, they could contribute to a more complex system.

The work of Peters and Peters (2000) concentrates on detecting figurative language in lexical resources. They mine WordNet (Fellbaum 1998) for examples of systematic

Table 1
Corpus statistics for linguistic cues.

Cue	BNC frequency	Sample size	Metaphors	Precision
"metaphorically speaking"	7	7	5	0.71
"literally"	1,936	50	13	0.26
"figurative"	125	50	9	0.18
"utterly"	1,251	50	16	0.32
"completely"	8,339	50	13	0.26
"so to speak"	353	49	35	0.71

polysemy, which allows them to capture metonymic and metaphorical relations. Their system searches for nodes that are relatively high in the WordNet hierarchy (i.e., are relatively general) and that share a set of common word forms among their descendants. Peters and Peters found that such nodes often happen to be in a metonymic (e.g., *publisher – publication*) or a metaphorical (e.g., *theory – supporting structure*) relation.

The CorMet system (Mason 2004) is the first attempt at discovering source–target domain mappings automatically. It does this by finding systematic variations in domain-specific selectional preferences, which are inferred from texts on the Web. For example, Mason collects texts from the LAB domain and the FINANCE domain, in both of which *pour* would be a characteristic verb. In the LAB domain *pour* has a strong selectional preference for objects of type *liquid*, whereas in the FINANCE domain it selects for *money*. From this Mason's system infers the domain mapping FINANCE – LAB and the concept mapping MONEY IS LIQUID. He compares the output of his system against the Master Metaphor List (MML; Lakoff, Espenson, and Schwartz 1991) and reports a performance of 77% in terms of accuracy (i.e., proportion of correctly induced mappings).

Birke and Sarkar (2006) present a sentence clustering approach for non-literal language recognition, implemented in the TroFi system (Trope Finder). The idea behind their system originates from a similarity-based word sense disambiguation method developed by Karov and Edelman (1998). The latter uses a set of seed sentences annotated with respect to word sense. The system computes similarity between the sentence containing the word to be disambiguated and all of the seed sentences and selects the sense corresponding to the annotation in the most similar seed sentences. Birke and Sarkar adapt this algorithm to perform a two-way classification (literal vs. non-literal), not aiming to distinguish between specific kinds of tropes. An example for the verb *pour* in their database is shown in Figure 2. They attain a performance of 0.54 in terms of F-measure (van Rijsbergen 1979).

The method of Gedigian et al. (2006) discriminates between literal and metaphorical use. The authors trained a maximum entropy classifier for this purpose. They collected their data using FrameNet (Fillmore, Johnson, and Petruck 2003) and PropBank (Kingsbury and Palmer 2002) annotations. FrameNet is a lexical resource for English containing information on words' semantic and syntactic combinatory possibilities, or valencies, in each of their senses. PropBank is a corpus annotated with verbal propositions and their arguments. Gedigian et al. (2006) extracted the lexical items whose frames are related to MOTION and CURE from FrameNet, then searched the PropBank Wall Street Journal corpus (Kingsbury and Palmer 2002) for sentences containing such lexical items and annotated them with respect to metaphoricity. For example, the verb *run* in the sentence "Texas Air has *run* into difficulty" was annotated as metaphorical, and in "I was doing the laundry and nearly broke my neck *running* upstairs to see" as literal. Gedigian et al. used PropBank annotation (arguments and their semantic

pour
nonliteral cluster
wsj04:7878 N As manufacturers get bigger, they are likely to pour more money into the battle for shelf space, raising the ante for new players.
wsj25:3283 N Salsa and rap music pour out of the windows.
wsj06:300 U Investors hungering for safety and high yields are pouring record sums into single-premium, interest-earning annuities.
literal cluster
wsj59:3286 L Custom demands that cognac be poured from a freshly opened bottle.

Figure 2
An example of the data of Birke and Sarkar (2006).

types) as features to train the classifier, and report an accuracy of 95.12%. This result is, however, only 2.22 percentage points higher than the performance of the naive baseline assigning majority class to all instances (92.90%). Such high performance of their system can be explained by the fact that 92.90% of the verbs of MOTION and CURE in their data are used metaphorically, thus making the data set unbalanced with respect to target categories and making the task easier.

Both Birke and Sarkar (2006) and Gedigian et al. (2006) focus only on metaphors expressed by a verb. The approach of Krishnakumaran and Zhu (2007) additionally covers metaphors expressed by nouns and adjectives. Krishnakumaran and Zhu use hyponymy relation in WordNet and word bigram counts to predict metaphors at a sentence level. Given a metaphor in copula constructions, or an IS-A metaphor (e.g., the famous quote by William Shakespeare "All the world's a *stage*") they verify if the two nouns involved are in hyponymy relation in WordNet, otherwise this sentence is tagged as containing a metaphor. They also treat expressions containing a verb or an adjective used metaphorically (e.g., "He *planted* good ideas in their minds" or "He has a *fertile* imagination"). For those cases, they calculate bigram probabilities of verb–noun and adjective–noun pairs (including the hyponyms/hypernyms of the noun in question). If the combination is not observed in the data with sufficient frequency, the system tags the sentence as metaphorical. This idea is a modification of the selectional preference view of Wilks, although applied at the bigram level. Alternatively, one could extract verb–object relations from parsed text. Compared to the latter, Krishnakumaran and Zhu (2007) lose a great deal of information. The authors evaluated their system on a set of example sentences compiled from the Master Metaphor List, whereby highly conventionalized metaphors are taken to be negative examples. Thus they do not deal with literal examples as such. Essentially, the distinction Krishnakumaran and Zhu are making is between the senses included in WordNet, even if they are conventional metaphors (e.g., "*capture* an idea"), and those not included in WordNet (e.g., "*planted* good ideas").

2.3.2 Automatic Metaphor Interpretation. One of the first computational accounts of metaphor interpretation is that of Martin (1990). In his metaphor interpretation, denotation and acquisition system (MIDAS), Martin models the hierarchical organization of conventional metaphors. The main assumption underlying this approach is that more specific conventional metaphors (e.g., COMPUTATIONAL PROCESS viewed as a LIVING BEING in "How can I *kill* a process?") descend from more general ones (e.g., PROCESS [general, as a sequence of events] is a LIVING BEING). Given an example of a metaphorical expression, MIDAS searches its database for a corresponding conceptual metaphor that would explain the anomaly. If it does not find any, it abstracts from the example to more general concepts and repeats the search. If a suitable general metaphor is found, it creates a new mapping for its descendant, a more specific metaphor, based on this example. This is also how novel conceptual metaphors are acquired by the system. The metaphors are then organized into a resource called MetaBank (Martin 1994). The knowledge is represented in MetaBank in the form of **metaphor maps** (Martin 1988) containing detailed information about source-target concept mappings and empirically derived examples. MIDAS has been integrated with Unix Consultant, a system that answers users' questions about Unix. The system first tries to find a literal answer to the question. If it is not able to, it calls MIDAS, which detects metaphorical expressions via selectional preference violation and searches its database for a metaphor explaining the anomaly in the question.

Another cohort of approaches aims to perform inference about entities and events in the source and target domains for the purpose of metaphor interpretation. These

include the KARMA system (Narayanan 1997, 1999; Feldman and Narayanan 2004) and the ATT-Meta project (Barnden and Lee 2002; Agerri et al. 2007). Within both systems the authors developed a metaphor-based reasoning framework in accordance with CMT. The reasoning process relies on manually coded knowledge about the world and operates mainly in the source domain. The results are then projected onto the target domain using the conceptual mapping representation. The ATT-Meta project concerns metaphorical and metonymic description of mental states; and reasoning about mental states is performed using first order logic. Their system, however, does not take natural language sentences as input, but hand-coded logical expressions that are representations of small discourse fragments. KARMA in turn deals with a broad range of abstract actions and events and takes parsed text as input.

Veale and Hao (2008) derive a "fluid knowledge representation for metaphor interpretation and generation" called Talking Points. Talking Points is a set of characteristics of concepts belonging to source and target domains and related facts about the world which are acquired automatically from WordNet and from the Web. Talking Points are then organized in *Slipnet*, a framework that allows for a number of insertions, deletions, and substitutions in definitions of such characteristics in order to establish a connection between the target and the source concepts. This work builds on the idea of *slippage* in knowledge representation for understanding analogies in abstract domains (Hofstadter and Mitchell 1994; Hofstadter 1995). The following is an example demonstrating how slippage operates to explain the metaphor *Make-up is a Western burqa*.

Make-up =>
$\equiv$ typically worn by women
$\approx$ expected to be worn by women
$\approx$ must be worn by women
$\approx$ must be worn by Muslim women
Burqa <=

By doing insertions and substitutions, the system arrives from the definition "typically worn by women" to that of "must be worn by Muslim women." Thus it establishes a link between the concepts of *make-up* and *burqa*. Veale and Hao, however, did not evaluate to what extent their system is able to interpret metaphorical expressions in real-world text.

The next sections of the paper are devoted to our own experiments on metaphor identification and interpretation.

3. Metaphor Identification Method and Experiments

The first task for metaphor processing within NLP is its identification in text. As discussed earlier, previous approaches to this problem either utilize hand-coded knowledge (Fass 1991; Krishnakumaran and Zhu 2007) or reduce the task to searching for metaphors of a specific domain defined a priori (e.g., MOTION metaphors) in a specific type of discourse (e.g., the *Wall Street Journal* [Gedigian et al. 2006]). In contrast, the search space in our experiments is the entire BNC and the domain of the expressions identified is unrestricted. In addition, the developed technique does not rely on any hand-crafted lexical or world knowledge, but rather captures metaphoricity by means of verb and noun clustering in a data-driven manner.

The motivation behind the use of clustering methods for the metaphor identification task lies in CMT. The patterns of conceptual metaphor (e.g., FEELINGS ARE LIQUIDS)

always operate on semantic classes, that is, groups of related concepts, defined by Lakoff and Johnson as conceptual domains (FEELINGS include *love, anger, hatred,* etc.; LIQUIDS include *water, tea, petrol, beer,* etc.). Thus modeling metaphorical mechanisms in accordance with CMT would involve capturing such semantic classes automatically. Previous research on corpus-based lexical semantics has shown that it is possible to automatically induce semantic word classes from corpus data via clustering of contextual cues (Pereira, Tishby, and Lee 1993; Lin 1998; Schulte im Walde 2006). The current consensus is that the lexical items showing similar behavior in a large body of text most likely have related meanings.

The second reason for the use of unsupervised and weakly supervised methods is suggested by the results of corpus-based studies of conceptual metaphor. The analysis of conceptual mappings in unrestricted text, conducted by Shutova and Teufel (2010), although confirming some aspects of CMT, uncovered a number of fundamental difficulties. One of these is the choice of the level of abstraction and granularity of categories (i.e., labels for source and target domains). This suggests that it is hard to define a comprehensive inventory of labels for source and target domains. Thus a computational model of metaphorical associations should not rely on explicit domain labels. Unsupervised methods allow us to recover patterns in data without assigning any explicit labels to concepts, and thus to model interconceptual mappings implicitly.

The method behind our metaphor identification system relies on distributional clustering. Noun clustering, specifically, is central to the approach. It is traditionally assumed that noun clusters produced using distributional clustering contain concepts that are similar to each other. This is true only in part, however. There exist two types of concepts: *concrete*, those concepts denoting physical entities or physical experiences (e.g., *chair, apple, house, rain*) and *abstract*, those concepts that do not physically exist at any particular time or place, but rather exist as a type of thing or as an idea (e.g., *justice, love, democracy*). It is the abstract concepts that tend to be described metaphorically, rather than concrete concepts. Humans use metaphor attempting to gain a better understanding of an abstract concept by comparing it to their physical experiences. As a result, abstract concepts expose different distributional behavior in a corpus. This in turn affects the application of clustering techniques and the obtained clusters for concrete and abstract concepts would be structured differently. Consider the example in Figure 3. The figure shows a cluster containing concrete concepts (on the right) that are various kinds of mechanisms; a cluster containing verbs co-occurring with mechanisms in the corpus (at the bottom); and a cluster containing abstract concepts (on the left) that tend to co-occur with these verbs. Such abstract concepts, albeit having quite distinct meanings (e.g., *marriage* and *democracy*), are observed in similar lexico-syntactic environments. This is due to the fact that they are systematically used metaphorically with the verbs from the domain of MECHANISM. Hence, they are automatically assigned to the same cluster. The following examples illustrate this phenomenon in textual data.

(18) Our relationship is not really *working*.

(19) Diana and Charles did not succeed in *mending* their marriage.

(20) The *wheels* of Stalin's regime were *well oiled* and already *turning*.

Such a structure of the abstract clusters can be explained by the fact that *relationships*, *marriages*, *collaborations*, and *political systems* are all cognitively mapped to the same

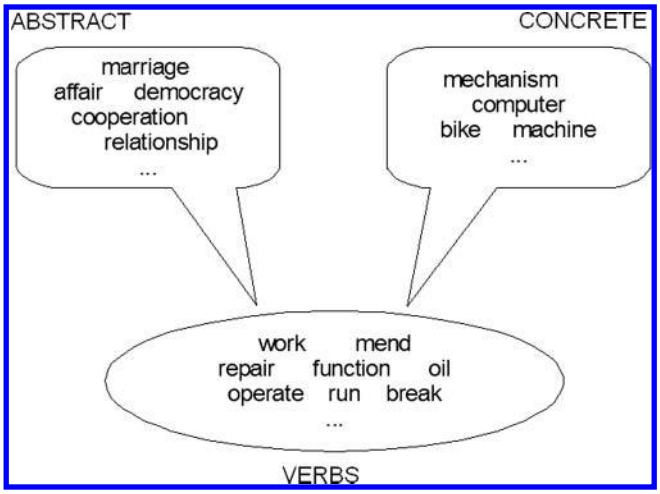

Figure 3
Cluster of target concepts associated with MECHANISM.

source domain of MECHANISM. In contrast to concrete concepts, such as *tea, water, coffee, beer, drink, liquid*, that are clustered together when they have similar meanings, abstract concepts tend to be clustered together if they are associated with the same source domain. We define this phenomenon as **clustering by association** and it becomes central to the system design. The expectation is that clustering by association would allow the harvesting of new target domains that are associated with the same source domain, and thus identify new metaphors.

The metaphor identification system starts from a small set of *seed* metaphorical expressions, that is, annotated metaphors (such as those in Examples (18) or (19)), which serve as training data. Note that seed annotation only concerns linguistic metaphors; metaphorical mappings are not annotated. The system then (1) creates source domains describing these examples by means of verb clustering (such as the verb cluster in Figure 3); (2) identifies new target domains associated with the same source domain by means of noun clustering (see, e.g., ABSTRACT cluster in Figure 3), and (3) establishes a link between the source and the target clusters based on the seed examples.

Thus the system captures metaphorical associations implicitly. It generalizes over the associated domains by means of verb and noun clustering. The obtained clusters then represent source and target concepts between which metaphorical associations hold. The knowledge of such associations is then used to identify new metaphorical expressions in a large corpus.

In addition to this, we build a selectional preference–based metaphor filter. This idea stems from the view of Wilks (1978), but is, however, a modification of it. The filter assumes that the verbs exhibiting weak selectional preferences, namely, verbs co-occurring with any argument class in linguistic data (*remember, influence,* etc.) generally have no or only weak potential for being a metaphor. It has been previously shown that it is possible to quantify verb selectional preferences on the basis of corpus data, using, for example, a measure defined by Resnik (1993). Once the candidate metaphors are identified in the corpus using clustering methods, those displaying weak selectional preferences can be filtered out.

Figures 4 and 5 depict the metaphor identification pipeline: first, the identification of metaphorical associations and then that of metaphorical expressions in text. In

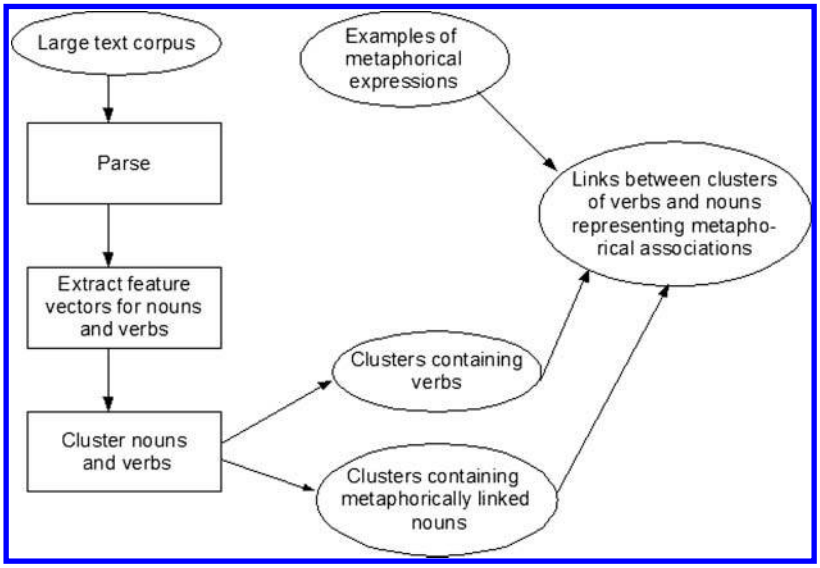

Figure 4
Learning metaphorical associations by means of verb and noun clustering and using the seed set.

summary, the system (1) starts from a seed set of metaphorical expressions exemplifying a range of source–target domain mappings; (2) performs noun clustering in order to harvest various target concepts associated with the same source domain; (3) creates a source domain verb lexicon by means of verb clustering; (4) searches the corpus for metaphorical expressions describing the target domain concepts using the verbs from the source domain lexicon; and (5) filters out the candidates exposing weak selectional preference strength as non-metaphorical.

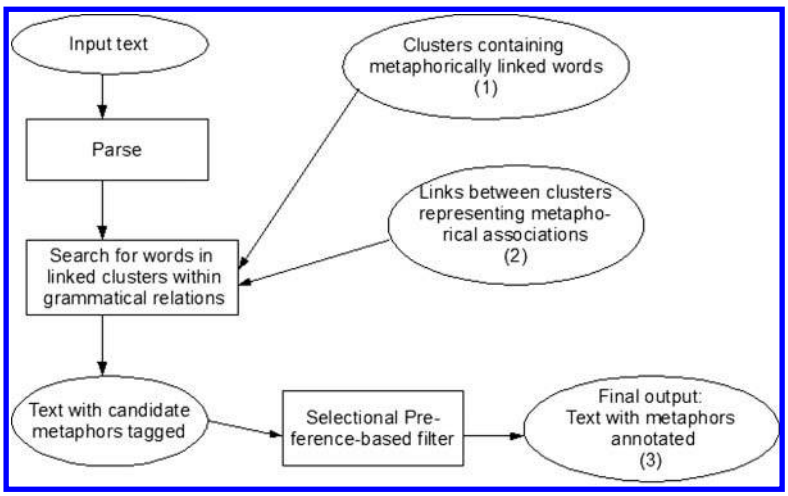

Figure 5
Identification of new metaphorical expressions in text.

3.1 Experimental Data

The identification system takes a list of seed phrases as input. Seed phrases contain manually annotated linguistic metaphors. The system generalizes from these linguistic metaphors to the respective conceptual metaphors by means of clustering. This generalization is then used to harvest a large number of new metaphorical expressions in unseen text. Thus the data needed for the identification experiment consist of a seed set, data sets of verbs and nouns that are subsequently clustered, and an evaluation corpus.

3.1.1 Metaphor Corpus and Seed Phrases. The data to test the identification module were extracted from the metaphor corpus created by Shutova and Teufel (2010). Their corpus is a subset of the BNC (Burnard 2007) and, as such, it provides a suitable platform for testing the metaphor processing system on real-world general-domain expressions in contemporary English. Our data set consists of verb–subject and verb–direct object metaphorical expressions. In order to avoid extra noise, we enforced some additional selection criteria. All phrases were included unless they fell in one of the following categories:

- Phrases where the subject or object referent is unknown (e.g., containing pronouns such as in "in which they [changes] *operated*") or represented by a named entity (e.g., "Then Hillary *leapt* into the conversation"). These cases were excluded from the data set because their processing would involve the use of additional modules for coreference resolution and named entity recognition, which in turn may introduce additional errors into the system.
- Phrases whose metaphorical meaning is realized solely in passive constructions (e.g., "sociologists have been *inclined* to [..]"). These cases were excluded because for many such examples it was hard for humans to produce a literal paraphrase realized in the form of the same syntactic construction. Thus their paraphrasing was deemed to be an unfairly hard task for the system.
- Multiword metaphors (e.g., "whether we *go on pilgrimage* with Raleigh or *put out to sea* with Tennyson"). The current system is designed to identify and paraphrase single-word, lexical metaphors. In the future the system needs to be modified to process multiword metaphorical expressions; this is, however, outside the scope of the current experiments.

The resulting data set consists of 62 phrases that are different single-word metaphors representing verb–subject and verb–direct object relations, where a verb is used metaphorically. The phrases include, for instance, "*stir* excitement," "*reflect* enthusiasm," "*grasp* theory," "*cast* doubt," "*suppress* memory," "*throw* remark" (verb–direct object constructions); and "campaign *surged*," "factor *shaped* [...]," "tension *mounted*," "ideology *embraces*," "example *illustrates*" (subject–verb constructions). This data set was used as a seed set in the identification experiments. The phrases in the data set were manually annotated for grammatical relations.

3.1.2 Verb and Noun Data Sets. The noun data set used for clustering consists of the 2,000 most frequent nouns in the BNC. The 2,000 most frequent nouns cover most common target categories and their linguistic realizations. BNC represents a suitable

source for such nouns because the corpus is balanced with respect to genre, style, and theme.

The verb data set is a subset of VerbNet (Kipper et al. 2006). VerbNet is the largest resource for general-domain verbs organized into semantic classes as proposed by Levin (1993). The data set includes all the verbs in VerbNet with the exception of highly infrequent ones. The frequency of the verbs was estimated from the data collected by Korhonen, Krymolowski, and Briscoe (2006) for the construction of the VALEX lexicon, which to date is one of the largest automatically created verb resources. The verbs from VerbNet that appear less than 150 times in this data were excluded. The resulting data set consists of 1,610 general-domain verbs.

3.1.3 Evaluation Corpus. The evaluation data for metaphor identification was the BNC parsed by the RASP parser (Briscoe, Carroll, and Watson 2006). We used the grammatical relation (GR) output of RASP for the BNC created by Andersen et al. (2008). The system searched the corpus for the source and target domain vocabulary within a particular grammatical relation (verb–direct object or verb–subject).

3.2 Method

The main components of the method include (1) distributional clustering of verbs and nouns, (2) search through the parsed corpus, and (3) selectional preference-based filtering.

3.2.1 Verb and Noun Clustering Method. The metaphor identification system relies on the clustering method of Sun and Korhonen (2009). They use a rich set of syntactic and semantic features (GRs, verb subcategorization frames [SCFs], and selectional preferences) and spectral clustering, a method particularly suitable for the resulting high dimensional feature space. This algorithm has proved to be effective in previous verb clustering experiments (Brew and Schulte im Walde 2002) and in other NLP tasks involving high dimensional data (Chen et al. 2006).

Spectral clustering partitions objects relying on their similarity matrix. Given a set of data points, the similarity matrix records similarities between all pairs of points. The system of Sun and Korhonen (2009) constructs similarity matrices using the **Jensen-Shannon divergence** as a measure. Jensen-Shannon divergence between two feature vectors w_i and w_j is defined as follows:

$$JSD(w_i, w_j) = \frac{1}{2}D(w_i||m) + \frac{1}{2}D(w_j||m) \quad (1)$$

where D is the Kullback-Leibler divergence, and m is the average of the w_i and w_j.

Spectral clustering can be viewed in abstract terms as the partitioning of a graph G over a set of words W. The weights on the edges of G are the similarities S_{ij}. The similarity matrix S thus represents the adjacency matrix for G. The clustering problem is then defined as identifying the optimal partition, or *cut*, of the graph into clusters, such that the intra-cluster weights are high and the inter-cluster weights are low. The system of Sun and Korhonen (2009) uses the MNCut algorithm of Meila and Shi (2001) for this purpose.

Sun and Korhonen (2009) evaluated their clustering approach on 204 verbs from 17 Levin classes and obtained an F-measure of 80.4, which is the state-of-the-art

performance level. The metaphor identification system uses the method of Sun and Korhonen to cluster both verbs and nouns (separately), however, significantly extending its coverage to unrestricted general-domain data and applying the method to a considerably larger data set of 1,610 verbs.

3.2.2 Feature Extraction and Clustering Experiments. For verb clustering, the best performing features from Sun and Korhonen (2009) were adopted. These include automatically acquired verb SCFs parameterized by their selectional preferences. These features were obtained using the SCF acquisition system of Preiss, Briscoe, and Korhonen (2007). The system tags and parses corpus data using the RASP parser (Briscoe, Carroll, and Watson 2006) and extracts SCFs from the produced grammatical relations using a rule-based classifier which identifies 168 SCF types for English verbs. It produces a lexical entry for each verb and SCF combination occurring in corpus data. The selectional preference classes were obtained by clustering nominal arguments appearing in the subject and object slots of verbs in the resulting lexicon.

Following previous works on semantic noun classification (Pantel and Lin 2002; Bergsma, Lin, and Goebel 2008), grammatical relations were used as features for noun clustering. More specifically, the frequencies of nouns and verb lemmas appearing in the subject, direct object, and indirect object relations in the RASP-parsed BNC were included in the feature vectors. For example, the feature vector for *banana* would contain the following entries: {`eat-dobj` n_1, `fry-dobj` n_2, `sell-dobj` n_3,..., `eat_with-iobj` n_i, `look_at-iobj` n_{i+1},..., `rot-subj` n_k, `grow-subj` n_{k+1},...}.

We experimented with different clustering granularities, subjectively examined the obtained clusters, and determined that the number of clusters set to 200 is the most suitable setting for both nouns and verbs in our task. This was done by means of qualitative analysis of the clusters as representations of source and target domains—that is, by judging how complete and homogeneous the verb clusters were as lists of potential source domain vocabulary and how many new target domains associated with the same source domain were found correctly in the noun clusters. This analysis was performed on a randomly selected set of 10 clusters taken from different granularity settings and none of the seed expressions were used for it. Examples of such clusters are shown in Figures 6 (nouns) and 7 (verbs), respectively. The noun clusters represent target concepts associated with the same source concept (some suggested source concepts are given in Figure 6, although the system only captures those implicitly). The verb clusters contain lists of source domain vocabulary.

3.2.3 Corpus Search. Once the clusters have been obtained, the system proceeds to search the corpus for source and target domain terms within verb–object (both direct and indirect) and verb–subject relations. For each seed expression, a cluster is retrieved for the verb to form the source concept, and a cluster is retrieved for the noun to form a list of target concepts. The retrieved verb and noun clusters are then linked, and such links represent metaphorical associations. The system then classifies grammatical relations in the corpus as metaphorical if the lexical items in the grammatical relation appear in the linked source (verb) and target (noun) clusters. This search is performed on the BNC parsed by RASP. Consider the following example sentence extracted from the BNC (the BNC text ID is given in brackets, followed by the hypothetical conceptual metaphor):

(21) Few would deny that in the nineteenth century change was greatly *accelerated*. (ACA) – CHANGE IS MOTION

Source: MECHANISM
Target Cluster: consensus relation tradition partnership resistance foundation alliance friendship contact reserve unity link peace bond myth identity hierarchy relationship connection balance marriage democracy defense faith empire distinction coalition regime division
Source: PHYSICAL OBJECT; LIVING BEING; STRUCTURE
Target Cluster: view conception theory concept ideal belief doctrine logic hypothesis interpretation proposition thesis assumption idea argument ideology conclusion principle notion philosophy
Source: STORY; JOURNEY
Target Cluster: politics practice trading reading occupation profession sport pursuit affair career thinking life
Source: LIQUID
Target Cluster: disappointment rage concern desire hostility excitement anxiety passion doubt panic delight anger fear curiosity shock terror surprise pride happiness pain enthusiasm alarm hope memory love satisfaction sympathy spirit frustration impulse instinct warmth beauty ambition thought guilt emotion sensation horror feeling laughter suspicion pleasure
Source: LIVING BEING; END
Target Cluster: defeat fall death tragedy loss collapse decline disaster destruction fate

Figure 6
Clustered nouns (the associated source domain labels are suggested by the authors for clarity; the system does not assign any labels, but models source and target domains implicitly).

Source Cluster: sparkle glow widen flash flare gleam darken narrow flicker shine blaze bulge
Source Cluster: gulp drain stir empty pour sip spill swallow drink pollute seep flow drip purify ooze pump bubble splash ripple simmer boil tread
Source Cluster: polish clean scrape scrub soak
Source Cluster: kick hurl push fling throw pull drag haul
Source Cluster: rise fall shrink drop double fluctuate dwindle decline plunge decrease soar tumble surge spiral boom
Source Cluster: initiate inhibit aid halt trace track speed obstruct impede accelerate slow stimulate hinder block
Source Cluster: work escape fight head ride fly arrive travel come run go slip move

Figure 7
Clustered verbs.

The relevant GRs identified by the parser are presented in Figure 8. The relation between the verb *accelerate* and its semantic object *change* in Example (21) is expressed in the passive voice and is, therefore, tagged by RASP as an `ncsubj` GR. Because this GR contains terminology from associated source (MOTION) and target (CHANGE) domains, it is marked as metaphorical and so is the term *accelerate*, which belongs to the source domain of MOTION.

3.2.4 Selectional Preference Strength Filter. In the previous step a set of candidate verb metaphors and the associated grammatical relations were extracted from the BNC. These now need to be filtered based on selectional preference strength. To do this, we

(21) *Change was greatly accelerated* – CHANGE IS MOTION
ncsubj head=accelerate+ed_VVN_25 dep=change_NN1_22
aux head=accelerate+ed_VVN_25 dep=be+ed_VBDZ_23
ncmod head=accelerate+ed_VVN_25 dep=greatly_RR_24
conj dep=accelerate+ed_VVN_25
passive head=accelerate+ed_VVN_25

Figure 8
Grammatical relations output for metaphorical expressions.

automatically acquire selectional preference distributions for verb–subject and verb–direct object relations from the RASP-parsed BNC. The noun clusters obtained using Sun and Korhonen's method as described earlier form the selectional preference classes. To quantify selectional preferences, we adopt the selectional preference strength (SPS) measure of Resnik (1993). Resnik models selectional preferences of a verb in probabilistic terms as the difference between the posterior distribution of noun classes in a particular relation with the verb and their prior distribution in that syntactic position irrespective of the identity of the verb. He quantifies this difference using the Kullback-Leibler divergence and defines **selectional preference strength** as follows:

$$S_R(v) = D(P(c|v)||P(c)) = \sum_c P(c|v) \log \frac{P(c|v)}{P(c)} \tag{2}$$

where $P(c)$ is the prior probability of the noun class, $P(c|v)$ is the posterior probability of the noun class given the verb, and R is the grammatical relation in question. In order to quantify how well a particular argument class fits the verb, Resnik defines another measure called **selectional association**:

$$A_R(v,c) = \frac{1}{S_R(v)} P(c|v) \log \frac{P(c|v)}{P(c)} \tag{3}$$

which stands for the contribution of a particular argument class to the overall selectional preference strength of a verb.

The probabilities $P(c|v)$ and $P(c)$ were estimated from the corpus data as follows:

$$P(c|v) = \frac{f(v,c)}{\sum_k f(v,c_k)} \tag{4}$$

$$P(c) = \frac{f(c)}{\sum_k f(c_k)} \tag{5}$$

where $f(v,c)$ is the number of times the predicate v co-occurs with the argument class c in the relation R, and $f(c)$ is the number of times the argument class occurs in the relation R regardless of the identity of the predicate.

Thus for each verb, its SPS can be calculated for specific grammatical relations. This measure was used to filter out the verbs with weak selectional preferences. The expectation is that such verbs are unlikely to be used metaphorically. The optimal selectional preference strength threshold was set experimentally for both verb–subject and verb–object relations on a small held-out data set (via qualitative analysis of the data). It approximates to 1.32. The system excludes expressions containing the verbs with preference strength below this threshold from the set of candidate metaphors. Examples of verbs with weak and strong direct object SPs are shown in Tables 2 and 3, respectively. Given the SPS threshold of 1.32, the filter discards 31% of candidate expressions initially identified in the corpus.

3.3 Evaluation

In order to show that the described metaphor identification method generalizes well over the seed set and that it operates beyond synonymy, its output was compared to

Table 2
Verbs with weak direct object SPs.

SPS	Verb
1.3175	undo
1.3160	bud
1.3143	deplore
1.3138	seal
1.3131	slide
1.3126	omit
1.3118	reject
1.3097	augment
1.3094	frustrate
1.3087	restrict
1.3082	employ
1.3081	highlight
1.3081	correspond
1.3056	dab
1.3053	assist
1.3043	neglect
...	

Table 3
Verbs with strong direct object SPs.

SPS	Verb	SPS	Verb
...			
3.0810	aggravate	2.9434	coop
3.0692	dispose	2.9326	hobble
3.0536	rim	2.9285	paper
3.0504	deteriorate	2.9212	sip
3.0372	mourn	...	
3.0365	tread	1.7889	schedule
3.0348	cadge	1.7867	cheat
3.0254	intersperse	1.7860	update
3.0225	activate	1.7840	belt
3.0085	predominate	1.7835	roar
3.0033	lope	1.7824	intensify
2.9957	bone	1.7811	read
2.9955	pummel	1.7805	unnerve
2.9868	disapprove	1.7776	arrive
2.9838	hoover	1.7775	publish
2.9824	beam	1.7775	reason
2.9807	amble	1.7774	bond
2.9760	diversify	1.7770	issue
2.9759	mantle	1.7760	verify
2.9730	pulverize	1.7734	vomit
2.9604	skim	1.7728	impose
2.9539	slam	1.7726	phone
2.9523	archive	1.7723	purify
2.9504	grease	...	

that of a baseline using WordNet. In the baseline system, WordNet synsets represent source and target domains. The quality of metaphor identification for both the system and the baseline was evaluated in terms of precision with the aid of human judges. To compare the coverage of the system to that of the baseline in quantitative terms we assessed how broadly they expand on the seed set. To do this, we estimated the number of word senses captured by the two systems and the proportion of identified metaphors that are not synonymous with any of those seen in the seed set, according to WordNet. This type of evaluation assesses how well clustering methods are suited to identify new metaphors not directly related to those in the seed set.

3.3.1 Comparison with WordNet Baseline. The baseline system was implemented using synonymy information from WordNet to expand on the seed set. Source and target domain vocabularies were thus represented as sets of synonyms of verbs and nouns in seed expressions. The baseline system then searched the corpus for phrases composed of lexical items belonging to those vocabularies. For example, given a seed expression "*stir* excitement," the baseline finds phrases such as "*arouse* fervour, *stimulate* agitation, *stir* turmoil," and so forth. It is not able to generalize over the concepts to broad semantic classes, however—for example, it does not find other FEELINGS such as *rage, fear, anger, pleasure*. This, however, is necessary to fully characterize the target domain. Similarly, in the source domain, the system only has access to direct synonyms of *stir*, rather than to other verbs characteristic of the domain of LIQUIDS (*pour, flow, boil,* etc.).

To compare the coverage achieved by the system using clustering to that of the baseline in quantitative terms, we estimated the number of WordNet synsets, that is, different word senses, in the metaphorical expressions captured by the two systems. We found that the baseline system covers only 13% of the data identified using clustering. This is due to the fact that it does not reach beyond the concepts present in the seed set. In contrast, most metaphors tagged by the clustering method (87%) are non-synonymous to those in the seed set and some of them are novel. Together, these metaphors represent a considerably wider range of meanings. Given the seed metaphors "*stir* excitement, *throw* remark, *cast* doubt," the system identifies previously unseen expressions "*swallow* anger, *hurl* comment, *spark* enthusiasm," and so on, as metaphorical. Tables 4 and 5 show examples of how the system and the baseline expand on the seed set, respectively. Full sentences containing metaphors annotated by the system are shown in Figure 9. Twenty-one percent of the expressions identified by the system do not have their corresponding metaphorical senses included in WordNet, such as "*spark* enthusiasm"; the remaining 79% are, however, more common conventional metaphors. Starting with a seed set of only 62 examples, the system expands significantly on the seed set and identifies a total of 4,456 metaphorical expressions in the BNC. This suggests that the method has the potential to attain a broad coverage of the corpus given a large and representative seed set.

3.3.2 Evaluation Against Human Judgments. In order to assess the quality of metaphor identification by both systems, their output was assessed by human judgments. For this purpose, we randomly sampled sentences containing metaphorical expressions as annotated by the system and by the baseline and asked human annotators to decide whether these were metaphorical or not.

Participants Five volunteers participated in the experiment. They were all native speakers of English and had no formal training in linguistics.

Table 4
Examples of seed set expansion by the system.

Seed phrase	Harvested metaphors	BNC frequency
reflect concern (V-O):	reflect concern	78
	reflect interest	74
	reflect commitment	26
	reflect preference	22
	reflect wish	17
	reflect determination	12
	reflect intention	8
	reflect willingness	4
	reflect sympathy	3
	reflect loyalty	2
	disclose interest	10
	disclose intention	3
	disclose concern	2
	disclose sympathy	1
	disclose commitment	1
	disguise interest	6
	disguise intention	3
	disguise determination	2
	obscure interest	1
	obscure determination	1
cast doubt (V-O):	cast doubt	197
	cast fear	3
	cast suspicion	2
	catch feeling	3
	catch suspicion	2
	catch enthusiasm	1
	catch emotion	1
	spark fear	10
	spark enthusiasm	3
	spark passion	1
	spark feeling	1
campaign surged (S-V):	campaign surged	1
	charity boomed	1
	effort decreased	1
	expedition doubled	1
	effort doubled	1
	campaign shrank	1
	campaign soared	1
	drive spiraled	1

Materials The subjects were presented with a set of 78 randomly sampled sentences annotated by the two systems. Fifty percent of the data set were the sentences annotated by the identification system and the remaining 50% were annotated by the baseline; and the sentences were randomized. The annotation was done electronically in Microsoft Word. An example of annotated sentences is given in Figure 10.

Task and guidelines The subjects were asked to mark which of the expressions were metaphorical in their judgment. The participants were encouraged to rely on their own intuition of what a metaphor is in the annotation process. Additional guidance,

Table 5
Examples of seed set expansion by the baseline.

Seed phrase	Harvested metaphors	BNC frequency
reflect concern (V-O):	reflect concern	78
	ponder business	1
	ponder headache	1
	reflect business	4
	reflect care	2
	reflect fear	19
	reflect worry	3
cast doubt (V-O):	cast doubt	197
	cast question	11
	couch question	1
	drop question	2
	frame question	21
	purge doubt	2
	put doubt	12
	put question	151
	range question	1
	roll question	1
	shed doubt	2
	stray question	1
	throw doubt	35
	throw question	17
	throw uncertainty	1
campaign surged (S-V):	campaign surged	1
	campaign soared	1

however, in the form of the following definition of metaphor (Pragglejaz Group 2007) was also provided:

1. For each verb establish its meaning in context and try to imagine a more basic meaning of this verb in other contexts. Basic meanings normally are: (1) more concrete; (2) related to bodily action; (3) more precise (as opposed to vague); (4) historically older.
2. If you can establish a basic meaning that is distinct from the meaning of the verb in this context, the verb is likely to be used metaphorically.

CKM 391 Time and time again he would stare at the ground, hand on hip, if he thought he had received a bad call, and then **swallow his anger** and play tennis.
AD9 3205 He tried to **disguise the anxiety** he felt when he found the comms system down, but Tammuz was nearly hysterical by this stage.
AMA 349 We will **halt the reduction** in NHS services for long-term care and community health services which support elderly and disabled patients at home.
ADK 634 **Catch their interest** and **spark their enthusiasm** so that they begin to see the product's potential.
K2W 1771 The committee heard today that gangs regularly **hurled** abusive **comments** at local people, making an unacceptable level of noise and leaving litter behind them.

Figure 9
Sentences tagged by the system (metaphors in **bold**).

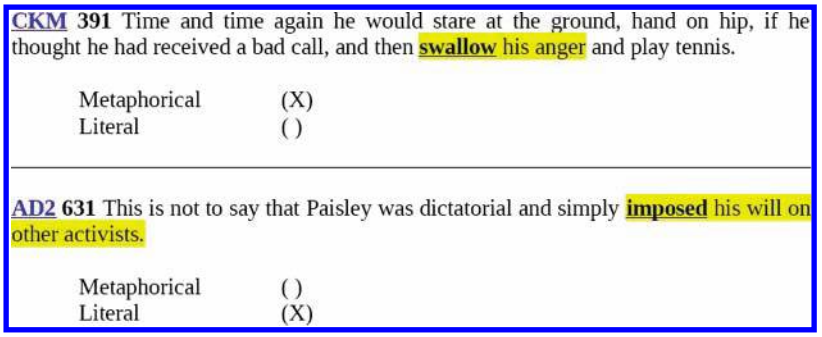

Figure 10
Evaluation of metaphor identification.

Interannotator agreement Reliability was measured at $\kappa = 0.63$ ($n = 2, N = 78, k = 5$). The data suggest that the main source of disagreement between the annotators was the presence of conventional metaphors (e.g., verbs such as *adopt, convey, decline*).

Results The system performance was then evaluated against the elicited judgments in terms of precision. The system output was compared to the gold standard constructed by merging the judgments, whereby the expressions tagged as metaphorical by at least three annotators were considered to be correct. This resulted in $P = 0.79$, with the baseline attaining $P = 0.44$. In addition, the system tagging was compared to that of each annotator pairwise, yielding an average $P = 0.74$ for the system and $P = 0.41$ for the baseline.

In order to compare system performance to the human ceiling, pairwise agreement was additionally calculated in terms of precision between the majority gold standard and each judge. This corresponds to an average of $P = 0.94$.

To show that the system performance is significantly different from that of the baseline, we annotated additional 150 instances identified by both systems for correctness and conducted a one-tailed t-test for independent samples. The difference is statistically significant with $t = 4.11$ ($df = 148, p < 0.0005$).

3.4 Discussion

We have shown that the method leads to a considerable expansion on the seed set and operates with high precision—namely, it produces high quality annotations, and identifies fully novel metaphorical expressions relying only on the knowledge of source–target domain mappings that it learns automatically. By comparing its coverage to that of a WordNet baseline, we showed that the method reaches beyond synonymy and generalizes well over the source and target domains.

The observed discrepancy in precision between the clustering approach and the baseline can be explained by the fact that a large number of metaphorical senses are included in WordNet. This means that in WordNet synsets source domain verbs appear together with more abstract terms. For instance, the metaphorical sense of *shape* in the phrase "*shape* opinion" is part of the synset "(determine, shape, mold, influence, regulate)." This results in the low precision of the baseline system, because it tags literal expressions (e.g., *influence opinion*) as metaphorical, assuming that all verbs from the synset belong to the source domain.

To perform a more comprehensive error analysis, we examined a larger subset of the metaphorical expressions identified by the system (200 sentences, equally covering verb–subject and verb–object constructions). System precision against the additional judgments by one of the authors was measured at 76% (48 instances were tagged incorrectly according to the judgments). The classification of system errors by type is presented in Table 6. Precision errors in the output of the system were also concentrated around the problem of conventionality of some metaphorical verbs, such as those in "*hold* views, *adopt* traditions, *tackle* a problem." This conventionality is reflected in the data in that such verbs are frequently used in their "metaphorical" contexts. As a result, they are clustered together with literally used terms. For instance, the verb *tackle* is found in a cluster with *solve, resolve, handle, confront, face*, and so forth. This results in the system tagging "resolve a problem" as metaphorical if it has previously seen "*tackle* a problem."

A number of system errors affecting its precision are also due to cases of general polysemy and homonymy of both verbs and nouns. For example, the noun *passage* can mean both "the act of passing from one state or place to the next" and "a section of text; particularly a section of medium length," as defined in WordNet. Sun and Korhonen's (2009) method performs hard clustering, that is, it does not distinguish between different word senses. Hence the noun *passage* occurred in only one cluster, containing concepts like *thought, word, sentence, expression, reference, address, description,* and so on. This cluster models the "textual" meaning of *passage*. As a result of sense ambiguity within the cluster, given the seed phrase "she *blocked* the thought," the system tags such expressions as "block passage," "impede passage," "obstruct passage," and "speed passage" as metaphorical.

The errors that may cause low recall of the system are of a different nature. Whereas noun clustering considerably expands the seed set by identifying new associated target concepts (e.g., given the seed metaphor "*sell* soul" it identifies "*sell* skin" and "*launch* pulse" as metaphorical), the verb clusters sometimes miss a certain proportion of source domain vocabulary. For instance, given the seed metaphor "example *illustrates*," the system identifies the following expressions: "history *illustrates*," "episode *illustrates*," "tale *illustrates*," "combination *illustrates*," "event *illustrates*," and so forth. It does not, however, capture obvious verb-based expansions, such as "episode *portrays*," present in the BNC. This is one of the problems that could lead to a lower recall of the system.

Nevertheless, in many cases the system benefits not only from dissimilar concepts within the noun clusters used to detect new target domains, but also from dissimilar concepts in the verb clusters. Verb clusters produced automatically relying on

Table 6
Common system errors by type.

Source of error	Subject–Verb	Verb–Object	Totals
Metaphor conventionality	7	14	21
General polysemy	9	6	15
Verb clustering	4	5	9
Noun clustering	2	1	3
SP filter	0	0	0
Totals	22	26	48

contextual features may contain lexical items with distinct, or even opposite meanings (e.g., *throw* and *catch*, *take off* and *land*). They tend to belong to the same semantic domain, however (e.g., verbs of dealing with LIQUIDS, verbs describing a FIGHT) It is the diversity of verb meanings within the domain cluster that allows the generalization from a limited number of seed expressions to a broader spectrum of previously unseen and novel metaphors, non-synonymous with those in the seed set.

The fact that the approach is seed-dependent is one of its possible limitations, affecting the coverage of the system. Wide coverage is essential for the practical use of the system. At this stage, however, it was impossible for us to reliably measure the recall of the system, because there is no large corpus annotated for metaphor available. In addition, because the current system was only tested with very few seeds (again, due to the lack of metaphor-annotated data), we expect the current overall recall of the system to be relatively low. In order to obtain a full coverage of the corpus, a large and representative seed set is necessary. Although it is hard to capture the whole variety of metaphorical language in a limited set of examples, it is possible to compile a seed set representative of all common source–target domain mappings. The learning capabilities of the system can then be used to expand on those to the whole range of conventional and novel metaphorical mappings and expressions. In addition, because the precision of the system was measured on the data set produced by expanding individual seed expressions, we would expect the expansion of other, new seed expressions to yield a comparable quality of annotations. Incorporating new seed expressions is thus likely to result in increasing recall without a significant loss in precision.

The current system harvests a large and relatively clean set of metaphorical expressions from the corpus. These annotations could provide a new platform for the development and testing of other metaphor systems.

4. Metaphor Interpretation Method and Experiments

As is the case in metaphor identification, the majority of existing approaches to metaphor interpretation also rely on task-specific hand-coded knowledge (Martin 1990; Fass 1991; Narayanan 1997, 1999; Barnden and Lee 2002; Feldman and Narayanan 2004; Agerri et al. 2007) and produce interpretations in a non-textual format (Veale and Hao 2008). The ultimate objective of automatic metaphor processing, however, is a type of interpretation that can be directly embedded into other systems to enhance their performance. We thus define metaphor interpretation as a paraphrasing task and build a system that automatically derives literal paraphrases for metaphorical expressions in unrestricted text. Our method is also distinguished from previous work in that it does not rely on any hand-crafted knowledge about metaphor, but in contrast is corpus-based and uses automatically induced selectional preferences.

The metaphor paraphrasing task can be divided into two subtasks: (1) generating *paraphrases*, that is, other ways of expressing the same meaning in a given context, and (2) discriminating between literal and metaphorical paraphrases. Consequently, the proposed approach is theoretically grounded in two ideas underlying each of these subtasks:

- The meaning of a word in context emerges through interaction with the meaning of the words surrounding it. This assumption is widely accepted in lexical semantics theory (Pustejovsky 1995; Hanks and Pustejovsky 2005) and has been exploited for lexical acquisition (Schulte im Walde 2006; Sun and Korhonen 2009). It suggests that the context itself imposes

certain semantic restrictions on the words which can occur within it. Given a large amount of linguistic data, it is possible to model these semantic restrictions in probabilistic terms (Lapata 2001). This can be done by deriving a ranking scheme for possible paraphrases that fit or do not fit in a specific context based on word co-occurrence evidence. This is how initial paraphrases are generated within the metaphor paraphrasing module.

- Literalness can be detected via strong selectional preference. This idea is a mirror-image of the selectional preference violation view of Wilks (1978), who suggested that a violation of selectional preferences indicates a metaphor. The key information that selectional preferences provide is whether there is an association between the predicate and its potential argument and how strong that association is. A literal paraphrase would normally come from the target domain (e.g., "understand the explanation") and be strongly associated with the target concept, whereas a metaphorical paraphrase would belong to the source domain (e.g., "*grasp* the explanation") and be associated with the concepts from this source domain more strongly than with the target concept. Hence we use a selectional preference model to measure the semantic fit of the generated paraphrases into the given context as opposed to all other contexts. The highest semantic fit then indicates the most literal paraphrase.

Thus the context-based probabilistic model is used for paraphrase generation and the selectional preference model for literalness detection. The key difference between the two models is that the former favors the paraphrases that co-occur with the words in the context more frequently than other paraphrases do, and the latter favors the paraphrases that co-occur with the words from the context more frequently than with any other lexical items in the corpus. This is the main intuition behind our approach.

The system thus incorporates the following components:

- **a context-based probabilistic model** that acquires paraphrases for metaphorical expressions from a large corpus;
- **a WordNet similarity component** that filters out the irrelevant paraphrases based on their similarity to the metaphorical term (similarity is defined as sharing a common hypernym within three levels in the WordNet hierarchy);
- **a selectional preference model** that discriminates literal paraphrases from the metaphorical ones. It re-ranks the paraphrases, de-emphasizing the metaphorical ones and emphasizing the literal ones.

In addition, the system disambiguates the sense of the paraphrases using the WordNet inventory of senses. The context-based model together with the WordNet filter constitute a metaphor paraphrasing baseline. By comparing the final system to this baseline, we demonstrate that simple context-based substitution, even supplied by extensive knowledge contained in lexical resources, is not sufficient for metaphor interpretation and that a selectional preference model is needed to establish the literalness of the paraphrases.

This section first provides an overview of paraphrasing and lexical substitution and relates these tasks to the problem of metaphor interpretation. It then describes the experimental data used to develop and test the paraphrasing system and the method itself, and finally, concludes with the system evaluation and the presentation of results.

4.1 Paraphrasing and Lexical Substitution

Paraphrasing can be viewed as a text-to-text generation problem, whereby a new piece of text is produced conveying the same meaning as the original text. Paraphrasing can be carried out at multiple levels (sentence-, phrase-, and word-levels), and may involve both syntactic and lexical transformations. Paraphrasing by replacing individual words in a sentence is known as **lexical substitution** (McCarthy 2002). Because, in this article, we address the phenomenon of metaphor at a single-word level, our task is close in nature to lexical substitution. The task of lexical substitution originates from word sense disambiguation (WSD). The key difference between the two is that whereas WSD makes use of a predefined sense-inventory to characterize the meaning of a word in context, lexical substitution is aimed at automatic induction of meanings. Thus the goal of lexical substitution is to generate the set of semantically valid substitutes for the word. Consider the following sentences from Preiss, Coonce, and Baker (2009).

(22) His parents felt that he was a bright boy.

(23) Our sun is a bright star.

Bright in Example (22) can be replaced by the word *intelligent*. The same replacement in the context of Example (23) will not produce an appropriate sentence. A lexical substitution system needs to (1) find a set of candidate synonyms for the word and (2) select the candidate that matches the context of the word best.

Both sentence- or phrase-level paraphrasing and lexical substitution find a wide range of applications in NLP. These include summarization (Knight and Marcu 2000; Zhou et al. 2006), information extraction (Shinyama and Sekine 2003), machine translation (Kurohashi 2001; Callison-Burch, Koehn, and Osborne 2006), text simplification (Carroll et al. 1999), question answering (McKeown 1979; Lin and Pantel 2001) and textual entailment (Sekine et al. 2007). Consequently, there has been a plethora of NLP approaches to paraphrasing (McKeown 1979; Meteer and Shaked 1988; Dras 1999; Barzilay and McKeown 2001; Lin and Pantel 2001; Barzilay and Lee 2003; Bolshakov and Gelbukh 2004; Quirk, Brockett, and Dolan 2004; Kauchak and Barzilay 2006; Zhao et al. 2009; Kok and Brockett 2010) and lexical substitution (McCarthy and Navigli 2007, 2009; Erk and Padó 2009; Preiss, Coonce, and Baker 2009; Toral 2009; McCarthy, Keller, and Navigli 2010).

Among paraphrasing methods one can distinguish (1) rule-based approaches, which rely on a set of hand-crafted (McKeown 1979; Zong, Zhang, and Yamamoto 2001) or automatically learned (Lin and Pantel 2001; Barzilay and Lee 2003; Zhao et al. 2008) paraphrasing patterns; (2) thesaurus-based approaches, which generate paraphrases by substituting words in the sentence by their synonyms (Bolshakov and Gelbukh 2004; Kauchak and Barzilay 2006); (3) natural language generation–based approaches (Kozlowski, McCoy, and Vijay-Shanker 2003; Power and Scott 2005), which transform a sentence into its semantic representation and generate a new sentence from it; and (4) SMT-based methods (Quirk, Brockett, and Dolan 2004), operating as monolingual

MT. A number of approaches to lexical substitution rely on manually constructed thesauri to find sets of candidate synonyms (McCarthy and Navigli 2007), whereas others address the task in a fully unsupervised fashion. In order to derive and rank candidate substitutes, the latter systems make use of distributional similarity measures (Pucci et al. 2009; McCarthy, Keller, and Navigli 2010), vector space models of word meaning (De Cao and Basili 2009; Erk and Padó 2009) or statistical learning techniques, such as hidden Markov models and *n*-grams (Preiss, Coonce, and Baker 2009).

The metaphor interpretation task is different from the WSD task, because it is impossible to predefine a set of senses of metaphorical words, in particular for novel metaphors. Instead, the correct substitute for the metaphorical term needs to be generated in a data-driven manner, as for lexical substitution. The metaphor paraphrasing task, however, also differs from lexical substitution in the following two ways. Firstly, a suitable substitute needs to be used literally in the target context, or at least more conventionally than the original word. Secondly, by definition, the substitution is not required to be a synonym of the metaphorical word. Moreover, for our task this is not even desired, because there is the danger that synonymous paraphrasing may result in another metaphorical expression, rather than the literal interpretation of the original one. Metaphor paraphrasing therefore presents an additional challenge in comparison to lexical substitution, namely, that of discriminating between literal and metaphorical substitutes. This second, harder, and not previously addressed task is the main focus of the work presented in this section. The remainder of the section is devoted to the description of the metaphor paraphrasing experiment.

4.2 Experimental Data

The paraphrasing system is first tested individually on a set of metaphorical expressions extracted from a manually annotated metaphor corpus of Shutova and Teufel (2010). This is the same data set as the one used for seeding the identification module (see Section 3.1.1 for description). Because the paraphrasing evaluation described in this section is conducted independently from the identification experiment, and no part of the paraphrasing system relies on the output of the identification system and vice versa, the use of the same data set does not give any unfair advantage to the systems. In the later experiment (Section 5) when the identification and paraphrasing system are evaluated jointly, again the same seed set will be used for identification; paraphrasing, however, will be performed on the output of the identification system (i.e., the new identified metaphors) and both the identified metaphors and their paraphrases will be evaluated by human judges not used in the previous and the current experiments.

4.3 Method

The system takes phrases containing annotated single-word metaphors as input; where a verb is used metaphorically, its context is used literally. It generates a list of possible paraphrases of the verb that can occur in the same context and ranks them according to their likelihood, as derived from the corpus. It then identifies shared features of the paraphrases and the metaphorical verb using the WordNet hierarchy and removes unrelated concepts. It then identifies the literal paraphrases among the remaining candidates based on the verb's automatically induced selectional preferences and the properties of the context.

4.3.1 Context-based Paraphrase Ranking Model. Terms replacing the metaphorical verb v will be called its interpretations i. We model the likelihood L of a particular paraphrase as a joint probability of the following events: the interpretation i co-occurring with the other lexical items from its context $w_1, ..., w_N$ in syntactic relations $r_1, ..., r_N$, respectively.

$$L_i = P(i, (w_1, r_1), (w_2, r_2), ..., (w_N, r_N)) \tag{6}$$

where $w_1, ..., w_N$ and $r_1, ..., r_N$ represent the fixed context of the term used metaphorically in the sentence. In the system output, the context $w_1, ..., w_N$ will be preserved, and the verb v will be replaced by the interpretation i.

We assume statistical independence between the relations of the terms in a phrase. For instance, for a verb that stands in a relation with both a subject and an object, the verb–subject and verb–direct object relations are considered to be independent events within the model. The likelihood of an interpretation is then calculated as follows:

$$P(i, (w_1, r_1), (w_2, r_2), ..., (w_N, r_N)) = P(i) \cdot P((w_1, r_1)|i) \cdot ... \cdot P((w_N, r_N)|i) \tag{7}$$

The probabilities can be calculated using maximum likelihood estimation

$$P(i) = \frac{f(i)}{\sum_k f(i_k)} \tag{8}$$

$$P(w_n, r_n|i) = \frac{f(w_n, r_n, i)}{f(i)} \tag{9}$$

where $f(i)$ is the frequency of the interpretation irrespective of its arguments, $\sum_k f(i_k)$ is the number of times its part of speech class is attested in the corpus, and $f(w_n, r_n, i)$ is the number of times the interpretation co-occurs with context word w_n in relation r_n. By performing appropriate substitutions into Equation (7) one obtains

$$P(i, (w_1, r_1), (w_2, r_2), ..., (w_N, r_N)) = \frac{f(i)}{\sum_k f(i_k)} \cdot \frac{f(w_1, r_1, i)}{f(i)} \cdot ... \cdot \frac{f(w_N, r_N, i)}{f(i)} = \frac{\prod_{n=1}^{N} f(w_n, r_n, i)}{(f(i))^{N-1} \cdot \sum_k f(i_k)} \tag{10}$$

This model is then used to rank the possible replacements of the term used metaphorically in the fixed context according to the data. The parameters of the model were estimated from the RASP-parsed BNC using the grammatical relations output created by Andersen et al. (2008).

4.3.2 WordNet Filter. The context-based model described in Section 4.3.1 overgenerates and hence there is a need to further narrow down the results. It is acknowledged in the linguistics community that metaphor is, to a great extent, based on similarity between the concepts involved (Gentner et al. 2001). We exploit this fact to refine paraphrasing. After obtaining the initial list of possible substitutes for the metaphorical term, the system filters out the terms whose meanings do not share any common properties with that of the metaphorical term. Consider the computer science metaphor "*kill* a process," which stands for "terminate a process." The basic sense of *kill* implies an *end*

Table 7
The list of paraphrases with the initial ranking (correct paraphrases are underlined).

Log-likelihood	Replacement
Verb–DirectObject	
hold back truth:	
−13.09	contain
−14.15	conceal
−14.62	suppress
−15.13	hold
−16.23	keep
−16.24	defend
stir excitement:	
−14.28	create
−14.84	provoke
−15.53	make
−15.53	elicit
−15.53	arouse
−16.23	stimulate
−16.23	raise
−16.23	excite
−16.23	conjure
leak report:	
−11.78	reveal
−12.59	issue
−13.18	disclose
−13.28	emerge
−14.84	expose
−16.23	discover
Subject–Verb	
campaign *surge*:	
−13.01	run
−15.53	improve
−16.23	soar
−16.23	lift

or *termination* of life. Thus *termination* is the shared element of the metaphorical verb and its literal interpretation.

Such an overlap of properties can be identified using the hyponymy relations in the WordNet taxonomy. Within the initial list of paraphrases, the system selects the terms that are hypernyms of the metaphorical term, or share a common hypernym with it. To maximize the accuracy, we restrict the hypernym search to a depth of three levels in the taxomomy. Table 7 shows the filtered lists of paraphrases for some of the test phrases, together with their log-likelihood. Selecting the highest ranked paraphrase from this list as a literal interpretation will serve as a baseline.

4.3.3 Re-ranking Based on Selectional Preferences. The lists which were generated contain some irrelevant paraphrases (e.g., "*contain* the truth" for "*hold back* the truth") and some paraphrases where the substitute itself is metaphorically used (e.g., "*suppress* the

truth"). As the task is to identify the literal interpretation, however, the system should remove these.

One way of dealing with both problems simultaneously is to use selectional preferences of the verbs. Verbs used metaphorically are likely to demonstrate semantic preference for the source domain, e.g., *suppress* would select for MOVEMENTS (political) rather than IDEAS, or TRUTH (the target domain), whereas the ones used literally for the target domain (e.g., *conceal*) would select for TRUTH. Selecting the verbs whose preferences the noun in the metaphorical expression matches best should allow filtering out non-literalness, as well as unrelated terms.

We automatically acquired selectional preference distributions of the verbs in the paraphrase lists (for verb–subject and verb–direct object relations) from the RASP-parsed BNC. As in the identification experiment, we derived selectional preference classes by clustering the 2,000 most frequent nouns in the BNC into 200 clusters using Sun and Korhonen's (2009) algorithm. In order to quantify how well a particular argument class fits the verb, we adopted the selectional association measure proposed by Resnik (1993), identical to the one we used within the selectional preference-based filter for metaphor identification, as described in Section 3.2.4. To remind the reader, selectional association is defined as follows:

$$A_R(v,c) = \frac{1}{S_R(v)} P(c|v) \log \frac{P(c|v)}{P(c)} \tag{11}$$

where $P(c)$ is the prior probability of the noun class, $P(c|v)$ is the posterior probability of the noun class given the verb, and S_R is the overall selectional preference strength of the verb in the grammatical relation R.

We use selectional association as a measure of semantic fitness (i.e., literalness) of the paraphrases. The paraphrases are re-ranked based on their selectional association with the noun in the context. Those paraphrases that are not well suited or used metaphorically are dispreferred within this ranking. The new ranking is shown in Table 8. The expectation is that the paraphrase in the first rank (i.e., the verb with which the noun in the context has the highest association) represents a literal interpretation.

4.4 Evaluation and Discussion

As in the case of identification, the paraphrasing system was tested on verb–subject and verb–direct object metaphorical expressions. These were extracted from the manually annotated metaphor corpus of Shutova and Teufel (2010), as described in Section 3.1.1. We compared the output of the final selectional-preference based system to that of the WordNet filter acting as a baseline. We evaluated the quality of paraphrasing with the help of human judges in two different experimental settings. The first setting involved direct judgments of system output by humans. In the second setting, the subjects did not have access to system output and had to provide their own literal paraphrases for the metaphorical expressions in the data set. The system was then evaluated against human judgments in Setting 1 and a paraphrasing gold standard created by merging annotations in Setting 2.

4.4.1 Setting 1: Direct Judgment of System Output. The subjects were presented with a set of sentences containing metaphorical expressions and the top-ranked paraphrases produced by the system and by the baseline, randomized. They were asked to mark as

Table 8
Paraphrases re-ranked by SP model (correct paraphrases are underlined).

Association	Replacement
Verb–DirectObject	
hold back truth:	
0.1161	conceal
0.0214	keep
0.0070	suppress
0.0022	contain
0.0018	defend
0.0006	hold
stir excitement:	
0.0696	provoke
0.0245	elicit
0.0194	arouse
0.0061	conjure
0.0028	create
0.0001	stimulate
≈ 0	raise
≈ 0	make
≈ 0	excite
leak report:	
0.1492	disclose
0.1463	discover
0.0674	reveal
0.0597	issue
≈ 0	emerge
≈ 0	expose
Subject–Verb	
campaign *surge*:	
0.0086	improve
0.0009	run
≈ 0	soar
≈ 0	lift

correct the paraphrases that have the same meaning as the term used metaphorically if they are used literally in the given context.

Subjects Seven volunteers participated in the experiment. They were all native speakers of English (one bilingual) and had little or no linguistics expertise.

Interannotator agreement The reliability was measured at $\kappa = 0.62$ ($n = 2$, $N = 95$, $k = 7$).

System evaluation against judgments We then evaluated the system performance against the subjects' judgments in terms of Precision at Rank 1, $P(1)$. Precision at Rank (1) measures the proportion of correct literal interpretations among the paraphrases in rank 1. The results are shown in Table 9. The system identifies literal paraphrases with a $P(1) = 0.81$ and the baseline with a $P(1) = 0.55$. We then conducted a one-tailed Sign test (Siegel and Castellan 1988) that showed that this difference in performance is statistically significant ($N = 15, x = 1, \mathrm{p} < 0.001$).

Table 9
System and baseline $P(1)$ and MAP.

Relation	System $P(1)$	Baseline $P(1)$	System MAP	Baseline MAP
Verb–DirectObject	0.79	0.52	0.60	0.54
Verb–Subject	0.83	0.57	0.66	0.57
Average	0.81	0.55	0.62	0.56

4.4.2 Setting 2: Creation of a Paraphrasing Gold Standard. The subjects were presented with a set of sentences containing metaphorical expressions and asked to write down all suitable literal paraphrases for the highlighted metaphorical verbs that they could think of.

Subjects Five volunteer subjects who were different from the ones used in the previous setting participated in this experiment. They were all native speakers of English and some of them had a linguistics background (postgraduate-level degree in English).

Gold Standard The elicited paraphrases combined together can be interpreted as a gold standard. For instance, the gold standard for the phrase "*brushed aside* the accusations" consists of the verbs *rejected*, *ignored*, *disregarded*, *dismissed*, *overlooked*, and *discarded*.

System evaluation by gold standard comparison The system output was compared against the gold standard using **mean average precision** (MAP) as a measure. MAP is defined as follows:

$$MAP = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N_j} \sum_{i=1}^{N_j} P_{ji} \tag{12}$$

where M is the number of metaphorical expressions, N_j is the number of correct paraphrases for the metaphorical expression j, P_{ji} is the precision at each correct paraphrase (the number of correct paraphrases among the top i ranks). First, average precision is estimated for individual metaphorical expressions, and then the mean is computed across the data set. This measure allows one to assess ranking quality beyond rank 1, as well as the recall of the system. As compared with the gold standard, MAP of the paraphrasing system is 0.62 and that of the baseline is 0.56, as shown in Table 9.

4.4.3 Discussion. Given that the metaphor paraphrasing task is open-ended, any gold standard elicited on the basis of it cannot be exhaustive. Some of the correct paraphrases may not occur to subjects during the experiment. As an example, for the phrase "*stir* excitement" most subjects suggested only one paraphrase "create excitement," which is found in rank 3, suggesting an average precision of 0.33 for this phrase. The top ranks of the system output are occupied by *provoke* and *stimulate*, however, which are intuitively correct, more precise paraphrases, despite none of the subjects having thought of them. Such examples contribute to the fact that the system's MAP is significantly lower than its precision at rank 1, because a number of correct paraphrases proposed by the system are not included in the gold standard.

The selectional preference-based re-ranking yields a considerable improvement in precision at rank 1 (26%) over the baseline. This component is also responsible for some errors of the system, however. One of the potential limitations of selectional preference-based approaches to metaphor paraphrasing is the presence of verbs exhibiting weak

selectional preferences. This means that these verbs are not strongly associated with any of their argument classes. As noted in Section 3, such verbs tend to be used literally, and are therefore suitable paraphrases. Our selectional preference model de-emphasizes them, however, and, as a result, they are not selected as literal paraphrases despite matching the context. This type of error is exemplified by the phrase "*mend* marriage." For this phrase, the system ranking overruns the correct top suggestion of the baseline, "improve marriage," and outputs "*repair* marriage" as the most likely literal interpretation, although it is in fact a metaphorical use. This is likely to be due to the fact that *improve* exposes a moderate selectional preference strength.

Table 10 provides frequencies of the common errors of the system by type. The most common type of error is triggered by the conventionality of certain metaphorical verbs. Because they frequently co-occur with the target noun class in the corpus, they receive a high association score with that noun class. This results in a high ranking of conventional metaphorical paraphrases. Examples of top-ranked metaphorical paraphrases include "*confront* a question" for "*tackle* a question," "*repair* marriage" for "*mend* marriage," "example *pictures*" for "example *illustrates*."

These errors concern non-literalness of the produced paraphrases. A less frequently occurring error was paraphrasing with a verb that has a different meaning. One such example was the metaphorical expression "tension *mounted*," for which the system produced a paraphrase "tension *lifted*," which has the opposite meaning. This error is likely to have been triggered by the WordNet filter, whereby one of the senses of *lift* would have a common hypernym with the metaphorical verb *mount*. This results in *lift* not being discarded by the filter, and subsequently ranked top due to the conventionality of the expression "tension *lifted*."

Another important issue that the paraphrase analysis brought to the foreground is the influence of wider context on metaphorical interpretation. The current system processes only the information contained within the GR of interest, discarding the rest of the context. For some cases, however, this is not sufficient and the analysis of a wider context is necessary. For instance, given the phrase "scientists *focus*" the system produces a paraphrase "scientists think," rather than the more likely paraphrase "scientists study." Such ambiguity of *focus* could potentially be resolved by taking its wider context into account. The context-based paraphrase ranking model described in Section 4.3.1 allows for the incorporation of multiple relations of the metaphorical verb in the sentence.

Although the paraphrasing system uses hand-coded lexical knowledge from WordNet, it is important to note that metaphor paraphrasing is not restricted to metaphorical senses included in WordNet. Even if a metaphorical sense is absent from WordNet, the system can still identify its correct literal paraphrase relying on the

Table 10
Common system errors by type.

Source of error	Subject–Verb	Verb–Object	Totals
Metaphor conventionality	0	5	5
General polysemy/WordNet filter	1	1	2
SP re-ranking	0	1	1
Lack of context	1	1	2
Totals	2	8	10

hyponymy relation and similarity between concepts, as described in Section 4.3.2. For example, the metaphorical sense of *handcuff* in "research is *handcuffed*" is not included in Wordnet, although the system correctly identifies its paraphrase *confine* ("research is confined").

5. Evaluation of Integrated System

Up to now, the identification and the paraphrasing systems were evaluated individually as modules. To determine to which extent the presented systems are applicable within NLP, we then ran the two systems together in a pipeline and evaluated the accuracy of the resulting text-to-text metaphor processing. First, the metaphor identification system was applied to naturally occurring text taken from the BNC and then the metaphorical expressions identified in those texts were paraphrased by the paraphrasing system. Some of the expressions identified and paraphrased by the integrated system are shown in Figure 11. The system output was compared against human judgments in two phases. In phase 1, a small sample of sentences containing metaphors identified and paraphrased by the system was judged by multiple judges. In phase 2, a larger sample of phrases was judged by only one judge (one of the authors of this article). Agreement of the judgments of the latter with the other judges was measured on the data from phase 1.

Because our goal was to evaluate both the accuracy of the integrated system and its usability by other NLP tasks, we assessed its performance in a two-fold fashion. Instances where metaphors were both correctly identified and paraphrased by the system were considered *strictly correct*, as they show that the system fully achieved

CKM 391 Time and time again he would stare at the ground, hand on hip, if he thought he had received a bad call, and then ***swallow*** **his anger** and play tennis.
CKM 391 Time and time again he would stare at the ground, hand on hip, if he thought he had received a bad call, and then **suppress his anger** and play tennis.

AD9 3205 He tried to ***disguise*** **the anxiety** he felt when he found the comms system down, but Tammuz was nearly hysterical by this stage.
AD9 3205 He tried to **hide the anxiety** he felt when he found the comms system down, but Tammuz was nearly hysterical by this stage.

AMA 349 We will ***halt*** **the reduction** in NHS services for long-term care and community health services which support elderly and disabled patients at home.
AMA 349 We will **prevent the reduction** in NHS services for long-term care and community health services which support elderly and disabled patients at home.

J7F 77 An economist would ***frame*** **this question** in terms of a cost-benefit analysis: the maximization of returns for the minimum amount of effort injected.
J7F 77 An economist would **phrase this question** in terms of a cost-benefit analysis: the maximization of returns for the minimum amount of effort injected.

EEC 1362 In it, Younger stressed the need for additional alternatives to custodial sentences, which had been implicit in the decision to ask the Council to ***undertake*** **the enquiry**.
EEC 1362 In it, Younger stressed the need for additional alternatives to custodial sentences, which had been implicit in the decision to ask the Council to **initiate the enquiry**.

A1F 24 Moreover, Mr Kinnock ***brushed aside*** **the suggestion** that he needed a big idea or unique selling point to challenge the appeal of Thatcherism.
A1F 24 Moreover, Mr Kinnock **dismissed the suggestion** that he needed a big idea or unique selling point to challenge the appeal of Thatcherism.

Figure 11
Metaphors identified (first sentences) and paraphrased (second sentences) by the system.

its goals. Instances where the paraphrasing retained the meaning and resulted in a literal paraphrase (including the cases where the identification module tagged a literal expression as a metaphor) were considered *correct lenient*. The intuition behind this evaluation setting is that correct paraphrasing of literal expressions by other literal expressions, albeit not demonstrating the positive contribution of metaphor processing, does not lead to any errors in system output and thus does not hamper the overall usability of the integrated system.

5.1 Phase 1: Small Sample, Multiple Judges

Three volunteer subjects participated in the experiment. They were all native speakers of English and had no formal training in linguistics.

Materials and task Subjects were presented with a set of sentences containing metaphorical expressions identified by the system and their paraphrases, as shown in Figure 12. There were 35 such sentences in the sample. They were asked to do the following:

1. Compare the sentences, decide whether the highlighted expressions have the same meaning, and record this in the box provided;
2. Decide whether the verbs in both sentences are used metaphorically or literally and tick the respective boxes.

For the second task, the same definition of metaphor as in the identification evaluation (cf. Section 3.3.2) was provided for guidance.

Interannotator agreement The reliability of annotations was evaluated independently for judgments on similarity of paraphrases and their literalness. The interannotator agreement on the task of distinguishing metaphoricity from literalness was measured at $\kappa = 0.53$ ($n = 2, N = 70, k = 3$). On the paraphrase (i.e., meaning retention) task, reliability was measured at $\kappa = 0.63$ ($n = 2, N = 35, k = 3$).

System performance We then evaluated the integrated system performance against the subjects' judgments in terms of accuracy (both strictly correct and correct

Example:

ACH **1081** His 'fascist' ideas had first been **shaped** by the First World War, which he felt Britain should not have entered.

ACH **1081** His 'fascist' ideas had first been **influenced** by the First World War, which he felt Britain should not have entered.

Do the highlighted expressions have the same meaning?
YES (X)
NO ( )

Is the **verb** in the first sentence used
metaphorically? (X)
literally? ( )

Is the **verb** in the second sentence used
metaphorically? ( )
literally? (X)

Figure 12
Evaluation of metaphor identification and paraphrasing.

Table 11
Integrated system performance.

Tagging case	Acceptability	Percentage
Correct paraphrase: metaphorical → literal	✔	53.8
Correct paraphrase: literal → literal	✔	13.5
Correct paraphrase: literal → metaphorical	✗	0.5
Correct paraphrase: metaphorical → metaphorical	✗	10.7
Incorrect paraphrase	✗	21.5

lenient). Strictly correct accuracy in this task measures the proportion of metaphors both identified and paraphrased correctly in the given set of sentences. Correct lenient accuracy, which demonstrates applicability of the system, is represented by the overall proportion of paraphrases that retained their meaning and resulted in a literal paraphrase (i.e., including literal paraphrasing of literal expressions in original sentences). Human judgments were merged into a majority gold standard, which consists of those instances that were considered correct (i.e., identified metaphor correctly paraphrased by the system) by at least two judges. Compared to this majority gold standard, the integrated system operates with a strictly correct accuracy of 0.66 and correct lenient accuracy of 0.71. The average human agreement with the majority gold standard in terms of accuracy is 0.80 on the literalness judgments and 0.89 on the meaning retention judgments.

5.2 Phase 2: Larger Sample, One Judge

The system was also evaluated on a larger sample of automatically annotated metaphorical expressions (600 sentences) using one person's judgments produced following the procedure from phase 1. We measured how far these judgments agree with the judges used in phase 1. The agreement on meaning retention was measured at $\kappa = 0.59$ ($n = 2, N = 35, k = 4$) and that on the literalness of paraphrases at $\kappa = 0.54$ ($n = 2, N = 70, k = 4$).

On this larger data set, the system achieved an accuracy of 0.54 (strictly correct) and 0.67 (correct lenient). The proportions of different tagging cases are shown in Table 11. The table also shows the acceptability of tagging cases. Acceptability indicates whether or not this type of system paraphrasing would cause an error when hypothetically integrated with an external NLP application. Cases where the system produces correct literal paraphrases for metaphorical expressions identified in the text would benefit another NLP application, whereas cases where literal expressions are correctly paraphrased by other literal expressions are considered neutral. Both such cases are deemed acceptable, because they increase or preserve literalness of the text. All other tagging cases introduce errors, thus they are marked as unacceptable. Examples of different tagging cases are shown in Table 12.

The accuracy of metaphor-to-literal paraphrasing (0.54) indicates the level of informative contribution of the system, and the overall accuracy of correct paraphrasing resulting in a literal expression (0.67) represents the level of its acceptability within NLP.

5.3 Discussion and Error Analysis

The results of integrated system evaluation suggest that the system is capable of providing useful information about metaphor for an external text processing application

Table 12
Examples of different tagging cases.

Tagging case	Examples
Correct paraphrase: met → lit	*throw* an idea → express an idea
Correct paraphrase: lit → lit	adopt a recommendation → accept a recommendation
Correct paraphrase: lit → met	arouse memory → *awaken* memory
Correct paraphrase: met → met	work *killed* him → work *exhausted* him

with a reasonable accuracy (0.67). It may, however, also introduce errors in the text by incorrect paraphrasing, as well as by producing metaphorical paraphrases. If the latter errors are rare (0.5%), the errors of the former type are sufficiently frequent (21.5%) to make the metaphor system less desirable for use in NLP. It is therefore important to address such errors.

Table 13 shows the contribution of the individual system components to the overall error. The identification system tags 28% of all instances incorrectly (170). This yields a component performance of 72%. This result is slightly lower than that obtained in its individual evaluation in a setting with multiple judges (79%). This can be explained by the fact that the integrated system was evaluated by one judge only, rather than using a majority gold standard. When compared with the judgments of each annotator pairwise the system precision was measured at 74% (cf. Section 3.3.2). Some of the literal instances tagged as a metaphor by the identification component are then correctly paraphrased with a literal expression by the paraphrasing component. Such cases do not change the meaning of the text, and hence are considered acceptable. The resulting contribution of the identification component to the overall error of the integrated system is thus 15%.

As Table 13 shows, the paraphrasing component failed in 32% of all cases (196 instances out of 600 were paraphrased incorrectly). As mentioned previously, this error can be further split into paraphrasing without meaning retention (21.5%) and metaphorical paraphrasing (11%). Both of these error types are unacceptable and lead to lower performance of the integrated system. This error rate is also higher than that of the paraphrasing system when evaluated individually on a manually created data set (19%). The reasons for incorrect paraphrasing by the integrated system are manifold

Table 13
System errors by component. Three categories are cases where the identification model incorrectly tagged a literal expression as metaphoric (false negatives from this module were not measured). The remaining two categories are for paraphrase errors on correctly identified metaphors.

Type of error	Identification	Paraphrasing
Correct paraphrase: lit → lit	81	0
Correct paraphrase: lit → met	3	3
Correct paraphrase: met → met	–	64
Incorrect paraphrase for literal	86	86
Incorrect paraphrase for metaphor	–	43
Totals	170	196

and concern both the metaphor identification and paraphrasing components. One of the central problems stems from the initial tagging of literal expressions as metaphorical by the identification system. The paraphrasing system is not designed with literal-to-literal paraphrasing in mind. When it receives literal expressions which have been incorrectly identified as input, it searches for a more literal paraphrase for them. Not all literally used words have suitable substitutes in the given context, however. For instance, the literal expression "approve conclusion" is incorrectly paraphrased as "evaluate conclusion."

Similar errors occur when metaphorical expressions do not have any single-word literal paraphrases, for example, "country *functions* according to...". This is, however, a more fundamental problem for metaphor paraphrasing as a task. In such cases, the system, nonetheless, attempts to produce a substitute with approximately the same meaning, which often leads to either metaphorical or incorrect paraphrasing. For instance, "country *functions*" is paraphrased by "country *runs*," with suggestions with lower rank being "country *works*" and "country *operates*."

Some errors that occur at the paraphrasing level are also due to the general word sense ambiguity of certain verbs or nouns. Consider the following paraphrasing example, where Example (24a) shows an automatically identified metaphor and Example (24b) its system-derived paraphrase:

(24) a. B71 852 Craig Packer and Anne Pusey of the University of Chicago have continued to *follow* the life and loves of these Tanzanian lions.

b. B71 852 Craig Packer and Anne Pusey of the University of Chicago have continued to succeed the life and loves of these Tanzanian lions.

This error results from the fact that the verb *succeed* has a high selectional preference for *life* in one of its senses ("attain success or reach a desired goal") and is similar to *follow* in WordNet in another of its senses ("be the successor [of]"). The system merges these two senses in one, resulting in an incorrect paraphrase.

One automatically identified example exhibited interaction of metaphor with metonymy at the interpretation level. In the phrase "*break* word," the verb *break* is used metaphorically (although conventionally) and the noun *word* is a metonym standing for *promise*. This affected paraphrasing in that the system searched for verbs denoting actions that could be done with words, rather than promises, and suggested the paraphrase "interrupt word(s)." This paraphrase is interpretable in the context of a person giving a speech, but not in the context of a person giving a promise. This was the only case of metonymy in the analyzed data, however.

Another issue that the evaluation on a larger data set revealed is the limitations of the WordNet filter used in the paraphrasing system. Despite being a wide-coverage general-domain database, WordNet does not include information about all possible relations that exist between particular word senses. This means that some of the correct paraphrases suggested by the context-based model get discarded by the WordNet filter due to missing information in WordNet. For instance, the system produces no paraphrase for the metaphors "*hurl* comment," "*spark* enthusiasm," and "*magnify* thought" that it correctly identified. This problem motivates the exploration of possible WordNet-free solutions for similarity detection in the metaphor paraphrasing task. The system could either rely entirely on such a solution, or back off to it in cases when the WordNet-based system fails.

Table 14 provides a summary of system errors by type. The most common errors are caused by metaphor conventionality resulting in metaphorical paraphrasing (e.g.,

Table 14
Errors of the paraphrasing component by type.

Source of error	Met→Met	Lit→Met	Incorr. for Lit	Incorr. for Met	Total
No literal paraphrase exists	11	0	5	2	18
Metaphor conventionality	53	3	0	0	56
General polysemy	0	0	13	10	23
WordNet filter	0	0	21	21	42
SP re-ranking	0	0	41	7	48
Lack of context	0	0	6	2	8
Interaction with metonymy	0	0	0	1	1
Totals	64	3	86	43	196

"*swallow* anger → *suppress* anger," "work *killed* him → work *exhausted* him"), followed by the WordNet filter– and general polysemy–related errors (e.g. "*follow* lives → *succeed* lives"), resulting in incorrect paraphrasing or the system not producing any paraphrase at all. Metaphor paraphrasing by another conventional metaphor instead of a literal expression is undesirable, although it may still be useful if the paraphrases are more lexicalized than the original expression. The word sense ambiguity– and WordNet-based errors are more problematic, however, and need to be addressed in the future. SP re-ranking is responsible for the majority of incorrect paraphrasing of literal expressions. This may be due to the fact that the model is ignorant of the meaning retention aspect, but rather favors the paraphrases that are used literally (albeit incorrectly) in the given context. This shows that when building an integrated system, it is necessary to adapt the metaphor paraphrasing module to be able to also handle literal expressions, because the identification module is likely to produce at least some of them.

5.4 Comparison to the CorMet System

It is hard to directly compare the performance of the presented system to the other recent approaches to metaphor, because all of these approaches assume different task definitions, and hence use data sets and evaluation techniques of their own. Among the data-driven methods, however, the closest in nature to ours is Mason's (2004) CorMet system. Mason's system does not perform metaphor interpretation or identification of metaphorical expressions in text, but rather focuses on the detection of metaphorical links between distant domains. Our system also involves such detection. Whereas Mason relies on domain-specific selectional preferences for this purpose, however, our system uses information about verb subcategorization, as well as general selectional preferences, to perform distributional clustering of verbs and nouns and then link the clusters based on metaphorical seeds. Another fundamental difference is that whereas CorMet assigns explicit domain labels, our system models source and target domains implicitly. In the evaluation of the CorMet system, the acquired metaphorical mappings are compared to those in the manually created Master Metaphor List demonstrating the accuracy of 77%. In our system, on the contrary, metaphor acquisition is evaluated via extraction of naturally occurring metaphorical expressions, achieving a performance of 79% in terms of precision. In order to compare the new mapping acquisition ability

of our system to that of CorMet, however, we performed an additional analysis of the mappings hypothesized by our noun clusters in relation to those in the MML. It was not possible to compare the new mappings discovered by our system to the MML directly as was done in Mason's experiments, because in our approach source domains are represented by clusters of their characteristic verbs. The analysis of the noun clusters with respect to expansion of the the seed mappings taken from the MML, however, allowed us to evaluate the mapping acquisition by our system in terms of both precision and recall. The goal was to confirm our hypothesis that abstract concepts get clustered together if they are associated with the same source domain and to evaluate the quality of the newly acquired mappings.

To do this, we randomly selected 10 target domain categories described in the MML and manually extracted all corresponding mappings (42 mappings in total). The categories included SOCIETY, IDEA, LIFE, OPPORTUNITY, CHANGE, LOVE, DIFFICULTY, CREATION, RELATIONSHIP, and COMPETITION. For the concept of OPPORTUNITY, for example, three mappings were present in the MML: OPPORTUNITIES ARE PHYSICAL OBJECTS, OPPORTUNITIES ARE MOVING ENTITIES, and OPPORTUNITIES ARE OPEN PATHS, whereas for the concept of COMPETITION the list describes only two mappings: COMPETITION IS A RACE, COMPETITION IS A WAR.

We then extracted the system-produced clusters containing the selected 10 target concepts. Examples of the mappings and the corresponding clusters are shown in Figure 13. Our goal was to verify whether other concepts in the cluster containing the target concept are associated with the source domains given in the mappings. Each member of these clusters was analyzed for possible association with the respective source domains. For each concept in a cluster, we verified that it is associated with the respective source domain by finding a corresponding metaphorical expression and annotating the concepts accordingly. The degree of association of the members of the clusters with a given source domain was evaluated in terms of precision on the set of hypothesized mappings. The precision of the cluster's association with the source

Conceptual mapping: RELATIONSHIP IS A MECHANISM (VEHICLE)
Cluster: consensus relation tradition partnership resistance foundation alliance friendship contact reserve unity link peace bond myth identity hierarchy relationship connection balance marriage democracy defense faith empire distinction coalition regime division
Conceptual mapping: LIFE IS A JOURNEY
Cluster: politics practice trading reading occupation profession sport pursuit affair career thinking life
Conceptual mapping: SOCIETY is a (HUMAN) BODY
Cluster: class population nation state country family generation trade profession household kingdom business industry economy market enterprise world community institution society sector
Conceptual mapping: DIFFICULTY is DIFFICULTY IN MOVING (OBSTACLE); PHYSICAL HARDNESS
Cluster: threat crisis risk problem poverty obstacle dilemma challenge prospect danger discrimination barrier difficulty shortage
Conceptual mapping: OPPORTUNITY is a PHYSICAL OBJECT; MOVING ENTITY; OPEN PATH
Cluster: incentive attraction scope remedy chance choice solution option perspective range possibility contrast opportunity selection alternative focus

Figure 13
Noun clusters.

concept was calculated as a proportion of the associated concepts in it. Based on these results we computed the average precision (AP) as follows:

$$AP = \frac{1}{M} \sum_{j=1}^{M} \frac{\#\text{associated concepts in cluster } c_j}{|c_j|} \tag{13}$$

where M is the number of hypothesized mappings and c_j is the cluster of target concepts corresponding to mapping j.

The annotation was carried out by one of the authors and its average precision is 0.82. This confirms the hypothesis of clustering by association and shows that our method favorably compares to Mason's system. This is only an approximate comparison, however. Direct comparison of metaphor acquisition by the two systems was not possible, as they produce the output in different formats and, as mentioned earlier, our system models conceptual mappings implicitly, both within the noun clusters, as well as by linking them to the verb clusters.

We then additionally evaluated the recall of mapping acquisition by our system against the MML. For each selected MML mapping, we manually extracted all alternative target concepts associated with the source domain in the mapping from the MML. For example, in case of LIFE IS A JOURNEY we identified all target concepts associated with JOURNEY according to the MML and extracted them. These included LIFE, CAREER, LOVE, and CHANGE. We then verified whether the relevant system-produced noun clusters contained these concepts. The recall was then calculated as a proportion of the concepts in this list within one cluster. For example, the concepts LIFE and CAREER are found in the same cluster, but not LOVE and CHANGE. The overall recall of mapping acquisition was measured at 0.50.

These results show that the system is able to discover a large number of metaphorical connections in the data with high precision. Although the evaluation against the Master Metaphor List is subjective, it suggests that the use of statistical data-driven methods in general, and distributional clustering in particular, is a promising direction for computational modeling of metaphor.

6. Conclusion and Future Directions

The 1980s and 1990s provided us with a wealth of ideas on the structure and mechanisms of metaphor. The computational approaches formulated back then are still highly influential, although their use of task-specific hand-coded knowledge is becoming increasingly less popular. The last decade witnessed a significant technological leap in natural language computation, whereby manually crafted rules gradually gave way to more robust corpus-based statistical methods. This is also the case for metaphor research. In this article, we presented the first integrated statistical system for metaphor processing in unrestricted text. Our method is distinguished from previous work in that it does not rely on any metaphor-specific hand-coded knowledge (besides the seed set in the identification experiments), operates on open-domain text, and produces interpretations in textual format. The system, consisting of independent metaphor identification and paraphrasing modules, operates with a high precision (0.79 for identification, 0.81 for paraphrasing, and 0.67 as an integrated system). Although the system has been tested only on verb–subject and verb–object metaphors at this stage, the described identification and paraphrasing methods should be similarly applicable to a wider range of syntactic constructions. This expectation rests on the fact that both distributional

clustering and selectional preference induction techniques have been shown to model the meanings of a range of word classes (Hatzivassiloglou and McKeown 1993; Boleda Torrent and Alonso i Alemany 2003; Brockmann and Lapata 2003; Zapirain, Agirre, and Màrquez 2009). Extending the system to deal with metaphors represented by other word classes and constructions as well as multi-word metaphors is part of future work.

Such an extension of the identification system would require the creation of a seed set exemplifying more syntactic constructions and the corpus search over further grammatical relations (e.g., verb–prepositional phrase [PP] complement relations: "Hillary *leapt* in the conversation;" adjectival modifier–noun relations "*slippery* mind, *deep* unease, *heavy* loss;" noun–PP complement relations: "a *fraction* of self-control, a *foot* of a mountain;" verb–VP complement relations: "*aching* to begin the day;" and copula constructions: "Death is the sorry end of the human story, not a mysterious prelude to a new one"). Besides noun and verb clustering, it would also be necessary to perform clustering of adjectives and adverbs. Clusters of verbs, adjectives, adverbs, and concrete nouns would then represent source domains within the model. The data study of Shutova and Teufel (2010) suggested that it is sometimes difficult to choose the optimal level of abstraction of domain categories that would generalize well over the data. Although the system does not explicitly assign any domain labels, its domain representation is still restricted by the fixed level of generality of source concepts, defined by the chosen cluster granularity. To relax this constraint, one could attempt to automatically optimize cluster granularity to fit the data more accurately and to ensure that the generated clusters explain the metaphorical expressions in the data more comprehensively. A hierarchical clustering algorithm, such as that of Yu, Yu, and Tresp (2006) or Sun and Korhonen (2011), could be used for this purpose. Besides this, it would be desirable to be able to generalize metaphorical associations learned from one type of syntactic construction across all syntactic constructions, without providing explicit seed examples for the latter. For instance, given the seed phrase "*stir* excitement," representing the conceptual mapping FEELINGS ARE LIQUIDS, the system should be able to discover not only that phrases such as "*swallow* anger" are metaphorical, but that phrases such as "*ocean* of happiness" are as well.

The extension of the paraphrasing system to other syntactic constructions would involve the extraction of further grammatical relations from the corpus, such as those listed herein, and their incorporation into the context-based paraphrase selection model. Extending both the identification system and the paraphrasing system would require the application of the selectional preference model to other word classes. Although Resnik's selectional association measure has been used to model selectional preferences of verbs for their nominal arguments, it is in principle a generalizable measure of word association. Information-theoretic word association measures (e.g., mutual information [Church and Hanks 1990]) have been continuously successfully applied to a range of syntactic constructions in a number of NLP tasks (Hoang, Kim, and Kan 2009; Baldwin and Kim 2010). This suggests that applying a distributional association measure, such as the one proposed by Resnik, to other part-of-speech classes should still result in a realistic model of semantic fitness, which in our terms corresponds to a measure of "literalness" of the paraphrases.

In addition, the selectional preference model can be improved by using an SP acquisition algorithm that can handle word sense ambiguity (e.g., Rooth et al. 1999; Ó Séaghdha 2010; Reisinger and Mooney 2010). The current approach relies on SP classes produced by hard clustering and fails to accurately model word senses of generally polysemous words. This resulted in a number of errors in metaphor paraphrasing and it therefore needs to be addressed in the future.

The current version of the metaphor paraphrasing system still relies on some hand-coded knowledge in the form of WordNet. WordNet has been criticized for a lack of consistency, high granularity of senses, and negligence with respect to some important semantic relations (Lenat, Miller, and Yokoi 1995). In addition, WordNet is a general-domain resource, which is less suitable if one wanted to apply the system to domain-specific data. For all of these reasons it would be preferable to develop a WordNet-free fully automated approach to metaphor resolution. Vector space models of word meaning (Erk 2009; Rudolph and Giesbrecht 2010; Van de Cruys, Poibeau, and Korhonen 2011) might provide a solution, as they have proved efficient in general paraphrasing and lexical substitution settings (Erk and Padó 2009). The feature similarity component of the paraphrasing system that is currently based on WordNet could be replaced by such a model.

Another crucial problem that needs to be addressed is the coverage of the identification system. To enable high usability of the system it is necessary to perform high-recall processing. One way to improve the coverage is the creation of a larger, more diverse seed set. Although it is hardly possible to describe the whole variety of metaphorical language, it is possible to compile a set representative of (1) all most common source–target domain mappings and (2) all types of syntactic constructions that exhibit metaphoricity. The existing metaphor resources, primarily the Master Metaphor List (Lakoff, Espenson, and Schwartz 1991), and examples from the linguistic literature about metaphor, could be a sensible starting point on a route to such a data set. Having a diverse seed set should enable the identification system to attain a broad coverage of the corpus.

The proposed text-to-text representation of metaphor processing is directly transferable to other NLP tasks and applications that could benefit from the inclusion of a metaphor processing component. Overall, our results suggest that the system can provide useful and accurate information about metaphor to other NLP tasks relying on lexical semantics. In order to prove its usefulness for external applications, however, an extrinsic task-based evaluation is outstanding. In the future, we intend to integrate metaphor processing with NLP applications, exemplified by MT and opinion mining, in order to demonstrate the contribution of this pervasive yet rarely addressed phenomenon to natural language semantics.

Acknowledgments

We would like to thank the volunteer annotators for their help in the evaluations, as well as the Cambridge Overseas Trust (UK), EU FP-7 PANACEA project, and the Royal Society (UK), who funded our work.

References

Abend, Omri and Ari Rappoport. 2010. Fully unsupervised core-adjunct argument classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 226–236, Uppsala.

Agerri, Rodrigo, John Barnden, Mark Lee, and Alan Wallington. 2007. Metaphor, inference and domain-independent mappings. In *Proceedings of RANLP-2007*, pages 17–23, Borovets.

Alonge, Antonietta and Margherita Castelli. 2003. Encoding information on metaphoric expressions in WordNet-like resources. In *Proceedings of the ACL 2003 Workshop on Lexicon and Figurative Language*, pages 10–17, Sapporo.

Andersen, Oistein, Julien Nioche, Ted Briscoe, and John Carroll. 2008. The BNC parsed with RASP4UIMA. In *Proceedings of LREC 2008*, pages 865–869, Marrakech.

Baldwin, Timothy and Su Nam Kim. 2010. Multiword expressions. In N. Indurkhya and F. J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, pages 267–292.

Barnden, John and Mark Lee. 2002. An artificial intelligence approach to metaphor understanding. *Theoria et Historia Scientiarum*, 6(1):399–412.

Barque, Lucie and François-Régis Chaumartin. 2009. *LDV Forum*, 24(2):5–18.

Barzilay, Regina and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 16–23, Edmonton.

Barzilay, Regina and Kathryn McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 50–57, Toulouse.

Bergsma, Shane, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 59–68, Honolulu, HI.

Birke, Julia and Anoop Sarkar. 2006. A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL-06*, pages 329–336, Trento.

Black, Max. 1962. *Models and Metaphors*. Cornell University Press, Ithaca, NY.

Boleda Torrent, Gemma and Laura Alonso i Alemany. 2003. Clustering adjectives for class acquisition. In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 9–16, Budapest.

Bolshakov, Igor and Alexander Gelbukh. 2004. Synonymous paraphrasing using Wordnet and Internet. In *Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems, NLDB 2004*, pages 312–323, Alicante.

Brew, Chris and Sabine Schulte im Walde. 2002. Spectral clustering for German verbs. In *Proceedings of EMNLP*, pages 117–124, Philadelphia, PA.

Briscoe, Ted, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, pages 77–80, Sydney.

Brockmann, Carsten and Mirella Lapata. 2003. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 27–34, Budapest.

Burnard, Lou. 2007. *Reference Guide for the British National Corpus (XML Edition)*. Available at `http://www.natcorp.ox.ac.uk/docs/URG`.

Callison-Burch, Chris, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of NAACL*, HLT-NAACL '06, pages 17–24, New York, NY.

Cameron, Lynne. 2003. *Metaphor in Educational Discourse*. Continuum, London.

Carroll, John, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 269–270, Bergen.

Chen, Jinxiu, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2006. Unsupervised relation disambiguation using spectral clustering. In *Proceedings of the COLING/ACL*, pages 89–96, Sydney.

Church, Kenneth and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Clark, Stephen and James Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Copestake, Ann and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67.

Davidov, Dmitry, Roi Reichart, and Ari Rappoport. 2009. Superior and efficient fully unsupervised pattern-based concept acquisition using an unsupervised parser. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 48–56, Boulder, CO.

De Cao, Diego and Roberto Basili. 2009. Combining distributional and paradigmatic information in a lexical substitution task. In *Proceedings of EVALITA Workshop, 11th Congress of Italian Association for Artificial Intelligence*, Reggie Emilia.

Dras, Mark. 1999. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. thesis, Macquarie University, Australia.

Erk, Katrin. 2009. Representing words as regions in vector space. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 57–65, Boulder, CO.

Erk, Katrin and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, Edinburgh.

Erk, Katrin and Sebastian Padó. 2009. Paraphrase assessment in structured vector space: exploring parameters and datasets. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 57–65, Athens.

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

Fass, Dan and Yorick Wilks. 1983. Preference semantics, ill-formedness, and metaphor. *Computational Linguistics*, 9(3-4):178–187.

Fauconnier, Gilles and Mark Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, New York, NY.

Feldman, Jerome. 2006. *From Molecule to Metaphor: A Neural Theory of Language*. The MIT Press, Cambridge, MA.

Feldman, Jerome and Srini Narayanan. 2004. Embodied meaning in a neural theory of language. *Brain and Language*, 89(2):385–392.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database* (ISBN: 0-262-06197-X). MIT Press, Cambridge, MA.

Fillmore, Charles, Christopher Johnson, and Miriam Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

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

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

Gentner, Dedre, Brian Bowdle, Phillip Wolff, and Consuelo Boronat. 2001. Metaphor is like analogy. In D. Gentner, K. J. Holyoak, and B. N. Kokinov, editors, *The Analogical Mind: Perspectives from Cognitive Science*. MIT Press, Cambridge, MA, pages 199–253.

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

Grady, Joe. 1997. *Foundations of Meaning: Primary Metaphors and Primary Scenes*. Ph.D. thesis, University of California at Berkeley.

Hanks, Patrick and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2):63–82.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, ACL '93, pages 172–182, Columbus, OH.

Hesse, Mary. 1966. *Models and Analogies in Science*. Notre Dame University Press, Notre Dame, IN.

Hoang, Hung Huu, Su Nam Kim, and Min-Yen Kan. 2009. A re-examination of lexical association measures. In *Proceedings of the Workshop on Multiword Expressions*, pages 31–39, Singapore.

Hofstadter, Douglas. 1995. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. HarperCollins Publishers, London.

Hofstadter, Douglas and Melanie Mitchell. 1994. The Copycat Project: A model of mental fluidity and analogy-making. In K. J. Holyoak and J. A. Barnden, editors, *Advances in Connectionist and Neural Computation Theory*. Ablex, New York, NY.

Karov, Yael and Shimon Edelman. 1998. Similarity-based word sense disambiguation. *Computational Linguistics*, 24(1):41–59.

Kauchak, David and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Main Conference on Human Language Technology, Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 455–462, New York, NY.

Kingsbury, Paul and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of LREC-2002*, pages 1989–1993, Gran Canaria, Canary Islands.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extensive classifications of English verbs. In *Proceedings of the 12th EURALEX International Congress*, pages 1–15, Torino.

Klein, Dan and Christopher Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo.

Knight, Kevin and Daniel Marcu. 2000. Statistics-based summarization—step one: Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial*

Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pages 703–710, Austin, TX.

Kok, Stanley and Chris Brockett. 2010. Hitting the right paraphrases in good time. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 145–153, Los Angeles, CA.

Korhonen, Anna, Yuval Krymolowski, and Ted Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proceedings of LREC 2006*, pages 1015–1020, Genoa.

Kozlowski, Raymond, Kathleen F. McCoy, and K. Vijay-Shanker. 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *Proceedings of the Second International Workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 1–8, Sapporo.

Krishnakumaran, Saisuresh and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 13–20, Rochester, NY.

Kurohashi, Sadao. 2001. SENSEVAL-2 Japanese translation task. In *Proceedings of the SENSEVAL-2 Workshop*, pages 37–44, Toulouse.

Lakoff, George, Jane Espenson, and Alan Schwartz. 1991. The master metaphor list. Technical report, University of California at Berkeley.

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

Lapata, Mirella. 2001. *The Acquisition and Modeling of Lexical Knowledge: A Corpus-Based Investigation of Systematic Polysemy*. Ph.D. thesis, University of Edinburgh.

Lenat, Doug, George Miller, and Toshio Yokoi. 1995. CYC, WordNet, and EDR: Critiques and responses. *Commun. ACM*, 38(11):45–48.

Levin, Beth. 1993. *English Verb Classes and Alternations*. University of Chicago Press, Chicago, IL.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 768–774, Montreal.

Lin, Dekang and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7:343–360.

Lönneker, Birte. 2004. Lexical databases as resources for linguistic creativity: Focus on metaphor. In *Proceedings of the LREC 2004 Workshop on Language Resources for Linguistic Creativity*, pages 9–16, Lisbon.

Lönneker, Birte and Carina Eilts. 2004. A current resource and future perspectives for enriching Wordnets with metaphor information. In *Proceedings of the Second International WordNet Conference—GWC 2004*, pages 157–162, Brno.

Martin, James. 1988. Representing regularities in the metaphoric lexicon. In *Proceedings of the 12th Conference on Computational Linguistics*, pages 396–401, Budapest.

Martin, James. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press Professional, Inc., San Diego, CA.

Martin, James. 1994. Metabank: A knowledge-base of metaphoric language conventions. *Computational Intelligence*, 10:134–149.

Martin, James. 2006. A corpus-based analysis of context effects on metaphor comprehension. In A. Stefanowitsch and S. T. Gries, editors, *Corpus-Based Approaches to Metaphor and Metonymy*. Mouton de Gruyter, Berlin, pages 214–236.

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

McCarthy, Diana. 2002. Lexical substitution as a task for WSD evaluation. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions - Volume 8*, WSD '02, pages 109–115, Philadelphia, PA.

McCarthy, Diana, Bill Keller, and Roberto Navigli. 2010. Getting synonym candidates from raw data in the English lexical substitution task. In *Proceedings of the 14th EURALEX International Congress*, Leeuwarden.

McCarthy, Diana and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague.

McCarthy, Diana and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

McKeown, Kathleen. 1979. Paraphrasing using given and new information in a question-answer system. In *Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics*, ACL '79, pages 67–72, La Jolla, CA.
Meila, Marina and Jianbo Shi. 2001. A random walks view of spectral segmentation. In *AISTATS*, Key West, FL.
Meteer, Marie and Varda Shaked. 1988. Strategies for effective paraphrasing. In *Proceedings of the 12th Conference on Computational Linguistics - Volume 2*, COLING '88, pages 431–436, Budapest.
Mitchell, Jeff and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL*, pages 236–244, Columbus, OH.
Murphy, Gregory. 1996. On metaphoric representation. *Cognition*, 60:173–204.
Narayanan, Srini. 1997. *Knowledge-based Action Representations for Metaphor and Aspect (KARMA)*. Ph.D. thesis, University of California at Berkeley.
Narayanan, Srini. 1999. Moving right along: A computational model of metaphoric reasoning about events. In *Proceedings of AAAI 99*, pages 121–128, Orlando, FL.
Nunberg, Geoffrey. 1987. Poetic and prosaic metaphors. In *Proceedings of the 1987 Workshop on Theoretical Issues in Natural Language Processing*, pages 198–201, Stroudsburg, PA.
Ó Séaghdha, Diarmuid. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444, Uppsala.
Orwell, George. 1946. Politics and the English Language. *Horizon*, 13(76):252–265.
Pantel, Patrick and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton.
Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of ACL-93*, pages 183–190, Morristown, NJ.
Peters, Wim and Ivonne Peters. 2000. Lexicalised systematic polysemy in Wordnet. In *Proceedings of LREC 2000*, Athens.
Pinker, Stephen. 2007. *The Stuff of Thought: Language as a Window into Human Nature*. Viking Adult, New York, NY.
Power, Richard and Donia Scott. 2005. Automatic generation of large-scale paraphrases. In *Proceedings of IWP*, pages 73–79.
Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22:1–39.
Preiss, Judita, Ted Briscoe, and Anna Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of ACL-2007*, volume 45, page 912, Prague.
Preiss, Judita, Andrew Coonce, and Brittany Baker. 2009. HMMs, GRs, and n-grams as lexical substitution techniques: are they portable to other languages? In *Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning*, MCTLLL '09, pages 21–27, Borovets.
Pucci, Dario, Marco Baroni, Franco Cutugno, and Alessandro Lenci. 2009. Unsupervised lexical substitution with a word space model. In *Proceedings of the EVALITA Workshop, 11th Congress of Italian Association for Artificial Intelligence*, Reggie Emilia.
Pustejovsky, James. 1995. *The Generative Lexicon.* MIT Press, Cambridge, MA.
Quirk, Chris, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 142–149, Barcelona.
Reining, Astrid and Birte Lönneker-Rodman. 2007. Corpus-driven metaphor harvesting. In *Proceedings of the HLT/NAACL-07 Workshop on Computational Approaches to Figurative Language*, pages 5–12, Rochester, NY.
Reisinger, Joseph and Raymond Mooney. 2010. A mixture model with sharing for lexical semantics. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1173–1182, Cambridge, MA.
Resnik, Philip. 1993. *Selection and Information: A Class-based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.
Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of ACL 99*, pages 104–111, Maryland.
Rudolph, Sebastian and Eugenie Giesbrecht. 2010. Compositional matrix-space

models of language. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 907–916, Uppsala.

Schulte im Walde, Sabine. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Sekine, Satoshi, Kentaro Inui, Ido Dagan, Bill Dolan, Danilo Giampiccolo, and Bernardo Magnini, editors. 2007. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Prague.

Shalizi, Cosma. 2003. Analogy and metaphor. Available at `http://masi.cscs.lsa.umich.edu/~crshalizi/notabene`.

Shinyama, Yusuke and Satoshi Sekine. 2003. Paraphrase acquisition for information extraction. In *Proceedings of the Second International Workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 65–71, Sapporo.

Shutova, Ekaterina. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of NAACL 2010*, pages 1029–1037, Los Angeles, CA.

Shutova, Ekaterina, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of COLING 2010*, pages 1,002–1,010, Beijing.

Shutova, Ekaterina and Simone Teufel. 2010. Metaphor corpus annotated for source–target domain mappings. In *Proceedings of LREC 2010*, pages 3,255–3,261, Malta.

Siegel, Sidney and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Book Company, New York, NY.

Sun, Lin and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of EMNLP 2009*, pages 638–647, Singapore.

Sun, Lin and Anna Korhonen. 2011. Hierarchical verb clustering using graph factorization. In *Proceedings of EMNLP*, pages 1,023–1,033, Edinburgh.

Toral, Antonio. 2009. The lexical substitution task at EVALITA 2009. In *Proceedings of EVALITA Workshop, 11th Congress of Italian Association for Artificial Intelligence*, Regio Emilia.

Tourangeau, Roger and Robert Sternberg. 1982. Understanding and appreciating metaphors. *Cognition*, 11:203–244.

Van de Cruys, Tim, Thierry Poibeau, and Anna Korhonen. 2011. Latent vector weighting for word meaning in context. In *Proceedings of EMNLP*, pages 1,012–1,022, Edinburgh.

van Rijsbergen, Keith. 1979. *Information Retrieval*, 2nd edition. Butterworths, London.

Veale, Tony and Yanfen Hao. 2008. A fluid knowledge representation for understanding and generating creative metaphors. In *Proceedings of COLING 2008*, pages 945–952, Manchester.

Wilks, Yorick. 1975. A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, 6:53–74.

Wilks, Yorick. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223.

Yu, K., S. Yu, and V. Tresp. 2006. Soft clustering on graphs. *NIPS*, pages 1553–1561, Vancouver.

Zapirain, Beñat, Eneko Agirre, and Lluís Màrquez. 2009. Generalizing over lexical features: selectional preferences for semantic role classification. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 73–76, Singapore.

Zhao, S., H. Wang, T. Liu, and S. Li. 2008. Pivot approach for extracting paraphrase patterns from bilingual corpora. In *Proceedings of ACL-08:HLT*, pages 780–788, Columbus, OH.

Zhao, Shiqi, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, ACL '09, pages 834–842, Suntec.

Zhou, Liang, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard H. Hovy. 2006. PARAEVAL: Using paraphrases to evaluate summaries automatically. In *Proceedings of HLT-NAACL*, pages 447–454, New York, NY.

Zong, Chengqing, Yujie Zhang, and Kazuhide Yamamoto. 2001. Approach to spoken Chinese paraphrasing based on feature extraction. In *Proceedings of NLPRS*, pages 551–556, Tokyo.