Design and Evaluation of Metaphor Processing Systems

Ekaterina Shutova*
University of California, Berkeley

System design and evaluation methodologies receive significant attention in natural language processing (NLP), with the systems typically being evaluated on a common task and against shared data sets. This enables direct system comparison and facilitates progress in the field. However, computational work on metaphor is considerably more fragmented than similar research efforts in other areas of NLP and semantics. Recent years have seen a growing interest in computational modeling of metaphor, with many new statistical techniques opening routes for improving system accuracy and robustness. However, the lack of a common task definition, shared data set, and evaluation strategy makes the methods hard to compare, and thus hampers our progress as a community in this area. The goal of this article is to review the system features and evaluation strategies that have been proposed for the metaphor processing task, and to analyze their benefits and downsides, with the aim of identifying the desired properties of metaphor processing systems and a set of requirements for their evaluation.

1. Introduction

Metaphor enriches our communication with a more diverse imagery and provides an important mechanism for reasoning about concepts. At the same time, it is also a very common linguistic device that has long become a part of our everyday language. Metaphors arise through systematic associations between distinct, and seemingly unrelated, concepts. For example, when we say "The *wheels* of Stalin's regime were *well-oiled* and already *turning*," we view a *political system* in terms of a *mechanism*, it can *function, break, have wheels*, and so forth. The existence of this association allows us to transfer knowledge and inferences from the domain of *mechanisms* to that of *political systems*. As a result, we reason about *political systems* in terms of *mechanisms* and discuss them using the *mechanism* terminology, giving rise to a variety of metaphorical expressions.

This view of metaphor is widely known as Conceptual Metaphor Theory (CMT). It was proposed by Lakoff and Johnson (1980), who claimed that metaphor is not merely a property of language, but rather a cognitive mechanism that structures our conceptual system in a certain way. Lakoff and Johnson explained metaphor through a presence of a mapping between two domains of experience: the target (e.g., *politics*) and the source (e.g., *mechanism*). Metaphor is thus not limited to meaning extensions of individual

* International Computer Science Institute, 1947 Center St. Ste. 600, Berkeley, CA 94704, USA.
E-mail: katia@icsi.berkeley.edu.

Submission received: 9 October 2013; revised version received: 4 March 2015; accepted for publication: 10 June 2015.

doi:10.1162/COLI_a_00233

words, but rather involves a complex cross-domain knowledge projection process. Let us consider a few more examples.

(1) "President Obama is rebuilding the campaign *machinery* that *vaulted* him into office" (*New York Times* 2011)

(2) 20 steps towards a modern, *working* democracy

(3) Time to *mend* our foreign policy.

(4) "She knows the *nuts* and *bolts*, and it's not the *nuts* and *bolts* inside legislation, it's the *nuts* and *bolts* of raising money, preparing the party for elections, a political consultant kind of politics." (Bzdek 2008)

These examples demonstrate how multiple properties and inferences from the domain of *mechanisms* are systematically projected onto our knowledge about politics. Lakoff and Johnson coined the term **conceptual metaphor** to describe such mappings from the source domain to the target. The view of an inter-conceptual mapping as a basis of metaphor was echoed by other prominent theories in the field. These include, most notably, the comparison view, formulated in the Structure–Mapping Theory of Gentner (1983), and the interaction view (Black 1962; Hesse 1966). However, the principles of CMT have inspired and influenced much of the computational work on metaphor, thus becoming more central to this paper. Conceptual metaphor manifests itself in language in the form of **linguistic metaphor**, or metaphorical expressions. These in turn include *lexical* metaphor, that is, single-word meaning extensions (as in Examples (2) and (3)); *multi-word* metaphorical expressions (e.g., "the government *turned a blind eye to* corruption"); or *extended* metaphor, that spans longer discourse fragments.

Manifestations of metaphor are frequent in language, appearing on average in every third sentence of general-domain text, according to corpus studies (Cameron 2003; Martin 2006; Shutova and Teufel 2010; Steen et al. 2010). This makes metaphor an important subject of linguistic research and makes its accurate processing essential for a range of practical NLP applications. These include, for example, (1) machine translation (MT): Because a large number of metaphorical expressions are culture-specific, they represent a considerable challenge for MT (e.g., the English metaphor "to *shoot down* someone's arguments" cannot be literally translated into German as "Argumente *abschießen*" and metaphor interpretation is required); (2) opinion mining: Metaphorical expressions tend to contain a strong emotional component—for example, compare the metaphorical expression "Government *loosened stranglehold* on business" and its literal counterpart *Government deregulated business* (Narayanan 1999); (3) information retrieval (IR): Non-literal language without appropriate disambiguation may lead to false positives in information retrieval (e.g., documents describing "*old school* gentlemen" should not be returned for the query *school* [Korkontzelos et al. 2013]); and many others.

Because metaphor interpretation requires complex analogical comparisons and the projection of inference structures across domains, the task of automatic metaphor processing is challenging. For many years, computational work on metaphor evolved around the use of hand-coded knowledge and rules to model metaphorical associations, making the systems hard to scale. Recent years have seen a growing interest in statistical modeling of metaphor (Mason 2004; Gedigian et al. 2006; Shutova 2010; Shutova, Sun, and Korhonen 2010; Turney et al. 2011; Dunn 2013a; Heintz et al. 2013; Hovy et al. 2013; Li, Zhu, and Wang 2013; Mohler et al. 2013; Shutova and Sun 2013;

Strzalkowski et al. 2013; Tsvetkov, Mukomel, and Gershman 2013), with many new techniques opening routes for improving system accuracy and robustness. A wide range of methods have been proposed and investigated by the community, including supervised (Gedigian et al. 2006; Dunn 2013a; Hovy et al. 2013; Mohler et al. 2013; Tsvetkov, Mukomel, and Gershman 2013) and unsupervised (Heintz et al. 2013; Shutova and Sun 2013) learning, distributional approaches (Shutova 2010, 2013; Shutova, Van de Cruys, and Korhonen 2012), lexical resource-based methods (Krishnakumaran and Zhu 2007; Wilks et al. 2013), psycholinguistic features (Turney et al. 2011; Gandy et al. 2013; Neuman et al. 2013; Strzalkowski et al. 2013), and Web search (Veale and Hao 2008; Bollegala and Shutova 2013; Li, Zhu, and Wang 2013). Although individual approaches tackling individual aspects of metaphor have met with success, the insights gained from these experiments are still difficult to integrate into a single computational metaphor modeling landscape, because of the lack of a unified task definition, a shared data set, and well-defined evaluation standards. This hampers our progress as a community in this area. In this paper we take a step towards closing this gap: We review the recent work on computational modeling of metaphor, the tasks addressed, the system features proposed, and the evaluations conducted, and analyze the relevance of different linguistic aspects of metaphor for system performance and applicability, with the aim of identifying the desired properties of metaphor processing systems and a set of requirements for their evaluation.

2. Considerations in the Design of a Metaphor Processing System

When designing a metaphor processing system, one faces a number of choices. Some stem from the linguistic and cognitive properties of metaphor, others concern the applicability and usefulness of the system in the wider NLP context. In this section, we analyze individual aspects of metaphor and their relevance to computational modeling, as well as their interplay in the design of a real-world system.

2.1 Linguistic Considerations and Levels of Analysis

Linguistic considerations that inform the design of metaphor processing systems concern primarily the choice of the level (or levels) of analysis. The levels of metaphor analysis include (1) linguistic metaphor (or metaphorical expressions), (2) conceptual metaphor, (3) extended metaphor, and (4) metaphorical inference. Let us consider an example of manifestations of the conceptual metaphor EUROPEAN INTEGRATION as a TRAIN JOURNEY, popular in the early 1990s, at various levels.

- **Conceptual:** EUROPEAN INTEGRATION as a TRAIN JOURNEY
- **Linguistic:** The *coupling of the carriages* may not be reliably secure, but the pan-European *express is in motion*.
- **Extended metaphor:** "There is a fear that the European train will thunder forward, laden with its customary cargo of gravy, towards a destination neither wished for nor understood by electorates. But the train can be stopped." (Margaret Thatcher, *Sunday Times*, 20 Sept 1992)
- **Metaphorical inference:** The fact that *expensive tracks have to be laid for the train to move forward* means that someone has to fund the process of European integration.

2.1.1 Linguistic metaphor. Linguistic metaphor, or metaphorical expressions, concern the surface realization of metaphorical mechanisms, and have been unsurprisingly central to metaphor processing research to date (Birke and Sarkar 2006; Gedigian et al. 2006; Krishnakumaran and Zhu 2007; Shutova, Sun, and Korhonen 2010; Turney et al. 2011; Dunn 2013a; Gandy et al. 2013; Heintz et al. 2013; Hovy et al. 2013; Neuman et al. 2013; Shutova 2013; Strzalkowski et al. 2013; Tsvetkov, Mukomel, and Gershman 2013). Metaphorical expressions represent the way in which text-processing systems encounter metaphor, and there is little doubt that any real-world metaphor processing system, whatever the approach or the application, ultimately needs to be able to identify and interpret them.

When focusing on linguistic metaphor, one needs to further take into account the level of conventionality of the expressions; how different syntactic constructions are used to convey metaphorical meanings and at what level metaphor annotation needs to be done (word, relation, or sentence level). Thus the following considerations become important for the task and system design:

- **Level of conventionality** of the metaphors accepted: Metaphor is a productive phenomenon, that is, its novel examples continue to emerge in language. However, a large number of metaphorical expressions become conventionalized over time (e.g., "I cannot *grasp* his way of thinking"). Although metaphorical in nature, their meanings are deeply entrenched in everyday use, and their comprehension is likened to that of literally used terms (Nunberg 1987). According to Gibbs (1984), metaphorical expressions are spread along a continuum from highly conventional, lexicalized metaphors to entirely novel and creative ones. Gibbs thus suggests that there is no clear demarcation line between literal and metaphorical language, and the distinction between them is rather governed by the level of conventionality of metaphorical expressions. From the usage perspective, metaphoricity may be viewed as a gradient phenomenon rather than a binary one (Dunn 2011), and conventionality becomes an important factor in the design and evaluation of metaphor processing systems. It is not yet clear where on the metaphorical–literal continuum the system should draw the line between what it considers metaphorical and what it considers literal. The answer to this question most likely depends on the NLP application in mind. However, generally speaking, real-world NLP applications are unlikely to be concerned with historical aspects of metaphor, but rather with the identification of figurative language that needs to be interpreted differently from the literal language. We therefore suggest that NLP applications do not necessarily need to address highly conventional and lexicalized metaphors that can be interpreted using standard word sense disambiguation techniques, but rather would benefit from the identification of less conventional and more creative language. Much of the metaphor processing work has focused on conventional metaphor, though in principle capable of identifying novel metaphor as well (Shutova, Sun, and Korhonen 2010; Turney et al. 2011; Dunn 2013a; Gandy et al. 2013; Heintz et al. 2013; Neuman et al. 2013; Shutova 2013; Strzalkowski et al. 2013; Tsvetkov, Mukomel, and Gershman 2013), with few approaches modeling only novel metaphor (Desalle, Gaume, and Duvignau 2009) or discriminating between conventional and novel metaphors (Krishnakumaran and Zhu 2007). Other approaches have

looked at literal versus non-literal distinction defined more broadly (Birke and Sarkar 2006; Li and Sporleder 2009, 2010).

- **Syntactic constructions covered**: Metaphors vary with respect to how they are expressed in language grammatically. The grammatical structure of metaphorical expressions is tightly coupled with the inference processes that produce them and are guided by the semantic frames they evoke (Sullivan 2007, 2013). According to Sullivan, conceptual metaphors are realized in language via mapping semantic roles in the source frame onto roles in the target frame. This suggests that both the source and the target domain impose a set of constraints on the roles that can be mapped, and thus on the syntactic options for expressing the metaphorical meaning. There has not yet been a computational approach investigating the interplay of frame semantics and surface structure of metaphorical language. However, the community has addressed modeling linguistic metaphor in a range of syntactic constructions. Verbal or adjectival metaphors are particularly widely embraced by NLP researchers (most approaches), with a few works focusing on copula constructions (Krishnakumaran and Zhu 2007; Gandy et al. 2013; Li, Zhu, and Wang 2013; Neuman et al. 2013) or other nominal metaphors (Heintz et al. 2013; Hovy et al. 2013; Li, Zhu, and Wang 2013; Mohler et al. 2013; Strzalkowski et al. 2013). A number of approaches to metaphor identification have also addressed multiword metaphors (Li and Sporleder 2010; Heintz et al. 2013; Hovy et al. 2013; Mohler et al. 2013). Corpus-linguistic research has shown that verbs and adjectives account for a large proportion of metaphorical expressions observed in the data (Cameron 2003; Shutova and Teufel 2010). However, a recent study (Jamrozik et al. 2013) has also shown that relational words tend to have a higher metaphorical potential. This corresponds to the data on verbal metaphor frequency, and it also suggests that relational nouns are important (e.g., "words are *friends of* translators").
- **Lexical, relation, or sentence level**: Finally, one needs to decide if metaphor should be annotated at the word level (i.e., tagging the source domain words alone), relation level (i.e., tagging both source and target words in a particular grammatical relation), or sentence level (i.e., tagging sentences that contain metaphorical language, without explicit annotation of source and target domain words). The word or relation levels provide the most information and have been the focus of the majority of approaches (Gedigian et al. 2006; Shutova, Sun, and Korhonen 2010; Turney et al. 2011; Gandy et al. 2013; Heintz et al. 2013; Hovy et al. 2013; Neuman et al. 2013; Shutova 2013; Shutova and Sun 2013; Wilks et al. 2013). However, some works annotated metaphor at the sentence level (Krishnakumaran and Zhu 2007; Dunn 2013a; Li, Zhu, and Wang 2013; Mohler et al. 2013; Strzalkowski et al. 2013; Tsvetkov, Mukomel, and Gershman 2013).

2.1.2 Conceptual metaphor. Conceptual metaphor represents a cognitive and conceptual mechanism by which humans produce and comprehend metaphorical expressions. Manifestations of conceptual metaphor are ubiquitous in language, communication, and even decision-making (Thibodeau and Boroditsky 2011). Here are a few 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"), 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*"); ARGUMENT IS A WAR (e.g., "He *shot down* all of my arguments," "He *attacked* every weak point in my argument").

On one hand, few would disagree that the metaphor processing systems capable of understanding and applying conceptual metaphor should be in a better position to accurately handle linguistic metaphors as well. However, a series of questions arise when designing a model of conceptual metaphor. For example, how does one represent conceptual metaphors in the system? What labels does one assign to source and target domains? Is it even possible to name all the conceptual metaphors that humans use and is it necessary to do so? A computational model needs a clear definition of what constitutes the source and target domains (whether they are manually listed or automatically learned) and the consistency and coverage of source and target domain categories would play a crucial role in how well the model can account for real-world data. Previous research on annotation of conceptual metaphor (Shutova and Teufel 2010) has shown that the annotators tend to disagree on the assignment of source and target domain categories. The most variation stems from the level of generality of the selected categories, indicating that while cross-domain mappings are intuitive to humans (i.e., they can be annotated in arbitrary text in principle), labeling source and target domains consistently appears to be a challenging task. What this suggests is that, although the mechanism of conceptual metaphor may be helpful to the system, the question of how to best represent source and target domains within the system remains open. A predefined set of categories, such as those widely discussed in the linguistic literature on CMT, may not be sufficient or even suitable for a computational model. And despite the validity of the main principles of CMT as a linguistic theory, it is not straightforward to port it to computational modeling of metaphor. A more flexible, and potentially data-driven, representation of source and target domain categories is needed for the latter purpose. A data-driven representation would also be better suited to account for the freedom of interpretation that some metaphors allow, since more flexible structures can be dynamically learned from the data. So far, the community has attempted assigning manually created labels to metaphorical mappings (Mason 2004; Baumer, Tomlinson, and Richland 2009), harvesting fine-grained mappings between individual nouns (Li, Zhu, and Wang 2013), using lexical resources to define or expand source and target domain categories (Gandy et al. 2013; Mohler et al. 2013), representing source and target concepts as word clusters (Shutova and Sun 2013) or automatically learned topics (Heintz et al. 2013), and learning metaphorical mappings implicitly within the model without explicit labeling (Shutova, Sun, and Korhonen 2010).

2.1.3 Extended metaphor. Extended metaphor refers to the use of metaphor at the discourse level. It manifests itself in discourse via a sequence of metaphorically used language, yielded by the same conceptual metaphor, whereby a continuous scenario from the source domain is metaphorically projected onto the target. For instance, viewing *European integration* as a *train journey* led to numerous metaphorical expressions in political discourse. Each of them mapped certain properties of *train journeys* to political processes—with countries as *loosely connected carriages*, peoples of different countries as *passengers of their respective carriages*, *expensive tracks that have to be laid for the train to move forward*, and the *final destination* not very well understood. This metaphor has been dominating the debate, with the European leaders arguing over its details

(Beigman Klebanov and Beigman 2010). One can see this metaphor frequently reappear and evolve over time, helping politicians defend their agendas and, not least, shedding some clarity on an otherwise uncertain future.

Research in linguistics and political science (Musolff 2000; Lakoff 2008; Lakoff and Wehling 2012) suggests that the use of a particular metaphor often guides the speakers' argumentation strategy throughout a piece of discourse, as well as participants' behavior in a dialogue. Beigman Klebanov and Beigman (2010) investigated extended metaphor within a game-theoretic framework, demonstrating that maintaining the metaphorical frame in a debate is rationalizable in terms of the gains the participants may get from doing so and their potential losses from swerving away from the metaphor. Their approach reverse-engineers the motivations behind the use of extended metaphor within a formal framework. Beigman Klebanov and Beigman's work was an important advance, enhancing our understanding of the inner workings of extended metaphor, motivation behind its use, and its effects on social dynamics. However, a computational method for identification and interpretation of extended metaphor in real-world discourse is yet to be proposed. A discourse-level metaphor processing system would need to identify a chain of metaphorical expressions in a text, which indicates a systematic association of the text topic with a particular domain. These chains would then demonstrate how continuous scenarios can be transferred across domains. Recovering this information from the data would allow us to better understand the structure behind metaphorical associations, as well as the inferential process by which knowledge is projected across domains. This system would also find application in social science, where metaphorical framing is widely studied as an indicator of the underlying cultural and moral models (Lakoff and Wehling 2012).

2.1.4 Metaphorical inference. When projecting knowledge from one domain to another, a set of complex inferences take place. Metaphorical inferences are grounded in the source domain and result in the production of surface structures we observe in language as metaphorical expressions. Metaphorical mappings are thus realized via projecting inference structures from the source domain onto the target. For example, when European integration is metaphorically viewed as a train journey, our knowledge of typical events and their consequences from the domain of train journey are projected onto our reasoning about the process of European integration. For instance, if we know that expensive tracks need to be laid before a train can move forward, we can infer that someone also needs to fund the process of European integration. Interestingly, in the presence of a conceptual metaphor such inference can take place even without any linguistic metaphor referring to the tracks being present, but rather on the basis of our common sense knowledge about the functioning of trains.

Besides allowing us to derive new information about the target domain, projecting the inferential structures from the source domain also invokes an emotional response coming from the source domain—for example, an unknown destination or a great expense make one feel uneasy when referred to both literally and metaphorically. Such a transfer of inferential processes and emotional content is believed by some to be one of the central purposes of metaphor (Kovecses 2005; Feldman 2006; Thibodeau and Boroditsky 2011). Psychologists Thibodeau and Boroditsky (2011) investigated how metaphor and metaphorical inference affect decision-making. Their hypothesis was that metaphors in language "instantiate frame-consistent knowledge structures [from the source domain] and invite structurally consistent inferences [about the target domain]." They used two groups of subjects, who were presented with two different texts about *crime*. In the first text crime was metaphorically portrayed as a VIRUS and

in the second as a BEAST. The two groups were then asked a set of questions on how to tackle crime in the city. It turned out that, whereas the first group tended to opt for preventive measures in tackling crime (e.g., stronger social policies), the second group converged on punishment- or restraint-oriented measures. According to Thibodeau and Boroditsky, their results demonstrate that metaphors have profound influence on how we conceptualize and act with respect to societal issues. Interestingly, the participants would explain their decisions via arguments unrelated to the metaphor, showing that the effect of metaphorical inference in guiding human reasoning is rather covert.

Being able to reproduce these inferences is likely to make automatic metaphor understanding better informed and hence more accurate. It may also provide a mechanism of representing source–target domain mappings, that themselves are generalizations over a set of inferences transferred from one domain to another. And finally, these inferences provide a platform for metaphor interpretation, namely, deriving the meaning of a metaphorical expression and the additional connotations it introduces (that are likely to originate from the source domain). There is a consensus among cognitive linguists that it is metaphorical inference that provides for the very texture behind the use of metaphor (Hobbs 1981; Carbonell 1982; Rohrer 1997; Turner and Fauconnier 2003; Feldman 2006). Although uncovering this texture is certainly one of the main objectives of computational metaphor understanding, it is at the same time a very challenging undertaking. Reproducing metaphorical inferences would require the ability to learn vast amounts of world knowledge from the data, as well as performing complex cross-domain comparisons. And despite being a very promising route, it has not yet been attempted in NLP.

2.2 Applicability

Another set of considerations in the design of metaphor processing systems stems from the needs of real-world NLP. The high frequency of metaphorical language in textual data makes accurate metaphor processing desired for a number of NLP applications. Thus, the format of metaphor understanding that metaphor processing systems provide should ideally be informed by the requirements of external NLP and a number of considerations arise in that respect:

- Metaphor processing typically involves two tasks: **metaphor identification** (distinguishing between literal and metaphorical language in text) and **metaphor interpretation** (identifying the intended meaning of the metaphor). Both of these provide useful information for language understanding and need to be addressed, either independently or as a single process.
- A metaphor processing system should 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 textual output, in which metaphors are interpreted.
- In order to be useful for real-world NLP, the system needs to be capable of processing real-world data, and thus **operate on unrestricted, continuous text**. Rather than only dealing with individual carefully selected clear-cut

examples, the system should be fully implemented and tested on free, naturally occurring text.

- To enable wide applicability, the system needs to **be open-domain**—that is, operate in all domains, genres, and topics. Therefore, ideally it should not rely on any domain-specific information or focus on individual types of instances (e.g., a limited set of hand-chosen source-target domain mappings).
- To be easily adaptable to new domains, the system should **not rely on task-specific hand-coded knowledge**. This means it needs to be either data-driven and be able to automatically acquire the knowledge it needs from text corpora, or rely only on large-scale, general-domain lexical resources (that are already in existence and do not need to be created in a costly manner). However, it would be an advantage if no such resource is required and the system can dynamically induce meanings in context.
- To be robust, the system needs to be able to deal with metaphors represented by **all word classes and syntactic constructions**. Many existing models are designed with specific kinds of metaphorical expressions in mind, for instance nominal metaphors in copula constructions or verbal metaphors in verb–object relations. To be applicable to support real-world NLP applications, these models need to be extended beyond those specific word classes and syntactic constructions, and be able to process any kind of metaphorical language.

When designing a metaphor processing task, methodology, and evaluation strategy, one thus needs to keep these criteria in mind. Although modeling all of the phenomena described in this section within a single system is by no means a requirement, it is critically important to be aware of all the guises that metaphor may take, both conceptually and empirically.

3. Metaphor Annotation and Resources

3.1 Corpora

Metaphor annotation studies have typically focused on one (or both) of the following tasks: (1) identification of metaphorical senses in text (i.e., distinguishing between literal and non-literal meanings), and (2) assignment of the corresponding source–target domain mappings. The majority of corpus-linguistic studies were concerned with metaphorical expressions and mappings within a limited domain—for example, WAR, BUSINESS, FOOD, or PLANT metaphors (Santa Ana 1999; Izwaini 2003; Koller 2004; Skorczynska Sznajder and Pique-Angordans 2004; Chung, Ahrens, and Huang 2005; Hardie et al. 2007; Gong, Ahrens, and Huang 2008; Lu and Ahrens 2008; Low et al. 2010), in a particular genre or type of discourse (Charteris-Black 2000; Cameron 2003; Izwaini 2003; Koller 2004; Skorczynska Sznajder and Pique-Angordans 2004; Martin 2006; Hardie et al. 2007; Lu and Ahrens 2008; Beigman Klebanov and Flor 2013), or in individual examples in isolation from wider context (Wikberg 2006; Lönneker-Rodman 2008). In addition, these approaches often focused on a small predefined set of source and target domains. Another vein of corpus-based research concerned cross-linguistic differences in the use of metaphor, also in a specific domain—for example,

financial discourse (Charteris-Black and Ennis 2001), metaphors describing FEELINGS (Stefanowitsch 2004; Diaz-Vera and Caballero 2013), or metaphorical expressions referring to body parts (Deignan and Potter 2004). Three recent studies are notable in that they moved away from investigating particular domains to a more general study of how metaphor behaves in unrestricted continuous text. Wallington et al. (2003), Shutova and Teufel (2010), and Steen et al. (2010) conducted consecutive metaphor annotation in open-domain texts.

Wallington et al. (2003) used two teams of annotators and compared externally prescribed definitions of metaphor with intuitive internal ones. Team A was asked to annotate "interesting stretches," whereby a phrase was considered interesting if (1) its significance in the document was non-physical, (2) it could have a physical significance in another context with a similar syntactic frame, and (3) this physical significance was related to the abstract one. Team B had to annotate phrases according to their own intuitive definition of metaphor. Apart from metaphorical expressions, the respective source–target domain mappings were also to be annotated. For this latter task, the annotators were given a set of mappings from the Master Metaphor List and were asked to assign the most suitable ones. However, the authors do not report the level of interannotator agreement, nor the coverage of the mappings in the Master Metaphor List on their data. The fact that the method is limited to a set of mappings exemplified in the Master Metaphor List suggests that it may not scale well to real-world data, because the predefined inventory of mappings is unlikely to be sufficient to cover the majority of metaphorical expressions in arbitrary text.

Steen and his colleagues (Pragglejaz Group 2007; Steen et al. 2010) proposed a metaphor identification procedure (MIP). In the framework of this procedure, the sense of every word in the text is considered as a potential metaphor. Every word is then tagged as literal or metaphorical, based on whether is has a "more basic, contemporary meaning" in other contexts than the current one. The summary of their annotation procedure is presented is Figure 1. In a sense, such annotation can be viewed as a form of word sense disambiguation with an emphasis on metaphoricity. Steen and colleagues ran a reliability study involving near-native speaker annotators (strongly relying on dictionary definitions) and report an interannotator agreement of 0.85 in terms of Fleiss' kappa. MIP laid the basis for the creation of the VU Amsterdam Metaphor Corpus[1] (Steen et al. 2010). This corpus is a subset of BNC Baby[2] annotated for linguistic metaphor. Its size is 200,000 words and it comprises four genres: news text, academic text, fiction, and conversations. The corpus has already found application in computational metaphor processing research (Dunn 2013b; Niculae and Yaneva 2013), as well as inspiring metaphor annotation efforts in other languages (Badryzlova et al. 2013).

The study of Shutova and Teufel (2010) was concerned with annotation of both metaphorical expressions and metaphorical mappings in continuous text. Their annotation procedure is based on MIP, modifying and extending it to the identification of conceptual metaphors along with the linguistic ones. Following MIP, the annotators were asked to identify the more basic sense of the word, and then label the context in which the word occurs in the basic sense as the source domain, and the current context as the target. They were provided with a list of suggested common source

1 http://www.ota.ox.ac.uk/headers/2541.xml.

2 BNC Baby is a 4-million-word subset of the British National Corpus (BNC) (Burnard 2007), comprising four different genres: academic, fiction, newspaper, and conversation. For more information, see http://www.natcorp.ox.ac.uk/corpus/babyinfo.html.

1. Read the entire text-discourse to establish a general understanding of the meaning.
2. Determine the lexical units in the text-discourse.
3.
   - For each lexical unit in the text, establish its meaning in context, that is, how it applies to an entity, relation, or attribute in the situation evoked by the text (contextual meaning). Take into account what comes before and after the lexical unit.
   - For each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be
     - More concrete [what they evoke is easier to imagine, see, hear, feel, smell, and taste];
     - Related to bodily action;
     - More precise (as opposed to vague);
     - Historically older;

     Basic meanings are not necessarily the most frequent meanings of the lexical unit.
   - If the lexical unit has a more basic current contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.
4. If yes, mark the lexical unit as metaphorical.

Figure 1
Metaphor identification procedure of the Pragglejaz Group (2007).

and target domains, but were also allowed to introduce domains of their own to match their intuitions. Shutova and Teufel's corpus is a subset of the BNC sampling various genres: fiction, newspaper/journal articles, essays on politics, international relations and sociology, and radio broadcast (transcribed speech). The size of the corpus is 13,642 words, containing 241 metaphorical expressions in total. Table 1 shows the breakdown of metaphors by type, as well as their variation across genres. Shutova and colleagues used the corpus data as a testbed in a number of computational experiments (Shutova 2010; Shutova, Van de Cruys, and Korhonen 2012; Shutova 2013; Shutova, Teufel, and Korhonen 2013).

Table 1
Corpus statistics from Shutova and Teufel (2010).

Text	ID	Genre	Sent.	Words	Met–rs	Met./Sent.	Verb m.
Hand in Glove, Goddard	G0N	Fiction	335	3,927	41	0.12	30
After Gorbachev, White	FYT	Politics	45	1,384	23	0.51	17
Today newspaper	CEK	News	116	2,086	48	0.41	30
Tortoise by Candlelight, Bawden	HH9	Fiction	79	1,366	12	0.15	10
The Masks of Death, Cecil	ACA	Sociology	60	1,566	70	1.17	42
Radio broadcast (current affairs)	HM5	Speech	58	1,828	10	0.17	7
Language and Literature journal	J85	Article	68	1,485	37	0.54	28
Total			761	13,642	241	0.32	164

3.2 Metaphor Lists and Databases

Lakoff and colleagues organized their ideas in a resource called the Master Metaphor List (MML) (Lakoff, Espenson, and Schwartz 1991). The list is a collection of source–target domain mappings (mainly those related to mind, feelings, and emotions) with corresponding examples of language use. The mappings in the list are organized in an ontology—for example, the metaphor PURPOSES ARE DESTINATIONS is a special case of a more general metaphor STATES ARE LOCATIONS. The resource has been criticized for the lack of clear structuring principles of the mapping ontology (Lönneker-Rodman 2008). However, to date MML is the most comprehensive resource for conceptual metaphor in the linguistic literature, and the examples from the list have been used by computational approaches (Mason 2004; Krishnakumaran and Zhu 2007; Li, Zhu, and Wang 2013), both for development and evaluation purposes. The MML also inspired the creation of other resources, including resources in multiple languages that could facilitate cross-linguistic research on metaphor. One such example is the Hamburg Metaphor Database (Lönneker 2004; Reining and Lönneker-Rodman 2007), which contains examples of metaphorical expressions in German and French. The expressions are mapped to senses from EuroWordNet[3] and annotated with source–target domain mappings taken from the MML.

4. Metaphor Identification Systems

Early approaches to metaphor relied on information in handcrafted knowledge bases, followed by metaphor identification in and with the help of lexical resources. Recent years have witnessed a growing interest in statistical and machine learning approaches to metaphor identification. As the field of computational semantics—in particular, robust parsing and lexical acquisition techniques—have progressed to the point where it is possible to accurately acquire lexical, domain, and relational information from corpora, this opened many new avenues for large-scale statistical metaphor identification. The vast majority of systems identify metaphor at the linguistic level (Birke and Sarkar 2006; Gedigian et al. 2006; Krishnakumaran and Zhu 2007; Shutova, Sun, and Korhonen 2010; Turney et al. 2011; Dunn 2013a; Heintz et al. 2013; Hovy et al. 2013; Neuman et al. 2013; Shutova 2013; Strzalkowski et al. 2013; Tsvetkov, Mukomel, and Gershman 2013), with very few focusing on the conceptual level (Mason 2004; Baumer, Tomlinson, and Richland 2009) or identifying both (Gandy et al. 2013; Li, Zhu, and Wang 2013; Shutova and Sun 2013). This section will first present computational approaches to linguistic metaphor identification, then move on to conceptual metaphor.

4.1 Identification of Linguistic Metaphors

4.1.1 Approaches Using Hand-Coded Knowledge and Lexical Resources. One of the first approaches to identify and interpret metaphorical expressions in text was proposed by Fass (1991) in his met* system. This system relies on the hypothesis that metaphors often represent a violation of selectional preferences in a given context (Wilks 1975,

3 EuroWordNet is a multilingual database with 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. http://www.illc.uva.nl/EuroWordNet/.

1978). Selectional preferences are the semantic constraints that a predicate places onto its arguments. Consider the following metaphorical expression.

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

The verb *drink* normally requires a grammatical subject of type ANIMATE and a grammatical object of type LIQUID. Therefore, *drink* taking a *car* as a subject in Example (5) is an anomaly, which, according to Wilks, indicates a metaphorical use of *drink*. met* detects non-literalness via selectional preference violation, utilizing handcrafted descriptions of selectional preferences. In case of a violation, the respective phrase is first tested for being metonymic, using hand-coded patterns (e.g., CONTAINER-FOR-CONTENT). If this fails, the system searches the knowledge base for a relevant analogy in order to discriminate metaphorical relations from the anomalous ones. For example, the sentence "My car *drinks* gasoline" 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. Fass (1991) presented the approach itself, but reported no evaluation results.

More recently, Wilks et al. (2013) revisited this idea, acquiring selectional preferences from lexical resources, namely VerbNet and WordNet. They focused on conventionalized metaphors included in lexical resources and proposed a technique for their automatic identification. They see this work as complementary to the approaches that perform data-driven learning of selectional preferences. The latter, according to the authors, is likely to miss conventional metaphors because of their widespread presence in the data. Wilks and colleagues expect that selectional preferences acquired from term definitions in lexical resources would circumvent this issue and enable them to efficiently detect highly conventionalized metaphors. The main hypothesis behind their approach is that if the first (main) WordNet sense of a word does not satisfy the preferences of its context in a given sentence, but has a lower (less frequent) sense in WordNet that satisfies the preference, then that use of the word and that WordNet sense are likely to be metaphorical. For instance, in the example "Mary married a *brick*", the first sense of *brick* is 'a physical object,' thus violating the preference of *marry* that selects for *people*, but the second sense of *brick* as 'a reliable person' satisfies this preference. To implement this approach, Wilks and colleagues acquire typical preferences of concepts (i.e., word senses) from WordNet glosses. They use a semantic parser (Allen, Swift, and de Beaumont 2008) to identify the nominal arguments of the verbs in glosses and their semantic roles and then abstract to their higher-level hypernyms in WordNet, which define the preferences. They compared the performance of their system to a baseline using hand-coded verb preferences in VerbNet. The evaluation was carried out on a set of 122 sentences from the domain of Governance, manually annotated for metaphoricity and selected so that the data set contains 50% metaphorical instances and 50% literal ones. They report an F-score of 0.49 for the VerbNet-based system and 0.67 for the WordNet-based one, the latter showing higher recall and the former higher precision. The approach of Wilks et al. rests on the assumption that WordNet sense ranking corresponds somewhat to the literal-to-metaphorical scale, as well as the assumption that there is only one literal sense for the given word. Although this may be true for the majority of senses, it is relatively easy to find counter-examples. For instance, the first WordNet sense of the verb *erase* is metaphorical, defined as "remove from memory or existence, e.g., The Turks *erased* the Armenians in 1915," with the literal sense ranked

second. The reliance on WordNet sense numbering is thus a limitation of the presented approach. Another issue (that the authors point out themselves) is that this approach is likely to detect metonymic uses along with metaphor, and a method to discriminate between the two is still needed.

The method of Krishnakumaran and Zhu (2007) uses hyponymy relation in WordNet and word bigram counts to annotate metaphor at the sentence level. Given an IS-A metaphor (e.g., *The world is a stage*[4]) they first verify if the two nouns involved are in hyponymy relation in WordNet, and if this is not the case then the sentence is tagged as containing a metaphor. Along with this they consider expressions containing a verb or an adjective used metaphorically (e.g., "He *planted* good ideas in their minds" or "He has a *fertile* imagination"). In such 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 containing it as metaphorical. This idea follows the intuition of Wilks. However, by using bigram counts over verb–noun pairs Krishnakumaran and Zhu (2007) lose a great deal of information compared with a system extracting selectional preferences for specific grammatical relations from parsed text. The authors evaluated their system on a set of example sentences compiled from the MML, whereby highly conventionalized metaphors (or dead metaphors) are taken to be negative examples, reporting an accuracy of 0.58. Thus Krishnakumaran and Zhu do not deal with literal examples as such: Essentially, the distinction they are making is between the senses included in WordNet, even if they are conventional metaphors, and those not included in WordNet.

4.1.2 Statistical Learning for Metaphor Identification. The first statistical approach to metaphor is the TroFi system (Trope Finder) of Birke and Sarkar (2006). Their method is based on sentence clustering, originating from a similarity-based word sense disambiguation method developed by Karov and Edelman (1998). The method uses a set of seed sentences, where the senses are annotated, 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 two-way classification: literal versus non-literal, and they do not clearly define the kinds of tropes they aim to discover. They evaluated their system on a set of 25 verbs (such as *absorb, die, touch, knock, strike, pour,* etc.), for each of which they extracted a set of sentences containing its literal and figurative uses, 1,298 in total, from the *Wall Street Journal* corpus. An example for the verb *pour* in their data set is shown in Figure 2. Two annotators annotated the sentences for literalness, achieving an agreement of $\kappa = 0.77$. The authors report a system performance of 53.8% in terms of F-score on this data set.

The metaphor identification system of Shutova, Sun, and Korhonen (2010) also uses clustering techniques, but performs word clustering to discover verb–subject and verb–object metaphors in unrestricted text. It starts from a small seed set of metaphorical expressions, learns the analogies involved in their production, and extends the set of analogies by means of verb and noun clustering. The method is based on the hypothesis of "clustering by association"—namely, that in the course of distributional noun clustering, abstract concepts tend to cluster together if they are associated with the same source domain, whereas concrete concepts cluster by meaning similarity. For

4 William Shakespeare.

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).

instance, *democracy* and *marriage* get clustered together, because both are associated with *mechanisms*, and as such appear with the *mechanism* terminology in the corpus. This allows the system to discover new, previously unseen conceptual and linguistic metaphors—for example, having seen the seed metaphor "*mend* marriage" it infers that "the *functioning* of democracy" is also used metaphorically. This is how the system expands from the seed set to new concepts. Shutova, Sun and Korhonen used a spectral clustering algorithm with lexico-syntactic features to cluster verbs and nouns. They applied their system to continuous text (the whole BNC) and evaluated its performance on a random sample of the extracted metaphors against human judgments. They report a precision of 0.79 with an inter-judge agreement of $k = 0.63$ among five annotators. Their data-driven system favorably compares to a WordNet-based baseline, where synsets are used in place of automatically derived clusters. Shutova and colleagues have shown that the clustering-based solution has a significantly wider coverage, capturing new metaphors rather than the synonymous ones, as well as yielding a 35% increase in precision. However, Shutova, Sun and Korhonen did not evaluate the recall of their system, which is likely to be dependent on the size of the seed set and a relatively large and representative seed set is needed to achieve full coverage.

Turney et al. (2011) classify verbs and adjectives as literal or metaphorical based on their level of concreteness or abstractness in relation to the noun they appear with. They learn concreteness rankings for words automatically (starting from a set of examples) and then search for expressions where a concrete adjective or verb is used with an abstract noun (e.g., "*dark* humor" is tagged as a metaphor and "dark hair" is not). They used the data set of Birke and Sarkar (2006) for evaluation of verb metaphors and attain an F-score of 0.68, which favorably compares to that of Birke and Sarkar. For adjectives, they have created their own data set of selected individual adjective–noun pairs for five adjectives: *dark*, *deep*, *hard*, *sweet*, and *warm*; 100 phrases in total. These were then manually annotated for metaphoricity. As compared to these annotations, the accuracy of adjective classification is 0.79. However, the adjective data set was constructed with the concreteness feature in mind, and therefore the results reported for verb metaphors are likely to be more objective.

Neuman et al. (2013) proposed an extension to the method of Turney et al. (2011) by incorporating the concept of selectional preferences into the concreteness-based model of metaphor. Their goal was to improve the performance of Turney's algorithm by covering metaphors formed of concrete concepts only (e.g., "*broken* heart") by detecting selectional preference violations. The authors address three types of metaphor introduced by Krishnakumaran and Zhu (2007) and they claim to have expanded on Turney's work by carrying out a more comprehensive evaluation of the abstractness–concreteness algorithm. However, the evaluation was done on only five

target concepts: *governance*, *government*, *God*, *mother*, and *father*. Sentences describing these concepts have been extracted from the Reuters (Lewis et al. 2004) and *New York Times* (Sandhaus 2008) corpora and annotated for metaphoricity. The authors measured the average precision of their system on this data set at 0.72 and the average recall at 0.80. The improvement over Turney's evaluation set-up was the annotation of complete sentences rather than isolated phrases. However, it should be noted that the system was evaluated on selected examples rather than continuous text.

Heintz et al. (2013) applied Latent Dirichlet Allocation (LDA) topic modeling (Blei, Ng, and Jordan 2003) to the problem of metaphor identification in experiments with English and Spanish. Their goal was to create a minimally supervised metaphor processing system that can be applied to low-resource languages. The hypothesis behind their system is that a sentence that contains both source and target domain vocabulary contains a metaphor. The authors focused on the target domain of *governance* and have manually compiled a set of source concepts with which governance can be associated. They use LDA topics as proxies for source and target concepts, and if vocabulary from both source and target topics is present in a sentence, this sentence is tagged as containing a metaphor. The topics are learned from Wikipedia and then aligned to source and target concepts using sets of human-created seed words. When the metaphorical sentences are retrieved, the source topics that are common in the document are excluded, thus ensuring that the source vocabulary is transferred from a new domain. Although this allows the authors to filter out some literal uses, this may also lead to discarding cases of extended metaphor. The authors collected the data for their experiments from news Web sites and governance-related blogs in English and Spanish. They ran their system on this data, and output a ranked set of metaphorical examples. They carried out two types of evaluation: (1) top five examples for each conceptual metaphor judged by two annotators, reporting an F-score of 0.59 for English ($\kappa = 0.48$); and (2) 250 top-ranked examples in system output annotated for metaphoricity using Amazon Mechanical Turk, yielding a mean metaphoricity of 0.41 (standard deviation = 0.33) in English and 0.33 (standard deviation = 0.23) in Spanish. One of the assumptions behind Heintz et al.'s method is that the same source–target domain mappings manifest themselves across languages. Although this is likely to be true for primary metaphors (Grady 1997), as the authors point out themselves, this assumption may not extend to a broader spectrum of metaphors, and thus may lead to limited coverage in some languages, as well as false positives. Another issue that comes to mind concerns the learning of the topics themselves: Because a large number of metaphors are used conventionally within a particular topic (e.g. "*cut* taxes"), in principle such an approach would learn them as part of the target domain topics and may thus fail to recognize them as source domain terms. However, the authors do not comment on how often this was observed in their data.

The method of Strzalkowski et al. (2013) also relies on modeling the topical structure of text, although using different techniques from Heintz et al. (2013). If Heintz et al. used LDA-acquired topics as approximations of concepts, Strzalkowski and colleagues identify topical chains (Broadwell et al. 2013) by looking for sequences of concepts in text. They also experiment within a limited domain, the target domain of *governance*. Their method first identifies sentences containing target domain vocabulary and extracts the surrounding five-sentence passage. They then identify topical chains in that passage, by linking the occurrences of nouns and verbs, including repetition, lexical variants, pronominal references, and WordNet synonyms and hyponyms. By virtue of this linking, the authors claim to "uncover the topical structure [of the text] that holds the narrative together." Their main hypothesis is that metaphorically used terms

typically occur outside the core topical structure of the text, because they represent vocabulary imported from a different domain. For all the words that are found outside the topical chains, Strzalkowski et al. compute imageability scores and retain the highest-scoring ones as candidate metaphors, if they are in a syntactic relation with any of the target domain terms. The authors then extract common contexts in which the candidates are used in text corpora and cluster these contexts in order to identify potential source domains, the so-called "proto-sources." Strzalkowski et al. (2013) evaluated the performance of their method on four languages: English, Spanish, Russian, and Farsi. The evaluation was carried out against human judgments of system output that were obtained using Amazon Mechanical Turk. The authors report a metaphor identification accuracy of 71% in English, 80% in Spanish, 69% in Russian, and 78% in Farsi. According to the paper, "hundreds" of instances were annotated in each language, although the exact number of instances is not reported. While the system performance is high, it should be noted that the experiments were carried out within a limited domain, and it is possible that the approach is not equally applicable to all domains. Because of its high reliance on imageability scores, it is likely to be able to delineate metaphorical language reasonably well for the abstract target domains, but less so for the concrete target domains. In the latter case, the target domain words may also exhibit high imageability, and the system would then rely solely on topic chain extraction to differentiate between literal and metaphorical language. The performance of the generalized system is thus dependent on the accuracy of topic chain extraction, which has not been evaluated independently. In addition, the current method ignores low-imageability metaphors, which abound even within the studied domain (e.g., "*invent* a new form of governance"). Despite the lack of generality, Strzalkowski et al.'s work, however, makes important contributions in that it addresses (though indirectly) the behavior of metaphor in discourse, and their framework can be viewed as a step towards modeling extended metaphor.

Many other statistical methods treated metaphor identification as a classification problem. Such methods are described in the following section.

4.1.3 Metaphor Identification as a Classification Problem. Gedigian et al. (2006) presented a method that discriminates between literal and metaphorical language, using a maximum entropy classifier. They obtained their training and test data by extracting the lexical items whose frames are related to MOTION and CURE from FrameNet (Fillmore, Johnson, and Petruck 2003). They 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. They used PropBank annotation (arguments and their semantic types) as features to train the classifier and report an accuracy of 95.12%. This result is, however, only a little higher than the performance of the naive baseline assigning majority class to all instances (92.90%). These numbers can be explained by the fact that 92.00% of the verbs of MOTION and CURE in the *Wall Street Journal* corpus are used metaphorically, thus making the data set unbalanced with respect to the target categories and the task easier.

The system of Li and Sporleder (2009, 2010) detects idioms by measuring semantic similarity within and between the literal and non-literal parts of an utterance. The non-literal language considered by their model includes metaphors, as well as other types of figurative language. Their main assumption is that figurative uses break cohesion in the sentence, which is defined by a similarity measure. This idea also goes back to Wilk's selectional preference violation approach to metaphor; however, combinations of word usages with larger sentential context are considered to determine the mismatch

(or violation). Li and Sporleder used Normalized Google Distance (Cilibrasi and Vitanyi 2007) as a similarity measure and a combination of classifiers (support vector machines [SVM] and Gaussian mixture models [GMM]) using similarity (or cohesion) information as features to learn idiomaticity scores. They evaluated their system on a data set of 17 idioms and their literal and non-literal contexts. For each expression, its occurrences were extracted from the Gigaword corpus along with five paragraphs of context. These examples were then annotated for literalness with an inter-annotator agreement of $\kappa = 0.7$. There were 3,964 examples in total, with approximately 80% of them being non-literal. They evaluated the method using 10-fold cross-validation and report an F-score of 0.75. However, they did not evaluate their system on metaphorical language independently.

Dunn (2013a, 2013b) presented an ontology-based domain interaction system MIMIL (Measuring and Identifying Metaphor in Language), which identifies metaphorical expressions at the utterance level. Dunn's system first maps the lexical items in the given utterance to concepts from SUMO ontology (Niles and Pease 2001, 2003), assuming that each lexical item is used in its default sense (i.e., no sense disambiguation is performed). The system then extracts the properties of concepts from the ontology, such as their domain type (ABSTRACT, PHYSICAL, SOCIAL, MENTAL) and event status (PROCESS, STATE, OBJECT). Those properties are then combined into feature-vector representations of the utterances. Dunn then applied a logistic regression classifier implemented in Weka (Witten and Frank 2005), using these features to perform metaphor identification. The work of Dunn (2013a, 2013b) is notable as he conducted evaluation of four types of approaches and compared their performance on the same task (identification of metaphorical expressions in continuous text) and on the same data (Corpus of Contemporary American English [CoCA] [Davies 2009] and VU Amsterdam Metaphor Corpus [Steen et al. 2010]). The evaluated approaches included the semantic similarity measurement method of Li and Sporleder (2009, 2010); the concreteness-based method of Turney et al. (2011); the clustering-based method of Shutova, Sun, and Korhonen (2010) modeling source–target domain mappings; and his own domain interaction method. Dunn re-implemented the four approaches as closely as possible to the original systems, although with some adjustments. A number of Dunn's adjustments were operational (e.g., using logistic regression instead of SVM for the implementation of the similarity-based method of Li and Sporleder (2009); using a *k*-means clustering approach instead of spectral clustering for the method of Shutova, Sun, and Korhonen (2010); and using a different semantic relatedness measure for the method of Li and Sporleder (2009)). However, some adjustments were conceptual, for instance, using bag-of-words based semantic relatedness instead of dependency-based distributional similarity in the re-implementation of the clustering system. Admitting that these adjustments may have impacted the results and, as such, may not be an accurate reflection of the performance of the original algorithms in full, Dunn's comparison of individual system features that were re-implemented nonetheless sheds light on the importance of particular properties of concepts for metaphor identification. In his first study (Dunn 2013a), he evaluated the systems on the CoCA data, where the sentences were annotated as metaphorical, literal, or humorous (however, neither the size of the data set nor the annotation procedure are described in Dunn's article). On this data set, the clustering-based system and the domain-interaction method significantly outperformed the other two systems, as shown in Figure 3. Dunn explains such discrepancy by the fact that the former systems are both theory-based and aim to model the underlying mechanisms of metaphor, while the similarity-based and abstractness-based systems model its surface realizations. In his second study, conducted on the VU Amsterdam Metaphor corpus, Dunn (2013b)

System	True Pos.	False Pos.	True Neg.	False Neg.	F-Meas.
Similarity	1	0	2,482	504	0.004
Abstractness	1	2	2,482	505	0.004
Joint	67	44	2,446	444	0.215
MIMIL	133	382	2,437	63	0.374
Source-Tar.	113	461	2,038	300	0.229

Figure 3
Results of Dunn (2013a). The "Joint" system integrated similarity, abstractness, and domain-interaction features in the feature vectors.

Table 2
Dunn's (2013b) results on the VU Amsterdam Metaphor Corpus with named-entity recognition.

System	True Positive	False Pos.	True Negative	False Neg.	F-Measure
Similarity	5,936	4,214	86	62	0.444
Abstractness	4,627	3,049	3,752	2,954	0.582
Source-Target	1,063	785	5,470	5,496	0.440
Domain Interaction	5,446	3,664	3,106	2,286	0.583

reports different results, however. He evaluated the systems on two versions of the data set, one where named entities have been recognized during pre-processing and one without named-entity recognition. The results are shown in Tables 2 and 3, respectively. Here, the domain-interaction and abstractness-based methods are leading, with the clustering-based method coming third. The difference in the results may be explained by the properties of the VU Amsterdam corpus. The corpus was compiled with an interest in historic aspects of metaphor, and, therefore, highly conventional and lexicalized metaphors account for a large proportion of the data. What this suggests is that the domain-interaction and abstractness-based approaches are perhaps better-suited for processing lexicalized metaphors, whereas the clustering and similarity-based systems may fail to identify those due to their high frequency and the near-literal behavior in the data. In contrast, the domain-interaction system, which is knowledge-based, and the abstractness system, which relies on a non-changing property of concepts (i.e., concreteness), thus appear to be well-suited for handling lexicalized metaphors.

Tsvetkov, Mukomel, and Gershman (2013) presented a supervised learning approach that makes use of coarse semantic features. They experimented with metaphor

Table 3
Dunn's (2013b) results on the VU Amsterdam Metaphor Corpus without named-entity recognition.

System	True Positive	False Pos.	True Negative	False Neg.	F-Measure
Similarity	5,658	3,973	63	56	0.444
Abstractness	5,882	4,205	441	354	0.482
Source-Target	1,725	1,342	2,171	2,677	0.487
Domain Interaction	6,561	4,205	1,462	676	0.573

identification in English and Russian, first training a classifier on English data only, and then projecting the trained model to Russian using a dictionary. They abstracted from the words in the English data to their higher level features, such as concreteness, animateness, named entity labels, and coarse-grained WordNet categories (corresponding to WN lexicographer files,[5] e.g., *noun.artifact*, *noun.body*, *verb.motion*, *verb.cognition*). They focused on subject–verb–object constructions and annotated metaphor at the sentence level. The authors used a logistic regression classifier and the combination of coarse semantic features for this purpose. They evaluated their model on the TroFi data set (Birke and Sarkar 2006) for English and a self-constructed data set of 140 sentences for Russian, attaining the F-scores of 0.78 and 0.76, respectively. Tsvetkov et al. (2014) extended this experiment to identify adjective–noun metaphors using similar features, as well as porting the model to two further languages (Spanish and Farsi), achieving F-scores in the range of 0.72 to 0.85. The results are encouraging and show that porting coarse-grained semantic knowledge across languages is feasible. However, it should be noted that the generalization to coarse semantic features inevitably only captures shallow behavior of metaphorical expressions in the data and bypasses conceptual information. In reality, as confirmed by corpus-linguistic studies (Charteris-Black and Ennis 2001; Kovecses 2005; Diaz-Vera and Caballero 2013), there is considerable variation in metaphorical language across cultures, which makes training only on one language and simply translating the model less suitable for modeling conceptual structure behind metaphor, which is one of the limitations of this approach. However, the experiments of Tsvetkov and colleagues suggest that coarse semantic features could be a useful component of a more complex system.

The approach of Mohler et al. (2013) relied on the concept of semantic signature of a text. The authors defined semantic signatures as a set of highly related and interlinked WordNet senses. They induced domain-sensitive semantic signatures of texts and then trained a set of classifiers to detect metaphoricity within a text by comparing its semantic signature to a set of known metaphors. The main intuition behind this approach is that the texts whose semantic signature closely matches the signature of a known metaphor is likely to represent an instance of the same conceptual metaphor. Mohler and colleagues conducted their experiments within a limited domain (the target domain of *governance*) and manually constructed an index of known metaphors for this domain. They then automatically created the target domain signature and a signature for each source domain among the known metaphors in the index. This was done by means of semantic expansion of domain terms using WordNet, Wikipedia links, and corpus co-occurrence statistics. Given an input text their method first identified all target domain terms using the target domain signature, then disambiguated the remaining terms using sense clustering and classified them according to their proximity to the source domains listed in the index. For the latter purpose, the authors experimented with a set of classifiers, including a maximum entropy classifier, an unpruned decision tree classifier, support vector machines, a random forest classifier, as well as the combination thereof. They evaluated their system on a balanced data set containing 241 metaphorical and 241 literal examples, and obtained the highest result of F-score of 0.70 using the decision tree classifier.

Hovy et al. (2013) used the idea of selectional preference violation as the indicator of metaphor, taking it to the next level. They trained an SVM classifier (Cortes and Vapnik 1995) with tree kernels (Moschitti, Pighin, and Basili 2006) to capture compositional

5 http://wordnet.princeton.edu/man/lexnames.5WN.html.

*A **bright** idea.*

" Peter is the **bright** , sympathetic guy when you 're doing a deal , " says one agent .	*yes*
Below he could see the **bright** torches lighting the riverbank .	*no*
Her **bright** eyes were twinkling .	*yes*
Washed , they came out surprisingly clear and **bright** .	*no*

Figure 4
Data annotation example from Hovy et al. (2013).

properties of metaphorical language. Their hypothesis is that unusual semantic compositions in the data may be indicative of the use of metaphor. They trained the model on labeled examples of literal and metaphorical uses of 329 words (3,872 sentences in total), with an expectation to learn the differences in their compositional behavior in the given lexico-syntactic contexts. The choice of dependency-tree kernels helped to capture such compositional properties, according to the authors. The authors constructed their data set by extracting sentences from the Brown corpus (Francis and Kucera 1979) that contained the words of interest, and annotating them for metaphoricity using Amazon Mechanical Turk. Example entries for the adjective *bright* is shown in Figure 4. Eighty percent of the data were used for training purposes, 10% for parameter tuning, and 10% for the evaluation. The learning was carried out using word vectors, as well as lexical, part-of-speech tags, and WordNet supersense representations of sentence trees as features, as shown in Figure 5. The authors reported encouraging results (F-score = 0.75), which is an indication of the importance of syntactic information and compositionality in metaphor identification.

4.2 Identification of Conceptual Metaphors

The first method for automatic identification of conceptual metaphor was the CorMet system of Mason (2004). CorMet induced metaphorical mappings by identifying systematic variations in domain-specific selectional preferences, which were learned in a data-driven way. For example, the verb *pour* has a strong selectional preference for

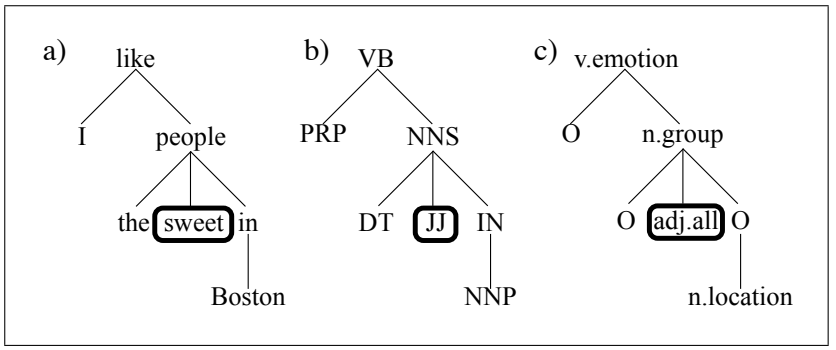

Figure 5
Dependency trees with lexical, part-of-speech, and WordNet supersense features from Hovy et al. (2013).

objects of type *liquid* in the LAB domain, and for *money* in the FINANCE domain. From this Mason's system inferred the domain mapping FINANCE—LAB and the concept mapping *money–liquid*. Mason used WordNet for acquisition of selectional preference classes and, therefore, the source and target domain categories were represented as clusters of WordNet synsets. The domain-specific corpora were obtained by searching the Web for specific terms of interest. Mason conducted two types of evaluation: (1) against the MML, where he manually mapped his output (WordNet synsets) to concrete concepts described in the MML (13 mappings in total) and then measured the accuracy at 77% (a mapping discovered by CorMet was considered correct if submappings specified in the MML were mostly present with high salience and incorrect submappings were present with relatively low salience); and (2) by compiling a list of mappings at random (assumed to be incorrect) and showing that the system assigned low scores to those. Baumer, Tomlinson, and Richland (2009) reimplemented the method of Mason (2004) in the framework of computational metaphor identification (CMI) procedure, and applied it to two types of corpora: student essays and political blogs. The authors presented some interesting examples of conceptual metaphors the system extracted, which they claim may foster critical thinking in social science. However, they did not carry out any quantitative evaluation.

Li, Zhu, and Wang (2013) proposed a method that performs metaphor identification using an "is-a" knowledge base. The authors automatically created two probabilistic knowledge bases by querying the Web using lexico-syntactic patterns. The first knowledge base contained hypernym–hyponym relations and was acquired using Hearst patterns (Hearst 1992). The second knowledge base contained metaphors in the form ⟨*target* is a *source*⟩ learned using a "*BE/VB like*" pattern. The second database was then filtered by removing the hypernym–hyponym relations present in the first database, as well as symmetric relations, to form a metaphor knowledge base. The authors applied the resulting metaphor knowledge base to perform metaphor recognition and explanation. They experimented with nominal metaphors (e.g., "Juliet is the *sun*") and verbal metaphors (e.g., "My car *drinks* gasoline"). In the case of nominal metaphors, the database was queried directly and the corresponding metaphor was either retrieved or not. In the case of verbal metaphors, where the noun denoting the source concept was not explicitly present in the sentence, it was derived based on the selectional preferences of the verbs. The authors computed selectional preferences of the given verb for the nouns present in the knowledge base, and "explained" the given metaphor by the noun exhibiting the highest selectional association with the metaphorical verb. For example, it outputs an explanation "car is a horse" for the metaphor in "my car *drinks* gasoline," since the conceptual metaphor CAR IS A HORSE is present in the knowledge base and *horse* satisfies the subject preference of *drink*. The authors evaluated their approach on a manually constructed data set of 200 randomly sampled sentences containing "is-a" constructions and 1,000 sentences containing metaphorical and literal uses of verbs. The annotation was carried out at the sentence level (i.e., complete sentences were annotated as metaphorical or not). The authors report an F-score of 69% on the recognition of "is-a" metaphors and that of 58% on the recognition of the verbal ones. Metaphor explanation performance (i.e., the source–target domain mappings generated for each recognized metaphor) was evaluated separately on 214 sentences extracted from linguistic literature (Lakoff and Johnson 1980) and the top-rank precision of 43% is reported. Intuitively, a purely simile-based approach to metaphor is likely to both undergenerate (a large number of metaphors would never be manifested in simile-like constructions) and overgenerate ("A is like B" pattern may describe other relations than metaphor). The key contribution of Li, Zhu, and Wang appears to be the filtering

method they introduce, as well as the selectional preference extension of the knowledge base to identify verbal metaphors.

The method of Shutova and Sun (2013) learns metaphorical associations between concepts from the data in an unsupervised way. They created a network (or a graph) of concepts, using hierarchical graph factorization clustering of nouns, and quantified the strength of association between concepts in this graph. Concrete concepts exhibited well-defined association patterns mainly based on subsumption within one domain, whereas abstract concepts tended to have both within-domain and cross-domain associations: the literal ones and the metaphorical ones. For example, the abstract concept of *democracy* was literally associated with a more general concept of *political system*, as well as metaphorically associated with the concept of *mechanism*. Because we often discuss *political systems* using the *mechanism* terminology, a corpus-based distributional learning approach learns that they share features with *political systems* (from their literal uses), as well as with *mechanisms* (from their metaphorical uses). The system of Shutova and Sun (2013) automatically discovered such association patterns within the graph and used them to identify metaphorical mappings. The mappings were represented in their system as cross-level, one-directional connections between clusters in the hierarchical graph (e.g., the *feeling* cluster was strongly associated with *fire*). Example output for the source concepts of *fire* and *disease* is shown in Figure 6. To identify metaphorical expressions representing a given mapping, Shutova and Sun used the features that resulted in strong metaphorical associations between the clusters in question (e.g., "passion *flared*" for FEELING IS FIRE), as shown in Figure 7. The authors evaluated the quality of metaphorical mappings and metaphorical expressions identified by the system against human judgments, as follows: (1) the human judges were presented with a random sample of system-produced metaphorical mappings between the clusters of nouns, as well as the corresponding metaphorical expressions, and asked to mark the ones they considered valid as correct; (2) the human annotators were presented with a set of source domain concepts and asked to write down all target concepts they associated with a given source, thus creating a gold standard. Shutova and Sun report the precision of 0.69 for metaphorical associations and 0.65 for metaphorical expressions, as evaluated against human judgments, and the recall of 0.61 for metaphorical associations, as evaluated against a human-created gold standard. These results are encouraging in that they show that it is possible to induce information about metaphorical mechanisms from distributional properties of concepts alone, without

SOURCE: fire
TARGET 1: sense hatred emotion passion enthusiasm sentiment hope interest feeling resentment optimism hostility excitement anger
TARGET 2: coup violence fight resistance clash rebellion battle drive fighting riot revolt war confrontation volcano row revolution struggle
TARGET 3: alien immigrant
TARGET 4: prisoner hostage inmate

SOURCE: disease
TARGET 1: fraud outbreak offense connection leak count crime violation abuse conspiracy corruption terrorism suicide
TARGET 2: opponent critic rival
TARGET 3: execution destruction signing
TARGET 4: refusal absence fact failure lack delay

Figure 6
Metaphorical associations discovered by the system of Shutova and Sun (2013).

FEELING IS FIRE hope *lit* (Subj), anger *blazed* (Subj), optimism *raged* (Subj), enthusiasm *engulfed* them (Subj), hatred *flared* (Subj), passion *flared* (Subj), interest *lit* (Subj), *fuel* resentment (Dobj), anger *crackled* (Subj), feelings *roared* (Subj), hostility *blazed* (Subj), *light* with hope (Iobj)
CRIME IS A DISEASE *cure* crime (Dobj), abuse *transmitted* (Subj), *eradicate* terrorism (Dobj), *suffer from* corruption (Iobj), *diagnose* abuse (Dobj), *combat* fraud (Dobj), *cope with* crime (Iobj), *cure* abuse (Dobj), *eradicate* corruption

Figure 7
Shutova and Sun (2013): Metaphorical expressions identified for the mappings FEELING IS FIRE and CRIME IS A DISEASE.

the use of hand-coded knowledge. Nevertheless, the fact that clustering techniques are typically applied to a limited set of concepts (i.e., cluster a limited set of nouns) somewhat constrains this approach. For instance, whereas common concepts (that are well represented in the data) can be clustered with a high accuracy, this is not always the case for rare concepts for which feature vectors are sparse. Thus an additional technique is needed to map new, unseen concepts to the concepts present in the graph.

Gandy et al. (2013) presented a system that first discovers metaphorical expressions using concreteness algorithm of Turney et al. (2011) and then assigns the corresponding metaphorical mappings using lexical resources and context clustering. They focused on the three types of metaphor defined by Krishnakumaran and Zhu (2007). Once the metaphorical expressions had been identified, Gandy et al. extracted the nouns that the metaphorical words, or *facets*, tend to co-occur within a large corpus (e.g., the nominal arguments of *open* in "*open* government"). The goal of this process was to form candidate *nominal analogies* between the target noun in the metaphor and the extracted nouns. For example, the expression "*open* government" suggests an analogy "government ∼ door," according to the authors. Figure 8 shows how nominal analogies are formed based on

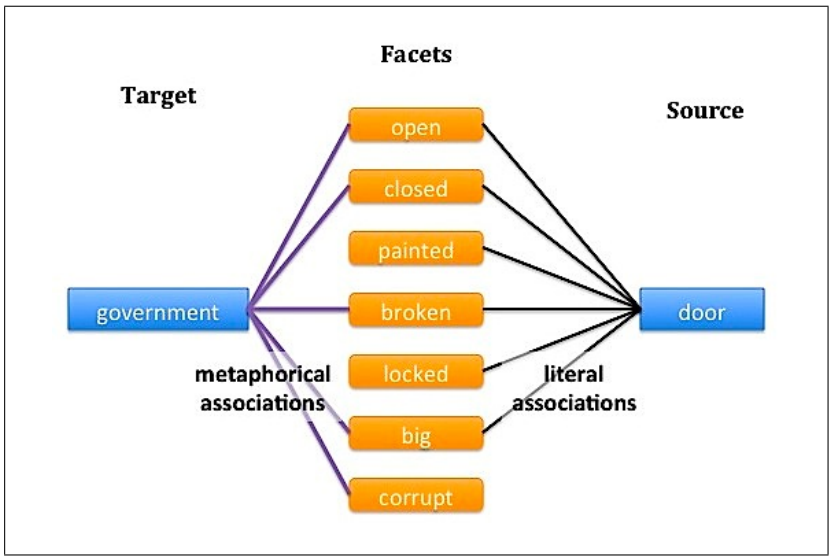

Figure 8
Nominal analogy induction from Gandy et al. (2013).

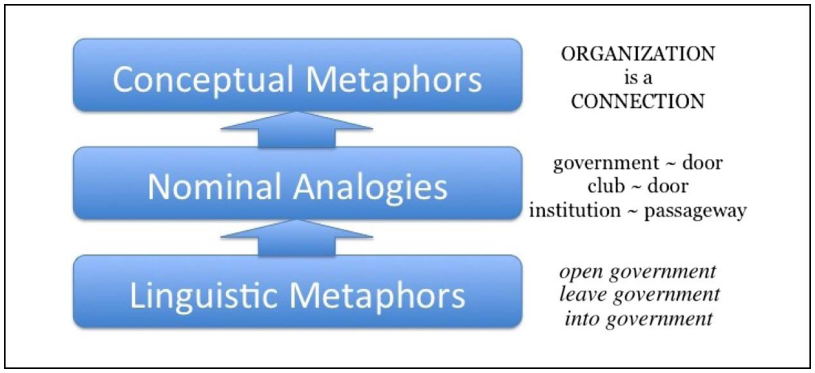

Figure 9
Gandy et al. (2013): Three levels of analysis.

a collection of metaphorical expressions. The individual (related) nominal analogies were then clustered together to identify conceptual metaphors, as shown in Figure 9. The authors evaluated their system by annotating metaphorical expressions for five target concepts (*government*, *governance*, *god*, *father*, and *mother*) in selected sentences from the Reuters corpus (Lewis et al. 2004). They report very encouraging results: Precision (P) = 0.76, Recall (R) = 0.82 for verb metaphors; P = 0.54, R = 0.43 for adjectival metaphors; and P = 0.84, R = 0.97 for copula constructions. The authors also evaluated the quality of conceptual metaphors produced by the system against human judgments and attained a precision of 0.65. However, the scope of the experiment is only limited to the given five concepts and it is not clear how well the method would generalize beyond these. Although the approach of Gandy et al. (2013) seems very promising, a comprehensive evaluation on open-domain corpus data is still necessary to prove its viability.

5. Metaphor Interpretation Systems

In one of the first approaches to metaphor interpretation, Martin (1990) presented a Metaphor Interpretation, Denotation, and Acquisition System (MIDAS), which explained linguistic metaphors through finding the corresponding conceptual metaphor. The method is based on the idea of hierarchical organization of conventional metaphors, namely, that more specific conventional metaphors descend from the general ones. Given an example of a metaphorical expression, MIDAS searched its database for a corresponding metaphor that would explain the anomaly. If it did not find any, it abstracted from the example to more general concepts and repeated the search. If it found a suitable general metaphor, it created a mapping for its descendant, a more specific metaphor, based on the given example. This was also how novel metaphors were acquired. MIDAS was integrated with the Unix Consultant (UC), the system that answers users' questions about Unix. The UC first tried to find a literal answer to the question. Failing to do so, it called MIDAS, which detected metaphorical expressions via selectional preference violation and searched its database for a metaphor explaining the anomaly in the question.

Another branch of early work on metaphor interpretation relied on performing inferences about entities and events in the source and target domains. The most prominent approaches 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 the theory of conceptual metaphor. The reasoning process relied on manually constructed knowledge about the world and operated mainly in the source domain. The results were then projected onto the target domain using the conceptual mapping representation. The ATT-Meta project concerned metaphorical and metonymic description of mental states and reasoning about mental states using first order logic. Their system, however, did not take natural language sentences as input, but logical expressions that are representations of small discourse fragments. KARMA in turn dealt with a broad range of abstract actions and events and took parsed text as input.

Since then the field moved towards acquiring the knowledge necessary for metaphor interpretation automatically (and at a larger scale) from lexical resources, corpora, and the Web. Veale and Hao (2008) derived a "fluid knowledge representation for metaphor interpretation and generation," called Talking Points. Talking Points are a set of characteristics of concepts belonging to source and target domains and related facts about the world which the authors acquired automatically from WordNet and from the Web. Talking Points were then organized in *Slipnet*, a framework that allowed 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 built on the idea of **slippage** in knowledge representation for understanding analogies in abstract domains (Hofstadter and Mitchell 1994; Hofstadter 1995). Example (6) demonstrates how slippage operates to explain the metaphor *Make-up is a Western burqa*.

(6) **Make-up** =>
≡ typically worn by women
≈ expected to be worn by women
≈ must be worn by women
≈ must be worn by Muslim women
Burqa <=

By doing insertions and substitutions the system arrived from the definition *typically worn by women* to that of *must be worn by Muslim women*, and thus established a link between the concepts of *make-up* and *burqa*. Veale and Hao (2008), however, did not evaluate to what extent their knowledge base of Talking Points and the associated reasoning framework are useful to interpret metaphorical expressions occurring in text.

Shutova (2010) defined metaphor interpretation as a paraphrasing task and presented a method for deriving literal paraphrases for metaphorical expressions from the BNC. For example, for the metaphors in "All of this *stirred* an unfathomable excitement in her" or "a carelessly *leaked* report," their system produced interpretations *All of this provoked an unfathomable excitement in her* and *a carelessly disclosed report*, respectively. They first applied a probabilistic model to rank all possible paraphrases for the metaphorical expressions, given the context; and then used automatically induced selectional preferences to discriminate between figurative and literal paraphrases. The selectional preference distribution was defined in terms of selectional association measures introduced by Resnik (1993) over the noun classes automatically produced by Sun and Korhonen (2009). Shutova (2010) tested her system only on metaphors expressed by a verb and reports an accuracy of 0.81, as evaluated on top-ranked paraphrases produced by the system. However, she used WordNet for supervision, which limits the number and range of paraphrases that can be identified by her method. Shutova, Van de Cruys,

and Korhonen (2012) and Bollegala and Shutova (2013) expanded on this work, addressing the metaphor paraphrasing task in an unsupervised setting and extending the coverage. The method of Shutova, Van de Cruys, and Korhonen (2012) first computed candidate paraphrases according to the context in which the metaphor appeared, using a vector space model. It then used a selectional preference model to measure the degree of literalness of the paraphrases. The authors evaluated their method on the metaphor paraphrasing data set of Shutova (2010) and reported a top-rank precision of 0.52. Bollegala and Shutova (2013) used a similar experimental set-up, however, their method extracted a set of candidate paraphrases from the Web using lexico-syntactic patterns as queries and ranked them based on search engine hits, attaining a precision of 0.42.

Shutova, Teufel, and Korhonen (2013) combined metaphor identification (Shutova, Sun, and Korhonen 2010) and interpretation (Shutova 2010) to perform text-to-text metaphor processing. The resulting system could take arbitrary text as input, parse it using a syntactic parser, identify metaphorical expressions in it, retrieve their literal paraphrases, and output a new version of the text in which metaphors were interpreted. The motivation behind such a set-up was that it allowed for a relatively straightforward integration with external NLP applications. To evaluate the system, the authors extracted a random sample of 200 metaphorical expressions the system identified in the BNC and applied the paraphrasing method to them. They evaluated the accuracy of metaphor identification and interpretation when performed simultaneously, as well as the system's applicability. The applicability was defined as the proportion of cases where the paraphrase was literal and the meaning of the phrase was retained, indicating whether this type of system paraphrasing would result in an error when hypothetically integrated with an external NLP application. In 54% of cases, the system both identified and interpreted the metaphor correctly, which is a promising result. In a further 13% of cases, the system produced a correct, literal paraphrase for a literal expression erroneously identified as a metaphor, leading to the overall applicability of integrated metaphor processing at 67%. Although the system is easy to integrate with external NLP applications that could benefit from metaphor resolution, it should be noted that some information conveyed by the metaphor is inevitably lost during literal paraphrasing. Metaphor paraphrasing as an approach thus rests on a crucial assumption that the benefit of correct metaphor understanding would outweigh the loss of additional connotations and rhetorical elements. This assumption is yet to be verified through an integration of this technology into real-world NLP, however.

Shutova (2013) presented a computational method that identified metaphorical expressions in unrestricted text by means of their interpretation. She again treated metaphor interpretation as paraphrasing and introduced the concept of **symmetric reverse paraphrasing** as a criterion for metaphor identification. The hypothesis behind the method is that literal paraphrases of literally used words should yield the original phrase when paraphrased in reverse. For example, when the expression *clean the house* is paraphrased as *tidy the house,* the reverse paraphrasing of *tidy* would generate *clean* as one of possible paraphrases. Shutova's expectation was that such symmetry in paraphrasing is indicative of literal use. The metaphorically used words are unlikely to exhibit this symmetry property when paraphrased in reverse. For example, the literal paraphrasing of the verb *stir* in "*stir* excitement" would yield "*provoke* excitement," but the reverse paraphrasing of *provoke* would not retrieve *stir*, indicating the non-literal use of *stir*. Shutova experimentally verified this hypothesis in a setting involving single-word metaphors expressed by a verb in verb–subject and verb–direct object relations. She applied the selectional preference-based metaphor paraphrasing method (Shutova 2010) to retrieve literal paraphrases of all input verbs and extended the method to

FYT Gorbachev **inherited** a Soviet state which was, in a celebrated Stalinist formulation, "national in form but socialist in content."
Paraphrase: Gorbachev received a Soviet state which was, in a celebrated Stalinist formulation, "national in form but socialist in content."

CEK The Clinton campaign **surged** again and he easily won the Democratic nomination.
Paraphrase: The Clinton campaign improved again and he easily won the Democratic nomination.

CEK Their views **reflect** a lack of enthusiasm among the British people at large for John Major's idea of European unity.
Paraphrase: Their views show a lack of enthusiasm among the British people at large for John Major's idea of European unity.

J85 [..] the reasons for this superiority are never **spelled out**.
Paraphrase [..] the reasons for this superiority are never specified.

J85 Anyone who has introduced speech act theory to students will know that these technical terms are not at all easy to **grasp**.
Paraphrase: Anyone who has introduced speech act theory to students will know that these technical terms are not at all easy to understand.

G0N The man's voice **cut in** .
Paraphrase: The man's voice interrupted.

Figure 10
Metaphors tagged by the system of Shutova (2013) (in **bold**) and their paraphrases.

perform metaphor identification by reverse paraphrasing. She evaluated the performance of the system on verb–subject and verb–object relations using the manually annotated metaphor corpus of Shutova and Teufel (2010), reporting a precision of 0.68 and a recall of 0.66. The system outperformed a baseline using selectional preference violation as an indicator of metaphor, that only attained a precision of 0.17 and a recall of 0.55. Some examples of metaphorical expressions identified by the system and their literal paraphrases are shown in Figure 10.

6. Investigated Techniques and Lessons Learned

The community has investigated a wide range of techniques and features for metaphor identification and interpretation in a variety of experimental settings. The majority of identification systems focus on the linguistic level, identifying either linguistic metaphor or non-literal language more generally, with a few identifying conceptual metaphor. Table 4 presents a summary of the tasks addressed. Some systems annotated metaphorical expressions at the word level, whereas others opted for the relation level or carried out sentence-level annotation. Individual approaches frequently limited the scope of their experiments to metaphors expressed by a particular part of speech and syntactic construction, as shown in Table 5. The majority of the systems focused on metaphorically used verbs or adjectives, with a few also considering nouns (in modifier or copula constructions) and multiword metaphors. The systems that identified conceptual metaphor also exhibit some variation in the representations they used. Source and target domains were represented as WordNet synsets (Mason 2004); individual nouns (Li, Zhu, and Wang 2013); or clusters of nouns (Shutova and Sun 2013; Gandy et al. 2013). Recent work on metaphor interpretation unfolded along two main axes: metaphor explanation (i.e., identifying the properties of concepts that the metaphor

Table 4
Identification systems: Task definition.

System	Ling.	Concept.	Non-lit.	Word	Relation	Sent.
Mason (2004)	–	✔	–	–	–	–
Birke and Sarkar (2006)	✔	–	✔	✔	–	–
Gedigian et al. (2006)	✔	–	–	✔	–	–
Krishnakumaran and Zhu (2007)	✔	–	–	–	–	✔
Shutova et al. (2010)	✔	–	–	✔	✔	–
Li and Sporleder (2009, 2010)	✔	–	✔	–	✔	–
Turney et al. (2011)	✔	–	–	✔	✔	–
Neuman et al. (2013)	✔	–	–	✔	✔	–
Dunn (2013a, b)	✔	–	–	–	–	✔
Tsvetkov et al. (2013, 2014)	✔	–	–	–	–	✔
Mohler et al. (2013)	✔	–	–	–	–	✔
Heintz et al. (2013)	✔	–	–	✔	–	–
Hovy et al. (2013)	✔	–	–	✔	–	–
Wilks et al. (2013)	✔	–	–	✔	✔	–
Strzalkowski et al. (2013)	✔	–	–	–	–	✔
Shutova and Sun (2013)	✔	✔	–	✔	✔	–
Gandy et al. (2013)	✔	✔	–	✔	✔	–
Li et al. (2013)	✔	✔	–	–	–	✔
Shutova (2013)	✔	–	–	✔	✔	–

Table 5
Identification systems: Parts of speech and constructions covered.

System	Verb	Adjective	Nominal	Copula	Multi-word
Mason (2004)	–	–	–	–	–
Birke and Sarkar (2006)	✔	–	–	–	–
Gedigian et al. (2006)	✔	–	–	–	–
Krishnakumaran and Zhu (2007)	✔	✔	–	✔	–
Shutova et al. (2010)	✔	–	–	–	–
Li and Sporleder (2009; 2010)	–	–	–	–	✔
Turney et al. (2011)	✔	✔	–	–	–
Neuman et al. (2013)	✔	✔	–	✔	–
Dunn (2013a, b)	✔	✔	✔	✔	✔
Tsvetkov et al. (2013, 2014)	✔	✔	–	–	–
Mohler et al. (2013)	✔	✔	✔	✔	✔
Heintz et al. (2013)	✔	✔	✔	✔	✔
Hovy et al. (2013)	✔	✔	✔	✔	✔
Wilks et al. (2013)	✔	–	–	–	–
Strzalkowski et al. (2013)	✔	–	✔	–	–
Shutova and Sun (2013)	✔	–	–	–	–
Gandy et al. (2013)	✔	✔	–	✔	–
Li et al. (2013)	✔	–	✔	✔	–
Shutova (2013)	✔	–	–	–	–

highlights and the comparisons it involves [Veale and Hao 2008]) and metaphor paraphrasing (i.e., identifying a literal [or more conventional] paraphrase of the metaphorical expression [Shutova 2010; Bollegala and Shutova 2013]).

Identification and interpretation approaches investigated a range of properties of metaphor and implemented them in a variety of system components. The most

Table 6
Identification systems: System features and techniques.

System / features	Sel. pref.	Violation	Concreteness	Topical structure	Clustering	ML Classifier	Lex resource	Web search	supervised	weakly-supervised	unsupervised
Mason (2004)	✓	–	–	–	✓	–	✓	✓	✓	–	–
Birke and Sarkar (2006)	–	–	–	–	✓	–	–	–	–	✓	–
Gedigian et al. (2006)	–	–	–	–	–	✓	✓	–	✓	–	–
Krishnakumaran and Zhu 2007)	–	✓	–	–	–	–	✓	–	✓	–	–
Shutova et al. (2010)	✓	–	–	–	✓	–	–	–	–	✓	–
Li and Sporleder (2009, 2010)	–	✓	–	–	–	✓	–	–	✓	–	✓
Turney et al. (2011)	–	–	✓	–	–	✓	–	–	✓	–	–
Neuman et al. (2013a, b)	✓	✓	✓	–	–	–	–	–	✓	–	–
Dunn (2013)	–	–	–	–	–	✓	–	–	✓	–	–
Tsvetkov et al. (2013, 2014)	–	–	✓	–	–	✓	✓	–	✓	–	–
Mohler et al. (2013)	–	–	–	✓	✓	✓	✓	–	✓	–	–
Heintz et al. (2013)	–	–	–	✓	–	–	–	–	–	✓	–
Hovy et al. (2013)	–	✓	–	–	–	✓	✓	–	✓	–	–
Wilks et al. (2013)	✓	✓	–	–	–	–	✓	–	✓	–	–
Strzalkowski et al. (2013)	–	–	✓	✓	✓	–	✓	–	✓	–	–
Shutova and Sun (2013)	✓	–	–	–	✓	–	–	–	–	–	✓
Gandy et al. (2013)	–	–	✓	–	✓	–	✓	–	✓	–	–
Li et al. (2013)	✓	–	–	–	–	–	–	✓	–	✓	–
Shutova (2013)	✓	–	–	–	–	–	✓	–	✓	–	–

prominent ones include selectional preferences (Martin 1990; Fass 1991; Mason 2004; Krishnakumaran and Zhu 2007; Li and Sporleder 2009, 2010; Shutova 2010; Shutova, Sun, and Korhonen 2010; Hovy et al. 2013; Li, Zhu, and Wang 2013; Wilks et al. 2013); semantic properties of concepts, such as imageability and concreteness (Turney et al. 2011; Gandy et al. 2013; Neuman et al. 2013; Strzalkowski et al. 2013); and topical structure of text (Heintz et al. 2013; Strzalkowski et al. 2013). The common methods used include supervised classification (Gedigian et al. 2006; Dunn 2013a; Hovy et al. 2013; Mohler et al. 2013; Tsvetkov, Mukomel, and Gershman 2013); clustering (Gandy et al. 2013; Shutova and Sun 2013; Shutova, Sun, and Korhonen 2010; Strzalkowski et al. 2013); vector space models (Shutova, Van de Cruys, and Korhonen 2012); the use of lexical resources and ontologies (Mason 2004; Krishnakumaran and Zhu 2007; Dunn 2013b; Gandy et al. 2013; Hovy et al. 2013; Mohler et al. 2013; Strzalkowski et al. 2013; Tsvetkov, Mukomel, and Gershman 2013; Wilks et al. 2013); and Web search (Veale and Hao 2008; Bollegala and Shutova 2013; Li, Zhu, and Wang 2013). A summary of techniques investigated by the community is presented in Table 6. In what follows we will discuss the main trends in metaphor processing research and the usefulness of individual types of techniques.

6.1 Selectional Preferences

Selectional preferences have long established themselves as one of the central components in metaphor-processing research. Wilks' (1978) selectional preference violation view of metaphor has been highly influential, with numerous approaches to metaphor identification implementing it directly or indirectly (Fass 1991; Martin 1990; Wilks et al. 2013). Other approaches modified this view and treated metaphor as a violation of semantic norm construed more broadly—for example, searching for expressions with low bi-gram probabilities (Krishnakumaran and Zhu 2007), identifying units that break sentence cohesion (Li and Sporleder 2009, 2010), or detecting unusual patterns in words' compositional behavior (Hovy et al. 2013).

Generally speaking, selectional preference violations (or other semantic violations mentioned above) are a property of surface realization of metaphor rather than its underlying conceptual mechanisms. One needs to bear this in mind when using this as a heuristic. On one hand, such violations are indicative of any kind of non-literalness (i.e., not only metaphor, but also, for instance, metonymy) or anomaly in language and the approach is likely to overgenerate. On the other hand, in the case of most conventional metaphors that are highly frequent, no statistically significant violation can be detected in the data, and the approach would bypass many such metaphors. Shutova (2013) conducted a data-driven study, where verb preferences were automatically acquired from the data and all the nominal arguments below a certain selectional association threshold were considered to represent a violation and were tagged as metaphorical. Such a technique attained a precision of 0.17 and a recall of 0.55, suggesting that the selectional preference violation hypothesis does not port well beyond handcrafted descriptions to large-scale, data-driven techniques.

In contrast, other, "non-violation" applications of selectional preferences have been fruitful in metaphor modeling. Mason (2004) automatically acquired domain-specific selectional preferences of verbs, and then, by mapping their common nominal arguments in different domains, arrived at the corresponding metaphorical mappings. Shutova (2010) presented a modification of Wilks' view, treating a strong selectional preference fit as a likely indicator of literalness or conventionality. In her metaphor paraphrasing system, Shutova ranks candidate paraphrases based on how well the

context fits their preferences, thus determining their literalness. In their metaphor identification system, Shutova, Sun, and Korhonen (2010) filtered out verbs that have weak selectional preferences, that is, that are equally associated with many argument classes (e.g., *choose* or *remember*), as having a lower metaphorical potential. Li, Zhu, and Wang (2013) used selectional preferences to assign the corresponding conceptual metaphor to metaphorical expressions. Although the idea of violation (i.e., treating metaphor as merely a context outlier) is controversial and should be applied with care, selectional preferences themselves are an important source of semantic information about the properties of concepts, which can be successfully exploited in metaphor processing in a variety of ways.

6.2 Topical Structure of Text

Two approaches (Heintz et al. 2013; Strzalkowski et al. 2013) focused on modeling topical structure of text to identify metaphor. The main hypothesis behind these methods is that metaphorical language (coming from a different domain) would represent atypical vocabulary within the topical structure of the text. This intuition is somewhat similar to the idea of semantic norm violation as an indicator of metaphor, although it is different in two crucial ways: (1) topical structure-based approaches explicitly model the interaction of vocabulary from two different domains (i.e., the source and the target); and (2) these approaches take into account domain interactions over extended discourse fragments, rather than individual expressions, thus utilizing information from wider context. Exploiting the wider topical structure of text is a promising avenue for metaphor processing. However, one needs to keep in mind that distributional similarity-based methods risk assigning frequent metaphors to target domains (as is the case for other semantic violation-based methods). For instance, *cut* may appear more frequently within the domain of economics and finance, rather than its original source domain. The choice of data for training such a model thus becomes crucial, and an appropriately balanced data set is needed. Investigating the topical structure of text is also an important step towards modeling extended metaphor, which interweaves the narrative in complex, but systematic ways.

6.3 Concreteness

Turney et al. (2011) introduced the idea of measuring concreteness of concepts to predict metaphorical use. The intuition behind their approach is that metaphor is commonly used to describe abstract concepts in terms of more concrete or physical experiences. Thus, Turney and colleagues expect that there would be some discrepancy in the level of concreteness of source and target terms in the metaphor. Neuman et al. (2013) and Gandy et al. (2013) followed in Turney's steps, reporting promising results. Tsvetkov, Mukomel, and Gershman (2013) took a different route, and used concreteness as one of the features to train a classifier. Strzalkowski et al. (2013) experimented with the imageability feature (that indicates how easy it is to visualize a concept) and demonstrated its relevance to metaphor identification.

Based on the results of these experiments, concreteness is likely to be a practically useful feature for metaphor processing. However, it should be noted that Turney's hypothesis (that target words tend to be abstract and source words tend to be concrete) explains only a fraction of metaphors and does not always hold. For example, one can use concrete–concrete metaphors (e.g., "*broken* heart"), abstract–abstract metaphors ("*diagnose* corruption"), and even abstract–concrete metaphors ("*invent* a soup").

However, it may be the case that within the concrete–abstract class of metaphor, the method operates with reasonable performance. Thus concreteness may become a useful feature of a more complex system that takes multiple factors into account, but is unlikely to be a reliable indicator of metaphor on its own.

6.4 Supervised Classification

A number of approaches trained classifiers on manually annotated data to recognize metaphor. The key question that supervised classification poses is what features are indicative of metaphor and how can one abstract from individual expressions to its high-level mechanisms? The community has experimented with a number of features, including lexical and syntactic information; higher-level features such as semantic roles, WordNet supersenses, named-entity types, and domain types extracted from ontologies; and semantic properties of concepts, such as animateness and concreteness. Gedigian et al. (2006) classified verb uses as literal or metaphorical, using the verbs' nominal arguments and their semantic roles (as annotated in PropBank) as features. They reported unusually high performance scores, although the narrow focus on specific lexical items makes it possible for the system to learn a model for individual words rather than performing generalization. In contrast, Dunn (2013a, 2013b) experimented with a wide range of metaphorical expressions from the VU Amsterdam Metaphor corpus, using high-level properties of concepts extracted an ontology, such as domain type and event status. Tsvetkov, Mukomel, and Gershman (2013) and Tsvetkov et al. (2014) used coarse semantic features, such as concreteness, animateness, named-entity types, and WordNet supersenses. What is particularly interesting about this work is that the authors have shown that the model learned with such coarse semantic features is portable across languages, thus suggesting that the chosen features successfully capture some of the properties of metaphor (even if the shallow ones). The work of Hovy et al. (2013) is notable as they focused on compositional rather than categorical features. They trained an SVM with dependency-tree kernels to capture compositional information using lexical, part-of-speech tags, and WordNet supersense representations of sentence trees, achieving successful results. The system of Mohler et al. (2013) aimed at modeling conceptual information in the form of semantic signatures of domains and metaphors. Such rich semantic information is likely to be a successful feature in metaphor recognition: however, Mohler and colleagues experimented within a limited domain and it is not clear how scalable such features would be.

To reliably capture the patterns of the use of metaphor in the data at a large scale, one needs to address conceptual properties of metaphor, along with the surface ones. Thus the models making generalizations at the level of metaphorical mappings and coarse-grained classes of concepts, in essence representing different domains, are likely to yield the optimal framework for the task. However, this hypothesis is yet to be experimentally verified.

6.5 Clustering

Clustering techniques were used in numerous approaches, predominantly to identify concepts similar or related to each other. Mason (2004) performed WordNet sense clustering to obtain selectional preference classes, whereas Mohler et al. (2013) used it to determine similarity between concepts and to link them in semantic signatures. Strzalkowski et al. (2013) and Gandy et al. (2013) clustered metaphorically used terms to

form potential source domains. Birke and Sarkar (2006) clustered sentences containing metaphorical and literal uses of verbs.

Another line of research focused on the use of clustering methods to investigate how metaphor partitions the linguistic feature space. Shutova, Sun, and Korhonen (2010) pointed out that the metaphorical uses of words constitute a large portion of the co-occurrence features extracted for abstract concepts from the data. For example, the feature vector for *politics* would contain GAME or MECHANISM terms among the frequent features. As a result, distributional clustering of abstract nouns with such features identifies groups of diverse concepts metaphorically associated with the same source domain (or sets of source domains). Shutova, Sun, and Korhonen exploit this property of co-occurrence vectors to identify new metaphorical mappings starting from a set of examples. The work of Shutova and Sun (2013) is based on the same observation. Through the use of hierarchical clustering techniques they derive a network of concepts in which metaphorical associations are exhibited at different levels of generality.

6.6 The Use of Lexical Resources

Peters and Peters (2000) and Wilks et al. (2013) detected metaphor directly in lexical resources. Peters and Peters mine WordNet for examples of systematic 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. Wilks et al. (2013) used WordNet glosses to learn selectional preferences of verbs, which were then used to annotate senses as literal or metaphorical, based on the selectional preference–violation hypothesis. Krishnakumaran and Zhu (2007) use hyponymy relation in WordNet to detect semantic violations. Shutova (2010, 2013) also relied on the hierarchical structure of WordNet, but to identify concepts that share common features (defined as sharing a common hypernym within three levels of the hierarchy).

WordNet synsets were also used to form selectional preference classes in SP-based methods (Mason 2004) or to detect semantically related concepts (Mohler et al. 2013; Strzalkowski et al. 2013; Gandy et al. 2013). Other researchers used WordNet to identify high-level properties of concepts, most notably WordNet supersenses, that served as features for classification (Tsvetkov, Mukomel, and Gershman 2013; Hovy et al. 2013).

6.7 Web Search

Because metaphor is a knowledge-intensive phenomenon, multiple approaches attempted to acquire the knowledge necessary for its identification and interpretation from the Web (Veale and Hao 2008; Bollegala and Shutova 2013; Li, Zhu, and Wang 2013). Web search engines provide a flexible tool for retrieving information that matches specific lexico-syntactic patterns (used as queries) and quantifying co-occurrence. Veale and Hao (2008) query the Web to harvest properties of concepts and cultural stereotypes, such as *has magical skill* for Wizard or *has brave spirit* for Lion, which are then used to perform metaphor interpretation through property comparison and substitution. Bollegala and Shutova (2013) use the Web to extract co-occurrence information for verbs and nouns, which allows them to generate a set of candidate paraphrases for metaphorical verbs. Li, Zhu, and Wang (2013) query the Web with Hearst patterns to acquire a large knowledge base of hyponymy relations; and simile patterns

(* *is like* *) to acquire a set of potential conceptual metaphors. Along with the flexibility and convenience of information retrieval tools, these three approaches also boast of wide coverage that the use of the Web allows them to achieve. The knowledge contained on the Web is not merely vast, but it is also constantly updated, which allows the system to stay on par with the current events and trends. Because metaphor is a productive and dynamic phenomenon (new metaphors arise as new events take place), such scalability and ongoing expansion of the Web make it an attractive corpus for metaphor research.

7. System Evaluation

System evaluation methodologies continue to be debated in many areas of computational semantics. Identifying a comprehensive and fair evaluation strategy for the task in mind is crucial for the development of fully functional NLP systems. In that light, a number of shared tasks have been proposed over the years at Workshops on Semantic Evaluation (SemEval) that enabled performance comparison across systems and methods. Such tasks as sentiment analysis, word similarity, word sense induction and disambiguation, coreference resolution, lexical substitution, and many others are commonly addressed at SemEval and have a number of benchmark data sets created for them (Agirre and Soroa 2007; McCarthy and Navigli 2007; Lefever and Hoste 2010; Manandhar et al. 2010; Mihalcea, Sinha, and McCarthy 2010; Recasens et al. 2010; Nakov et al. 2013), against which the systems are evaluated and compared. Computational work on metaphor, on the contrary, is considerably more fragmented than similar research efforts in other areas of NLP. With the lack of an established data set, the community has utilized a variety of evaluation strategies, including the use of annotated corpora, human judgments of system output, evaluation against the MML, and annotation of individual selected examples (usually phrases or sentences) via Amazon Mechanical Turk. With a few exceptions, the majority of approaches created their own test sets, making the results not directly comparable.

The most desirable type of evaluation is that conducted against an annotated full-text corpus, namely, naturally occurring, continuous text manually annotated for metaphor. Ideally, such a corpus should be open-domain and representative of a range of genres, making the results indicative of the likely performance of the system on arbitrary text. Another benefit of this type of evaluation is that it allows one to assess both the precision and the recall of the system. However, only two of the presented approaches (Dunn 2013b; Shutova 2013) conducted this type of evaluation, as shown in Table 7. More typically, approaches were instead evaluated on a random sample of system output against human judgments (Mason 2004; Shutova, Sun, and Korhonen 2010; Heintz et al. 2013; Shutova and Sun 2013; Strzalkowski et al. 2013). Although this type of evaluation allows one to measure the precision of the system on a random sample, it does not provide any information about the possible recall. Shutova, Sun, and Korhonen (2010) and Shutova and Sun (2013) applied their methods to a general-domain corpus (the BNC), from which a random sample of metaphorical expressions annotated by the system was then extracted for evaluation. In contrast, Heintz et al. (2013) and Strzalkowski et al. (2013) collected their data with a focus on a limited domain. The experiments of Mason (2004) and Shutova and Sun (2013) were concerned with conceptual metaphor, and a random sample of the metaphorical mappings identified by the systems was extracted and evaluated against human judgments in terms of precision. However, the latter two approaches also measured recall, by manually compiling a gold-standard of metaphorical mappings for the concepts of interest (Shutova and Sun 2013) or against the MML (Mason 2004).

Table 7
Identification systems: Evaluation set-up.

System	Annotated corpus	Continuous text	Human judgments	Individual selected examples	MML	AMT
Mason (2004)	–	–	✔	✔	✔	–
Birke and Sarkar (2006)	–	–	–	✔	–	–
Gedigian et al. (2006)	–	–	–	✔	–	–
Krishnakumaran and Zhu (2007)	–	–	–	✔	✔	–
Shutova et al. (2010)	–	✔	✔	–	–	–
Li and Sporleder (2010)	–	–	–	✔	–	–
Turney et al. (2011)	–	–	–	✔	–	–
Neuman et al. (2013)	–	–	–	✔	–	–
Dunn (2013)	✔	✔	–	–	–	–
Tsvetkov et al. (2013, 2014)	–	–	–	✔	–	–
Mohler et al. (2013)	–	–	–	✔	–	–
Heintz et al. (2013)	–	–	✔	–	–	✔
Hovy et al. (2013)	–	–	–	✔	–	✔
Wilks et al. (2013)	–	–	–	✔	–	–
Strzalkowski et al. (2013)	–	–	✔	–	–	✔
Shutova and Sun (2013)	–	✔	✔	–	–	–
Gandy et al. (2013)	–	–	–	✔	–	–
Li et al. (2013)	–	–	–	✔	✔	✔
Shutova (2013)	✔	✔	–	–	–	–

The majority of approaches (see Table 7) did not apply their systems to continuous text, but rather to a set of pre-selected examples (phrases, sentences, or occasionally paragraphs) in isolation from wider context. Such examples were annotated as metaphorical or literal by independent expert annotators or via Amazon Mechanical Turk. The benefit of this set-up (as opposed to the evaluation on a random sample of system output) is that it allows one to measure both precision and recall. However, the method used for selection of individual examples may introduce a bias into the evaluation and provide an unfair advantage to the system, unless the test sample was selected randomly from arbitrary text. In other words, this type of evaluation is likely to be less objective than the evaluation on continuous corpus text or a random sample.

A number of approaches (Gedigian et al. 2006; Krishnakumaran and Zhu 2007; Gandy et al. 2013; Heintz et al. 2013; Mohler et al. 2013; Neuman et al. 2013; Strzalkowski et al. 2013; Wilks et al. 2013) conducted their experiments within a limited domain. Despite allowing for an in-depth investigation of domain-specific patterns of metaphor use, such evaluations are problematic as they provide no indication of the scalability of the method beyond the studied domain to real-world data. This criticism also applies to the evaluations against MML, as the list is limited in domain coverage and the type of metaphors it provides.

Two data sets stand out as having been repeatedly adopted for metaphor research, enabling direct system comparison. These include the TroFi data set of Birke and Sarkar (2006) and the metaphor paraphrasing data set of Shutova (2010). The TroFi dataset consists of 25 verbs and example sentences, containing their metaphorical and literal use. It was adopted by Turney et al. (2011) and Tsvetkov, Mukomel, and Gershman (2013) in their metaphor identification experiments. The paraphrasing data set and gold standard of Shutova (2010) consists of 52 metaphorically used verbs and their human-derived literal (or more conventional) paraphrases in the given context. The data set

Table 8
Identification systems: Measures used and results.

System	Precision	Recall	F-score	Acc	Lim.	Open
Mason (2004)	–	–	–	0.77	–	✔
Birke and Sarkar (2006)	–	–	0.54	–	–	✔
Gedigian et al. (2006)	–	–	–	0.95*	✔	–
Krishnakumaran, Zhu (2007)	–	–	–	0.58	✔	–
Shutova et al. (2010)	0.79	–	–	–	–	✔
Li and Sporleder (2010)	–	–	0.75	0.78	–	✔
Turney et al. (2011)	–	–	0.68**	0.79**	–	✔
Neuman et al. (2013)	0.71	0.43–0.97	–	–	✔	–
Dunn (2013)	–	–	0.58	–	–	✔
Tsvetkov et al. (2013; 2014)	0.78	0.79	0.78	–	–	✔
Mohler et al. (2013)	0.56	0.93	0.7	–	✔	–
Heintz et al. (2013)	0.54	0.64	0.59	–	✔	–
Hovy et al. (2013)	0.7	0.8	0.75	0.75	–	✔
Wilks et al. (2013)	0.57	0.82	0.67	–	✔	–
Strzalkowski et al. (2013)	–	–	–	0.71	✔	–
Shutova and Sun (2013)	0.65 (LM); 0.69 (CM)	0.61 (CM)	–	–	–	✔
Gandy et al. (2013)	0.76 (LM); 0.65 (CM)	0.82 (LM)	–	–	✔	–
Li et al. (2013)	0.65–0.73	0.52–0.66	0.58–0.69	–	–	✔
Shutova (2013)	0.68	0.66	0.67	–	–	✔

* The results of Gedigian et al. (2006) should be interpreted with a reference to the performance of an *all metaphor* baseline attaining 0.92.
** Turney et al. (2011) report results on the verb dataset in terms of F-score and on the adjective dataset in terms of accuracy. LM stands for linguistic metaphor and CM for conceptual metaphor.

has been used in multiple metaphor interpretation experiments (Shutova 2010; Shutova, Van de Cruys, and Korhonen 2012; Bollegala and Shutova 2013).

The evaluations of metaphor identification tend to be conducted in terms of precision and recall, and occasionally, accuracy. Table 8 presents a summary of results of metaphor identification experiments, classified by domain coverage. The most successful systems attain an F-score in the range of 70–78%, with the highest precision reported for the methods of Tsvetkov, Mukomel, and Gershman (2013), Shutova, Sun, and Korhonen (2010), and Gandy et al. (2013); and the highest recall for the methods of Mohler et al. (2013), Wilks et al. (2013), Gandy et al. (2013) (limited domain) and Tsvetkov, Mukomel, and Gershman (2013) and Hovy et al. (2013) (open domain). However, because the methods were evaluated on data sets of different size and balance of categories, as well as created with different criteria in mind, this comparison is only approximate. A comprehensive evaluation on the same large data set is needed to determine the best performing techniques. For example, the accuracy of 95% reported by Gedigian et al. (2006) was measured on a data set dominated by metaphorical expressions (all metaphor baseline achieves 92%). This result cannot be directly compared to that of a system evaluated on a test set with a balance of metaphorical and literal instances.

One of the key difficulties metaphor processing research is facing today is that of a lack of annotated data. The data sets used in the experiments are typically too

small to obtain a generalizable result. Another issue is that many data sets created to evaluate individual methods were designed for a specific task, focusing on a particular type of metaphor and using an annotation scheme of their own. This makes the results not directly comparable and the overall performance landscape difficult to interpret. This situation calls for the consolidation of the insights gained from the current experiments into a single task definition and the creation of a large data set with this task in mind. Ideally, the systems should be evaluated on an expert-annotated corpus, containing continuous, general-domain text. In order to be indicative of the likely performance of the system on real-world data, the corpus needs to have a comprehensive coverage of registers and genres. Finally, a large corpus annotated for metaphor would enable reliable evaluation both in terms of precision and recall.

The work of Dunn (2013b) was exemplary in that he evaluated and compared four different methods on the same data set. Dunn used the VU Amsterdam Metaphor Corpus (Steen et al. 2010), which is currently the largest metaphor corpus in existence, containing approximately 200,000 words. However, because Steen and colleagues were interested in historical aspects of metaphor along with its use in modern language, the VU metaphor corpus contains a large proportion of lexicalized metaphors. Their meanings are ingrained in everyday use and can be interpreted via established techniques (e.g., word sense disambiguation), and their metaphorical nature may or may not be of interest to wider NLP. Whether metaphor processing systems should address highly conventional and dead metaphors depends on the task and application in mind, and corpus annotation should reflect this task definition in a consistent way. As the field moves forward, it would also be desirable to conduct extrinsic evaluations of the metaphor processing systems, in order to determine their usefulness for external NLP applications. One such experiment has already been carried out by Agerri (2008), who has demonstrated that metaphor interpretation plays an important role in textual entailment resolution.

8. Conclusion

Metaphor makes our thoughts move vivid and enriches our communication with novel imagery, but most importantly it plays a fundamental structural role in our cognition, helping us organize and project knowledge. As a result, its manifestations are pervasive in language and reasoning, making its computational processing an imperative task within NLP and intelligent systems engineering at large. Despite involving complex comparisons and information transfers, metaphor is a well-structured and systematic phenomenon, highly suitable for computational modeling. Focusing primarily on linguistic metaphor, the community has investigated a range of its aspects, implemented in a variety of system features. Among the most successful features are concreteness, distributional behavior of source and target domain vocabulary, selectional preferences, textual coherence, and topical properties of source and target words. The field has evolved from the widespread use of hand-coded knowledge to mainly data-driven research. Balanced corpora, the Web, Wikipedia and, sometimes, domain-specific corpora have become the primary source of knowledge for metaphor processing. The community has investigated supervised learning, clustering, topic modeling, and pattern-based search to acquire lexical, relational, and domain knowledge from these corpora. Yielding promising results, this was a significant advance in computational modeling of metaphor, allowing for the application of the systems to real-world data.

These experiments also provided new insights on the behavior of metaphor across domains, genres, and types of discourse.

The research on mining metaphorical associations from the data sheds light on how metaphors structure our conceptual system, as well as how specific conceptual metaphors are realized in language, revealing new information about their cognitive processing. Large-scale, automatic identification of conceptual metaphor thus provides a bridge between computational and cognitive research in this area, and has a wider scientific relevance beyond NLP. An interdisciplinary approach, leveraging knowledge from linguistics, cognitive science, psychology, neuroscience, and computer science, would be well-positioned to advance our understanding of metaphorical mechanisms and take the performance of metaphor processing systems to the next level. Two areas that are particularly likely to benefit from an interdisciplinary approach are metaphorical inference and extended metaphor, which have so far escaped attention in NLP. Recent advances in processing linguistic and conceptual metaphor, however, bring us a step closer to understanding and modeling these phenomena.

Despite the promising experimental results reported by the community, little attention has yet been given to real-world applications of metaphor processing. Possible applications include other semantic tasks within NLP and data mining, as well as social science and educational applications. Within NLP, most applications that need to access semantic knowledge would benefit from robust and accurate metaphor resolution. These include, for instance, machine translation, sentiment analysis, or text classification. Because the metaphors we use are known to be indicative of our underlying viewpoints, metaphor processing is likely to be fruitful in determining political affiliation from text or pinning down cross-cultural and cross-population differences, and thus become a useful tool in data mining. In social science, metaphor is extensively studied as a way to frame cultural and moral models, and to predict social choice (Landau, Sullivan, and Greenberg 2009; Thibodeau and Boroditsky 2011; Lakoff and Wehling 2012). Metaphor is also widely viewed as a creative tool. Its knowledge projection mechanisms help us to grasp new concepts and generate innovative ideas. This opens many avenues for the creation of computational tools that foster creativity (Veale 2011, 2014) and support assessment in education (Burstein et al. 2013).

The design of the metaphor processing task should thus be informed by the possible applications. The application in mind may place particular requirements on the types of metaphor the system needs to address and the output representations it is expected to produce. Whereas an NLP application, such as machine translation, would be primarily concerned with linguistic metaphors and, possibly, the more creative instances thereof, a data mining application, aiming to detect a set of trends, may find the identification of conceptual metaphors prominent in the data more informative. The formulation of the task and experimental design, in turn, predetermine how system performance is best evaluated. So far, the lack of a common task definition and a shared data set have hampered our progress as a community in this area. This calls for a unification of the task definition and a large-scale annotation effort that would provide a data set for metaphor system evaluation, built with the insights gained from the present studies. The main purpose of this paper is to provide a platform for debate that would assist us in formulating the overall goals of metaphor processing and devising an optimal experimental strategy, enabling us as a community to make significant progress in this important and fascinating area.

References

Agerri, Rodrigo. 2008. Metaphor in textual entailment. In *Proceedings of COLING 2008*, pages 3–6, Manchester, UK.

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.

Agirre, Eneko and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague.

Allen, James F., Mary Swift, and Will de Beaumont. 2008. Deep semantic analysis of text. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pages 343–354, Venice.

Badryzlova, Yulia, Natalia Shekhtman, Yekaterina Isaeva, and Ruslan Kerimov. 2013. Annotating a Russian corpus of conceptual metaphor: A bottom–up approach. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 77–86, Atlanta, GA.

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

Baumer, Eric, Bill Tomlinson, and Lindsey Richland. 2009. Computational metaphor identification: A method for identifying conceptual metaphors in written text. In *Proceedings of Analogy '09*, pages 20–29, Sofia.

Beigman Klebanov, Beata and Eyal Beigman. 2010. A game-theoretic model of metaphorical bargaining. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 698–709, Uppsala.

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

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.

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

Bollegala, Danushka and Ekaterina Shutova. 2013. Metaphor interpretation using paraphrases extracted from the web. *PLoS ONE*, 8(9):e74304.

Broadwell, George Aaron, Jennifer Stromer-Galley, Tomek Strzalkowski, Samira Shaikh, Sarah M. Taylor, Ting Liu, Umit Boz, Alana Elia, Laura Jiao, and Nick Webb. 2013. Modeling sociocultural phenomena in discourse. *Natural Language Engineering*, 19(2):213–257.

Burnard, Lou. 2007. *Reference Guide for the British National Corpus (XML Edition)*. Available from `www.natcorp.ox.ac.uk/corpus/babyinfo.html`.

Burstein, Jill, John Sabatini, Jane Shore, Brad Moulder, and Jennifer Lentini. 2013. A user study: Technology to increase teachers' linguistic awareness to improve instructional language support for English language learners. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 1–10, Atlanta, GA.

Bzdek, Vincent. 2008. *Woman of the House: The Rise of Nancy Pelosi*. Palgrave Macmillan, Basingstoke, New York.

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

Carbonell, Jaime. 1982. Metaphor: An inescapable phenomenon in natural language comprehension. In Wendy Lehnert and Martin Ringle, editors, *Strategies for Natural Language Processing*. Lawrence Erlbaum, pages 415–434.

Charteris-Black, Jonathan. 2000. Metaphor and vocabulary teaching in ESP economics. *English for Specific Purposes*, 19(2):149–165.

Charteris-Black, Jonathan and Timothy Ennis. 2001. A comparative study of metaphor in Spanish and English financial reporting. *English for Specific Purposes*, 20(3):249–266.

Chung, Siaw-Fong, Kathleen Ahrens, and Chu-Ren Huang. 2005. Source domains as concept domains in metaphorical expressions. *International Journal of Computational Linguistics and Chinese Language Processing*, 10(4):553–570.

Cilibrasi, Rudi L. and Paul M. B. Vitanyi. 2007. The Google similarity distance. *IEEE Transactional on Knowledge and Data Engineering*, 19(3):370–383.

Cortes, Corinna and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Davies, Mark. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Deignan, A. and L. Potter. 2004. A corpus study of metaphors and metonyms in English and Italian. *Journal of Pragmatics*, 36:1231–1252.

Desalle, Yann, Bruno Gaume, and Karine Duvignau. 2009. Slam: Solutions lexicales automatique pour métaphores. *Traitement Automatique des Langues*, 50(1):145–175.

Diaz-Vera, Javier and Rosario Caballero. 2013. Exploring the feeling-emotions continuum across cultures: Jealousy in English and Spanish. *Intercultural Pragmatics*, 10(2):265–294.

Dunn, Jonathan. 2011. Gradient semantic intuitions of metaphoric expressions. *Metaphor and Symbol*, 26(1):53–67.

Dunn, Jonathan. 2013a. Evaluating the premises and results of four metaphor identification systems. In *Proceedings of CICLing'13*, pages 471–486, Samos.

Dunn, Jonathan. 2013b. What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 1–10, Atlanta, GA.

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

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.

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

Francis, W. Nelson and Henry Kucera. 1979. A standard corpus of present-day edited American English. Technical report, Brown University, Providence, RI. Available from `www.helsinki.fi/varieng/CoRD/corpora/BROWN`.

Gandy, Lisa, Nadji Allan, Mark Atallah, Ophir Frieder, Newton Howard, Sergey Kanareykin, Moshe Koppel, Mark Last, Yair Neuman, and Shlomo Argamon. 2013. Automatic identification of conceptual metaphors with limited knowledge. In *Proceedings of AAAI 2013*, pages 328–334, Bellevue, WA.

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.

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

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

Gong, Shu-Ping, Kathleen Ahrens, and Chu-Ren Huang. 2008. Chinese word sketch and mapping principles: A corpus-based study of conceptual metaphors using the building source domain. *International Journal of Computer Processing of Oriental Languages*, 21(2):3–17.

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

Hardie, Andrew, Veronika Koller, Paul Rayson, and Elena Semino. 2007. Exploiting a semantic annotation tool for metaphor analysis. In *Proceedings of the Corpus Linguistics Conference*, pages 1–12, Birmingham, UK.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Nantes.

Heintz, Ilana, Ryan Gabbard, Mahesh Srivastava, Dave Barner, Donald Black, Majorie Friedman, and Ralph Weischedel. 2013. Automatic extraction of linguistic metaphors with lda topic modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66, Atlanta, GA.

Hesse, Mary. 1966. *Models and Analogies in Science*. Notre Dame University Press, South Bend, NA.

Hobbs, Jerry R. 1981. Metaphor interpretation as selective inferencing. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'81, pages 85–91, Vancouver.

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

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: Norwood, NJ.
Hovy, Dirk, Shashank Shrivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57, Atlanta, GA.
Izwaini, Sattar. 2003. Corpus-based study of metaphor in information technology. In *Proceedings of the Workshop on Corpus-based Approaches to Figurative Language, Corpus Linguistics 2003*, pages 1–8, Lancaster.
Jamrozik, Anja, Eyal Sagi, Micah Goldwater, and Dedre Gentner. 2013. Relational words have high metaphoric potential. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 21–26, Atlanta, GA.
Karov, Yael and Shimon Edelman. 1998. Similarity-based word sense disambiguation. *Computational Linguistics*, 24(1):41–59.
Kingsbury, Paul and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of LREC-2002*, pages 1989–1993, Gran Canaria.
Koller, Veronika. 2004. *Metaphor and Gender in Business Media Discourse: A Critical Cognitive Study*. Palgrave Macmillan, Basingstoke, New York.
Korkontzelos, Ioannis, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, GA.
Kovecses, Z. 2005. *Metaphor in Culture: Universality and Variation*. Cambridge University Press, Cambridge.
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.
Lakoff, George. 2008. *The Political Mind*. Viking, New York.
Lakoff, George, Jane Espenson, and Alan Schwartz. 1991. The master metaphor list. Technical report, University of California at Berkeley. Available from `araw.mede.uic.edu/valansz/metaphor/METAPHORLIST.pdf`.
Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.
Lakoff, George and Elisabeth Wehling. 2012. *The Little Blue Book: The Essential Guide to Thinking and Talking Democratic*. Free Press, New York.
Landau, Mark J., Daniel Sullivan, and Jeff Greenberg. 2009. Evidence that self-relevant motives and metaphoric framing interact to influence political and social attitudes. *Psychological Science*, 20(11):1421–1427.
Lefever, Els and Véronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala.
Lewis, David D., Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.
Li, Hongsong, Kenny Q. Zhu, and Haixun Wang. 2013. Data-driven metaphor recognition and explanation. *Transactions of the Association for Computational Linguistics*, 1:379–390.
Li, Linlin and Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 315–323.
Li, Linlin and Caroline Sporleder. 2010. Using Gaussian mixture models to detect figurative language in context. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 297–300, Singapore.
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-Rodman, Birte. 2008. The Hamburg metaphor database project: Issues in resource creation. *Language Resources and Evaluation*, 42(3):293–318.
Low, Graham, Zazie Todd, Alice Deignan, and Lynne Cameron. 2010. *Researching and Applying Metaphor in the Real World*. John Benjamins, Amsterdam/Philadelphia.
Lu, Louis and Kathleen Ahrens. 2008. Ideological influences on building metaphors in Taiwanese presidential

speeches. *Discourse and Society*, 19(3):383–408.
Manandhar, Suresh, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala.
Martin, James. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press Professional, Inc., San Diego, CA.
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.
Mason, Zachary. 2004. Cormet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.
McCarthy, Diana and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague.
Mihalcea, Rada, Ravi Sinha, and Diana McCarthy. 2010. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 9–14, Uppsala.
Mohler, Michael, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35, Atlanta, GA.
Moschitti, Ro, Daniele Pighin, and Roberto Basili. 2006. Tree kernel engineering for proposition re-ranking. In *Proceedings of Mining and Learning with Graphs (MLG)*, pages 165–172, Berlin.
Musolff, Andreas. 2000. *Mirror Images of Europe: Metaphors in the Public Debate about Europe in Britain and Germany*. Iudicium, Muenchen.
Nakov, Preslav, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, GA.
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.
Neuman, Yair, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. 2013. Metaphor identification in large texts corpora. *PLoS ONE*, 8(4):e62343.
Niculae, Vlad and Victoria Yaneva. 2013. Computational considerations of comparisons and similes. In *Proceedings of ACL (Student Research Workshop)*, pages 89–95, Sophia.
Niles, Ian and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, New York, NY.
Niles, Ian and Adam Pease. 2003. Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology. In *Proceedings of the International Conference on Information and Knowledge Engineering*, pages 412–416, Las Vegas, NV.
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.
Peters, Wim and Ivonne Peters. 2000. Lexicalised systematic polysemy in WordNet. In *Proceedings of LREC 2000*, pages 1–7, Athens.
Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22:1–39.
Recasens, Marta, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 1–8, Uppsala.
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.
Resnik, Philip. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Rohrer, Tim. 1997. Conceptual blending on the information highway: How metaphorical inferences work. In W. Liebert, G. Redeker, and L. Waugh, editors, *Discourse and Perspective in Cognitive Linguistics*. John Benjamins Publishing Company, Amsterdam/ Philadelphia, pages 185–204.

Sandhaus, Evan. 2008. *The New York Times Annotated Corpus*. Available from `https://catalog/lde.upenn.edu/LDC2008T19`.

Santa Ana, Otto. 1999. Like an animal I was treated?: Anti-immigrant metaphor in US public discourse. *Discourse Society*, 10(2):191–224.

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

Shutova, Ekaterina. 2013. Metaphor identification as interpretation. In *Proceedings of *SEM 2013*, pages 276–285, Atlanta, GA.

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

Shutova, Ekaterina, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of Conference on Computational Linguistics 2010*, pages 1002–1010, Association for Computational Linguistics. Beijing.

Shutova, Ekaterina and Simone Teufel. 2010. Metaphor corpus annotated for source–target domain mappings. In *Proceedings of LREC 2010*, pages 3255–3261, Malta.

Shutova, Ekaterina, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Shutova, Ekaterina, Tim Van de Cruys, and Anna Korhonen. 2012. Unsupervised metaphor paraphrasing using a vector space model. In *Proceedings of COLING 2012*, pages 1121–1130, Mumbai.

Skorczynska Sznajder, Hanna and Jordi Pique-Angordans. 2004. A corpus-based description of metaphorical marking patterns in scientific and popular business discourse. In *Proceedings of European Research Conference on Mind, Language and Metaphor (Euresco Conference)*, pages 112–129, Granada.

Steen, Gerard J., Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins, Amsterdam/Philadelphia.

Stefanowitsch, A. 2004. Happiness in English and German: A metaphorical-pattern analysis. In M. Altenberg and S. Kemmer, editors, *Language, Culture, and Mind*. CSLI Publications Stanford, CA.

Strzalkowski, Tomek, George Aaron Broadwell, Sarah Taylor, Laurie Feldman, Samira Shaikh, Ting Liu, Boris Yamrom, Kit Cho, Umit Boz, Ignacio Cases, and Kyle Elliot. 2013. Robust extraction of metaphor from novel data. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 67–76, Atlanta, GA.

Sullivan, Karen. 2007. *Grammar in metaphor: A construction grammar account of metaphoric language*. Ph.D. thesis, University of California, Berkeley.

Sullivan, Karen. 2013. *Frames and constructions in metaphoric language*. John Benjamins, Amsterdam.

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

Thibodeau, Paul H. and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PLoS ONE*, 6(2):e16782, 02.

Tsvetkov, Yulia, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, MD.

Tsvetkov, Yulia, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, Atlanta, GA.

Turner, Mark and Gilles Fauconnier. 2003. Metaphor, metonymy, and binding. In A. Barcelona, editor, *Metaphor and Metonymy at the Crossroads: A Cognitive Perspective*. Walter de Gruyter GmbH, Berlin, pages 133–145.

Turney, Peter D., Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*,

EMNLP '11, pages 680–690, Stroudsburg, PA.

Veale, Tony. 2011. Creative language retrieval: A robust hybrid of information retrieval and linguistic creativity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 278–287, Portland, OR.

Veale, Tony. 2014. A service-oriented architecture for metaphor processing. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 52–60, Baltimore, MD.

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, UK.

Wallington, A. M., J. A. Barnden, P. Buchlovsky, L. Fellows, and S. R. Glasbey. 2003. Metaphor annotation: A systematic study. Technical Report CSRP-03-04, School of Computer Science, University of Birmingham.

Wikberg, Kay. 2006. The role of corpus studies in metaphor research. In *2006 Stockholm Metaphor Festival*, pages 33–48, Stockholm.

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.

Wilks, Yorick, Adam Dalton, James Allen, and Lucian Galescu. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 36–44, Atlanta, GA.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2nd edition.