Multilingual Metaphor Processing: Experiments with Semi-Supervised and Unsupervised Learning

Ekaterina Shutova[*]
University of Cambridge

Lin Sun[**]
Greedy Intelligence

Elkin Darío Gutiérrez[†]
University of California, San Diego

Patricia Lichtenstein[‡]
University of California, Merced

Srini Narayanan[§]
Google Research

Highly frequent in language and communication, metaphor represents a significant challenge for Natural Language Processing (NLP) applications. Computational work on metaphor has traditionally evolved around the use of hand-coded knowledge, making the systems hard to scale. Recent years have witnessed a rise in statistical approaches to metaphor processing. However, these approaches often require extensive human annotation effort and are predominantly evaluated within a limited domain. In contrast, we experiment with weakly supervised and unsupervised techniques—with little or no annotation—to generalize higher-level mechanisms of metaphor from distributional properties of concepts. We investigate different levels and types of supervision (learning from linguistic examples vs. learning from a given set of metaphorical mappings vs. learning without annotation) in flat and hierarchical, unconstrained and constrained clustering settings. Our aim is to identify the optimal type of supervision for a learning algorithm that discovers patterns of metaphorical association from text. In order to investigate

[*] Computer Laboratory, William Gates Building, Cambridge CB3 0FD, UK. E-mail: es407@cam.ac.uk.
[**] Greedy Intelligence Ltd, Hangzhou, China. E-mail: lin.sun@greedyint.com.
[†] Department of Cognitive Science, 9500 Gilman Dr, La Jolla, CA 92093, USA. E-mail: e4gutier@ucsd.edu.
[‡] Department of Cognitive and Information Sciences, UC Merced, 5200 Lake Rd Merced, CA 95343, USA. E-mail: plichtenstein@ucmerced.edu.
[§] Google, Brandschenkestrasse 110, 8002 Zurich, Switzerland. E-mail: srinin@google.com.

Submission received: 28 September 2015; revised version received: 19 February 2016; accepted for publication: 29 May 2016.

doi:10.1162/COLI_a_00275

the scalability and adaptability of our models, we applied them to data in three languages from different language groups—English, Spanish, and Russian—achieving state-of-the-art results with little supervision. Finally, we demonstrate that statistical methods can facilitate and scale up cross-linguistic research on metaphor.

1. Introduction

Metaphor brings vividness, distinction, and clarity to our thought and communication. At the same time, it plays an important structural role in our cognition, helping us to organize and project knowledge (Lakoff and Johnson 1980; Feldman 2006) and guide our reasoning (Thibodeau and Boroditsky 2011). Metaphors arise from systematic associations between distinct, and seemingly unrelated, concepts. For instance, when we talk about "the *turning wheels* of a political regime," "*rebuilding* the campaign *machinery*" or "*mending* foreign policy," we view *politics* and *political systems* in terms of *mechanisms*—they can *function, break, be mended, 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 in a variety of metaphorical expressions. The view of metaphor as a mapping between two distinct domains was echoed by numerous theories in the field (Black 1962; Hesse 1966; Lakoff and Johnson 1980; Gentner 1983). The most influential of these was the Conceptual Metaphor Theory of Lakoff and Johnson (1980). Lakoff and Johnson claimed that metaphor is not merely a property of language, but rather a cognitive mechanism that structures our conceptual system in a certain way. They coined the term *conceptual metaphor* to describe the mapping between the target concept (e.g., *politics*) and the source concept (e.g., *mechanism*), and *linguistic metaphor* to describe the resulting metaphorical expressions. Other examples of common metaphorical mappings include: TIME IS MONEY (e.g., "That flat tire *cost* me an hour"); IDEAS ARE PHYSICAL OBJECTS (e.g., "I can not *grasp* his way of thinking"); VIOLENCE IS FIRE (e.g., "violence *flares* amid curfew"); 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*").

Manifestations of metaphor are pervasive in language and reasoning, making its computational processing an imperative task within Natural Language Processing (NLP). Explaining up to 20% of all word meanings according to corpus studies (Shutova and Teufel 2010; Steen et al. 2010), metaphor is currently a bottleneck, particularly in semantic tasks. An accurate and scalable metaphor processing system would become an important component of many practical NLP applications. These include, for instance, machine translation (MT): A large number of metaphorical expressions are culture-specific and therefore represent a considerable challenge in translation (Schäffner 2004; Zhou, Yang, and Huang 2007). Shutova, Teufel, and Korhonen (2013) conducted a study of metaphor translation in MT. Using Google Translate,[1] a state-of-the-art MT system, they found that as many as 44% of metaphorical expressions in their data set were translated incorrectly, resulting in semantically infelicitous sentences. A metaphor

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

processing component could help to avoid such errors. Other applications of metaphor processing include, for instance, opinion mining: metaphorical expressions tend to contain a strong emotional component (e.g., compare the metaphor "Government *loosened its stranglehold* on business" and its literal counterpart "Government deregulated business" [Narayanan 1999]); or information retrieval: 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 the metaphors we use are also 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).

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; 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; Beigman Klebanov et al. 2014; Mohler et al. 2014), 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 classification (Gedigian et al. 2006; Dunn 2013a; Hovy et al. 2013; Mohler et al. 2013; Tsvetkov, Mukomel, and Gershman 2013), unsupervised learning (Heintz et al. 2013; Shutova and Sun 2013), distributional approaches (Shutova 2010; Shutova, Van de Cruys, and Korhonen 2012; Shutova 2013; Mohler et al. 2014), 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 using lexico-syntactic patterns (Veale and Hao 2008; Bollegala and Shutova 2013; Li, Zhu, and Wang 2013). However, even the statistical methods have been predominantly applied in limited-domain, small-scale experiments. This is mainly due to the lack of general-domain corpora annotated for metaphor that are sufficiently large for training wide-coverage supervised systems. In addition, supervised methods tend to rely on lexical resources and ontologies for feature extraction, which limits the robustness of the features themselves and makes the methods dependent on the coverage (and the availability) of these resources. This also makes these methods difficult to port to new languages, for which such lexical resources or corpora may not exist. In contrast, we experiment with minimally supervised and unsupervised learning methods that require little or no annotation; and use robust, dynamically mined lexico-syntactic features that are well suited for metaphor processing. This makes our methods scalable to new data and portable across languages, domains, and tasks, bringing metaphor processing technology a step closer to a possibility of integration with real-world NLP.

Our methods use distributional clustering techniques to investigate how metaphorical cross-domain mappings partition the semantic space in three different languages—English, Russian, and Spanish. In a distributional semantic space, each word is represented as a vector of contexts in which it occurs in a text corpus.[2] Because of the high frequency and systematicity with which metaphor is used in language, it is naturally and systematically reflected in the distributional space. As a result of metaphorical cross-domain mappings, the words' context vectors tend to be non-homogeneous in structure and to contain vocabulary from different domains. For instance, the context vector for the noun *idea* would contain a set of literally used terms (e.g., *understand* [an idea]) and a set of metaphorically used terms, describing *ideas* as PHYSICAL OBJECTS (e.g., *grasp* [an idea], *throw* [an idea]), LIQUIDS (e.g., [ideas] *flow*), or FOOD (e.g., *digest* [an idea]), and so on. Similarly, the context vector for *politics* would contain MECHANISM terms (e.g., *operate* or *refuel* [politics]), GAME terms (e.g., *play* or *dominate* [politics]), SPACE terms (e.g., *enter* or *leave* [politics]), as well as the literally used terms (e.g., *explain* or *understand* [politics]), as shown in Figure 1. This demonstrates how metaphorical usages, abundant in the data, structure the distributional space. As a result, the context vectors of different concepts contain a certain degree of cross-domain overlap, thus implicitly encoding cross-domain mappings. Figure 1 shows such a term overlap in the direct object vectors for the concepts of GAME and POLITICS. We exploit such composition of the context vectors to induce information about metaphorical mappings directly from the words' distributional behavior in an unsupervised or a minimally supervised way. We then use this information to identify metaphorical language. Clustering methods model modularity in the structure of the semantic space, and thus naturally provide a suitable framework to capture metaphorical information. To our knowledge, the metaphorical cross-domain structure of the distributional space has not yet been explicitly exploited in wider NLP. Instead, most NLP approaches tend to treat all types of distributional features as identical, thus possibly losing important conceptual information that is naturally encoded in the distributional semantic space.

The focus of our experiments is on the identification of metaphorical expressions in verb–subject and verb–object constructions, where the verb is used metaphorically. In the first set of experiments, we apply a flat clustering algorithm, spectral clustering (Ng et al. 2002), to learn metaphorical associations from text. The system clusters verbs and nouns to create representations of source and target domains. The verb clustering is used to harvest source domain vocabulary and the noun clustering is used to identify groups of target concepts associated with the same source. For instance, the nouns *democracy* and *marriage* are clustered together (in the target noun cluster), because both are metaphorically associated with (for example) *mechanisms* or *games* and, as such, appear with *mechanism* and *game* terms in the corpus (the source verb cluster). The obtained clusters represent source and target concepts between which metaphorical associations hold. We first experiment with the unconstrained version of spectral clustering using the method of Shutova, Sun, and Korhonen (2010), where metaphorical patterns are derived from the distributional information alone and the clustering process is fully unsupervised. We then extend this method to perform constrained clustering, where a small number of example metaphorical mappings are used to guide the learning process, with the expectation of changing the cluster structure towards capturing metaphorically associated concepts. We then analyze and compare

[2] In our experiments we use a syntax-aware distributional space, where the vectors are constructed using the words' grammatical relations.

N: game	N: politics
1170 **play**	31 **dominate**
202 **win**	30 **play**
99 miss	28 **enter**
76 watch	16 **discuss**
66 **lose**	13 **leave**
63 start	12 **understand**
42 enjoy	8 **study**
22 finish	6 **explain**
...	5 **shape**
20 **dominate**	4 **influence**
18 quit	4 **change**
17 host	4 **analyse**
17 follow	...
17 control	2 **transform**
...	...

Figure 1
Context vectors for *game* and *politics* (verb–direct object relations) extracted from the British National Corpus. The context vectors demonstrate how metaphor structures the distributional semantic space through cross-domain vocabulary projection.

the structure of the clusters obtained with or without the use of constraints. The learning of metaphorical associations is then boosted from a small set of example metaphorical expressions that are used to connect the verb and noun clusters. Finally, the acquired set of associations is used to identify new, unseen metaphorical expressions in a large corpus.

Although we believe that these methods would capture a substantial amount of information about metaphorical associations from distributional properties of concepts, they are still dependent on the seed expressions to identify new metaphorical language. In our second set of experiments, we investigate to what extent it is possible to acquire information about metaphor from distributional properties of concepts alone, without any need for labeled examples. For this purpose, we apply the hierarchical clustering method of Shutova and Sun (2013) to identify both metaphorical associations and metaphorical expressions in a fully unsupervised way. We use hierarchical graph factorization clustering (Yu, Yu, and Tresp 2006) of nouns to create a network (or a graph) of concepts and to quantify the strength of association between concepts in this graph. The metaphorical mappings are then identified based on the association patterns between concepts in the graph. The mappings are represented as cross-level, one-directional connections between clusters in the graph. The system then uses salient features of the metaphorically connected clusters to identify metaphorical expressions in text. Given a source domain, the method outputs a set of target concepts associated with this source, as well as the corresponding metaphorical expressions.

We then compare the ability of these methods (that require different kinds and levels of supervision) to identify metaphor. In order to investigate the scalability and adaptability of the methods, we applied them to unrestricted, general-domain text in three typologically different languages—English, Spanish, and Russian. We evaluated the performance of the systems with the aid of human judges in precision- and recall-oriented settings, achieving state-of-the-art results with little supervision. Finally, we analyze the differences in the use of metaphor across languages, as discovered by the systems, and demonstrate that statistical methods can facilitate and scale up cross-linguistic research on metaphor.

2. Related Work

2.1 Metaphor Annotation Studies

Metaphor annotation studies have typically been corpus-based and involved either continuous annotation of metaphorical language (i.e., distinguishing between literal and metaphorical uses of words in a given text), or search for instances of a specific metaphor in a corpus and an analysis thereof. 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; Hardie et al. 2007; Lu and Ahrens 2008; Low et al. 2010), or in a particular genre or type of discourse, such as financial (Charteris-Black and Ennis 2001; Martin 2006), political (Lu and Ahrens 2008), or educational (Cameron 2003; Beigman Klebanov and Flor 2013) discourse.

Two studies (Steen et al. 2010; Shutova and Teufel 2010) moved away from investigating particular domains to a more general study of how metaphor behaves in unrestricted continuous text. Steen and colleagues (Pragglejaz Group 2007; Steen et al. 2010) proposed a metaphor identification procedure (MIP), in which every word is tagged as literal or metaphorical, based on whether it has a "more basic meaning" in other contexts than the current one. The basic meaning was defined as "more concrete; related to bodily action; more precise (as opposed to vague); historically older" and its identification was guided by dictionary definitions. The resulting VU Amsterdam Metaphor Corpus[3] is a 200,000-word subset of the British National Corpus (BNC) (Burnard 2007) annotated for linguistic metaphor. The corpus has already found application in computational metaphor processing research (Dunn 2013b; Niculae and Yaneva 2013; Beigman Klebanov et al. 2014), as well as inspiring metaphor annotation efforts in other languages (Badryzlova et al. 2013). Shutova and Teufel (2010) extended MIP 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. Shutova and Teufel's corpus is a 13,000-word subset of the BNC sampling a range of genres, and it has served as a testbed in a number of computational experiments (Shutova 2010; Shutova, Sun, and Korhonen 2010; Bollegala and Shutova 2013).

Lönneker (2004) investigated metaphor annotation in lexical resources. The resulting Hamburg Metaphor Database contains examples of metaphorical expressions in German and French, which are mapped to senses from EuroWordNet[4] and annotated with source-target domain mappings.

2.2 Computational Approaches to Metaphor Identification

Early computational work on metaphor tended to be theory-driven and utilized hand-coded descriptions of concepts and domains to identify and interpret metaphor.

[3] http://www.ota.ox.ac.uk/headers/2541.xml.

[4] http://www.illc.uva.nl/EuroWordNet/.

The system of Fass (1991), for instance, was an implementation of the selectional preference violation view of metaphor (Wilks 1978) and detected metaphor and metonymy as a violation of a common preference of a predicate by a given argument. Another branch of approaches (Martin 1990; Narayanan 1997; Barnden and Lee 2002) implemented some aspects of the conceptual metaphor theory (Lakoff and Johnson 1980), reasoning over hand-crafted representations of source and target domains. The system of Martin (1990) explained linguistic metaphors through finding the corresponding metaphorical mapping. The systems of Narayanan (1997) and Barnden and Lee (2002) performed inferences about entities and events in the source and target domains in order to interpret a given metaphor. The reasoning processes relied on manually coded 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 reliance on task- and domain-specific hand-coded knowledge makes these systems difficult to scale to real-world text. Later research thus turned to general-domain lexical resources and ontologies, as well as statistical methods, in order to design more scalable solutions. Mason (2004) introduced the use of statistical techniques for metaphor processing; however, his approach had a considerable reliance on WordNet (Fellbaum 1998). His CorMet system discovered source–target domain mappings automatically, by searching for systematic variations in domain-specific verb preferences. For example, *pour* is a characteristic verb in both LAB and FINANCE domains. In the LAB domain it has a strong preference for *liquids* and in the FINANCE domain for *money*. From this information, Mason's system inferred the domain mapping FINANCE–LAB and the concept mapping *money–liquid*. The system of Krishnakumaran and Zhu (2007) used hyponymy relations in WordNet and word bigram counts to predict verbal, nominal, and adjectival metaphors. For instance, given an IS-A construction (e.g., "The world is a *stage*") the system verified that the two nouns were in hyponymy relation in WordNet, and if this was not the case the expression was tagged as metaphorical. Given a verb–noun or an adjective–noun pair (such as "*planting* ideas" or "*fertile* imagination"), the system computed the bigram probability of this pair (including the hyponyms/hypernyms of the noun) and if the combination was not observed in the data with sufficient frequency, it was tagged as metaphorical.

These systems have demonstrated that statistical methods, when combined with broad-coverage lexical resources, can be successfully used to model at least some aspects of metaphor, increasing the system coverage. As statistical NLP, lexical semantics, and lexical acquisition techniques developed over the years, it has become possible to build larger-scale statistical metaphor processing systems that promise a step forward both in accuracy and robustness. Numerous approaches (Li and Sporleder 2010; Shutova 2010; Turney et al. 2011; Hovy et al. 2013; Shutova and Sun 2013; Shutova, Teufel, and Korhonen 2013; Tsvetkov, Mukomel, and Gershman 2013) used machine learning and statistical techniques to address a wider range of metaphorical language in general-domain text. For instance, the method of Turney et al. (2011) classified verbs and adjectives as literal or metaphorical based on their level of concreteness or abstractness in relation to the noun they appear with. They learned concreteness rankings for words automatically (starting from a set of examples) and then searched for expressions where a concrete adjective or verb was used with an abstract noun (e.g., "*dark* humor" was tagged as a metaphor and *dark hair* was not). The method of Turney et al. (2011) has served as a foundation for the later approaches of Neuman et al. (2013) and Gandy et al. (2013), who extended it through the use of selectional preferences and the identification of source domains, respectively.

Another branch of research focused on applying statistical learning to the problem of metaphor identification (Gedigian et al. 2006; Shutova, Sun, and Korhonen 2010; Dunn 2013a; Heintz et al. 2013; Hovy et al. 2013; Mohler et al. 2013; Shutova and Sun 2013; Tsvetkov, Mukomel, and Gershman 2013; Beigman Klebanov et al. 2014). The learning techniques they have investigated include supervised classification, clustering, and Latent Dirichlet Allocation (LDA) topic modeling. We review these methods in more detail subsequently.

2.2.1 Metaphor Identification as Supervised Classification. A number of approaches trained classifiers on manually annotated data to recognize metaphor (Gedigian et al. 2006; Dunn 2013a; Hovy et al. 2013; Mohler et al. 2013; Tsvetkov, Mukomel, and Gershman 2013; Beigman Klebanov et al. 2014). The method of Gedigian et al. (2006), for instance, discriminated between literal and metaphorical uses of the verbs of MOTION and CURE using a maximum entropy classifier. The authors obtained their data by extracting the lexical items whose frames are related to MOTION and CURE from FrameNet (Fillmore, Johnson, and Petruck 2003). To construct their training and test sets, they searched the PropBank *Wall Street Journal* corpus (Kingsbury and Palmer 2002) for sentences containing such lexical items and manually annotated them for metaphoricity. They used PropBank annotation (arguments and their semantic types) as features to train the classifier and reported an accuracy of 95.12%. This result was, however, only a little higher than the performance of the naive baseline assigning majority class to all instances (92.90%).

Dunn (2013a, 2013b) presented an ontology-based domain interaction approach that identified metaphorical expressions at the utterance level. Dunn's system first mapped the lexical items in the given utterance to concepts from SUMO ontology (Niles and Pease 2001, 2003), assuming that each lexical item was used in its default sense—that is, no sense disambiguation was performed. The system then extracted the properties of concepts from the ontology, such as their domain type (ABSTRACT, PHYSICAL, SOCIAL, MENTAL) and event status (PROCESS, STATE, OBJECT). Those properties were then combined into feature-vector representations of the utterances. Dunn trained a logistic regression classifier using these features to perform metaphor identification, reporting an F-score of 0.58 on general-domain data.

Tsvetkov, Mukomel, and Gershman (2013) experimented with metaphor 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 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*]). The authors used a logistic regression classifier and a combination of the listed features to annotate metaphor at the sentence level. The model was trained on the TroFi data set (Birke and Sarkar 2006) of 1,298 sentences containing literal and metaphorical uses of 25 verbs. Tsvetkov and colleagues evaluated their method on self-constructed data sets of 98 sentences for English and 140 sentences for Russian, attaining F-scores of 0.78 and 0.76, respectively. 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 is likely to focus on shallow properties of metaphorical language and to bypass conceptual information. Corpus-linguistic

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

research (Charteris-Black and Ennis 2001; Kovecses 2005; Diaz-Vera and Caballero 2013) suggests that there is considerable variation in metaphorical language across cultures, which makes training only on one language and translating the model problematic for modeling conceptual structure behind metaphor.

The approach of Mohler et al. (2013) relied on the concept of semantic signature of a text, defined 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 intuition behind this approach was that the texts whose semantic signature closely matched the signature of a known metaphor would be likely to contain 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 maximum entropy classifier, unpruned decision tree classifier, support vector machines, 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 F-score of 0.70 using the decision tree classifier.

Hovy et al. (2013) trained a support vector machine classifier (Cortes and Vapnik 1995) with tree kernels (Moschitti, Pighin, and Basili 2006) to capture the compositional properties of metaphorical language. Their hypothesis was that unusual semantic compositions in the data would be indicative of the use of metaphor. The system was trained 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. Hovy et al. used word vectors, as well as lexical, part-of-speech tags and WordNet supersense representations of sentence trees as features. They report encouraging results, F-score = 0.75, which is an indication of the importance of syntactic information and compositionality in metaphor identification.

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 described approaches experimented with a number of features, including lexical and syntactic information and higher-level features such as semantic roles, WordNet supersenses, and domain types extracted from ontologies. The results that came out of these studies suggest that in order to reliably capture the patterns of the use of metaphor in the data on a large scale, one needs to address conceptual properties of metaphor, along with the surface ones. Thus the model would need to make generalizations at the level of metaphorical mappings and coarse-grained classes of concepts, in essence representing different domains (such as *politics* or *machines*). Although our intention in this article is to model such domain structure in a minimally supervised or unsupervised way and to learn it from the data directly, the clusters produced by our models provide a representation of conceptual domains that could also be a useful feature within a supervised classification framework.

2.2.2 The Use of Clustering for Metaphor Processing. We first introduced the use of clustering techniques to learn metaphorical associations in our earlier work (Shutova, Sun, and Korhonen 2010; Shutova and Sun 2013). The metaphor identification system of Shutova, Sun, and Korhonen (2010) 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 spectral clustering of verbs and nouns. Shutova, Sun, and Korhonen (2010) introduced the hypothesis of "clustering by association," stating 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. In the course of distributional clustering, concrete concepts (e.g., *water, coffee, beer, liquid*) tend to be clustered together when they have similar meanings. In contrast, abstract concepts (e.g., *marriage, democracy, cooperation*) tend to be clustered together when they are metaphorically associated with the same source domain(s) (e.g., both *marriage* and *democracy* can be viewed as *mechanisms* or *games*). Because of this shared association structure they share common contexts in the corpus. For instance, Figure 2 shows a more concrete cluster of *mechanisms* and a more abstract cluster containing both *marriage* and *democracy*, along with their associated verb cluster. Such clustering patterns allow the system to discover new, previously unseen conceptual and linguistic metaphors starting from a small set of examples, or seed metaphors. For instance, having seen the seed metaphor "*mend* marriage" it infers that "the *functioning* of democracy" is also used metaphorically, since *mend* and *function* are both MECHANISM verbs and *marriage* and *democracy* are in the same cluster. This is how the system expands from a small set of seed metaphorical expressions to cover new concepts and new metaphors.

Shutova, Sun, and Korhonen (2010) experimented with unconstrained spectral clustering and applied their system to English data. In this article, we extend their method to perform constrained clustering, and thus investigate the effectiveness of additional supervision in the form of annotated metaphorical mappings. We then also apply the

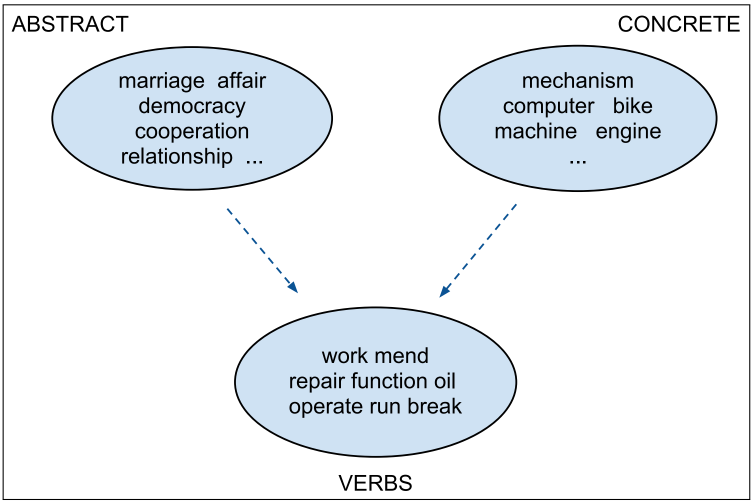


Figure 2
Clusters of abstract and concrete nouns. On the right is a cluster containing concrete concepts that are various kinds of *mechanisms*; at the bottom is a cluster containing verbs co-occurring with *mechanisms* in the corpus; and on the left is a cluster containing abstract concepts that tend to co-occur with these verbs.

original unconstrained method and its new constrained variant to three languages—English, Spanish, and Russian—thus testing the approach in a multilingual setting.

The second set of experiments in this article are based on the method of Shutova and Sun (2013), which is inspired by the same observation about distributional clustering. Through the use of hierarchical soft clustering techniques, Shutova and Sun (2013) derive a network of concepts in which metaphorical associations are exhibited at different levels of granularity. If, in the method of Shutova, Sun, and Korhonen (2010), the source and target domain clusters were connected through the use of the seed expressions, the method of Shutova and Sun (2013) learns both the clusters and the connections between them automatically from the data, in a fully unsupervised fashion. Because one of the aims of this article is to investigate the level and type of supervision optimally required to generalize metaphorical mechanisms from text, we adapt and apply the method of Shutova and Sun (2013) to our languages of interest and compare its performance to that of the spectral clustering based methods across languages. We thus also test the method, which has been previously evaluated only on English data, in a multilingual setting.

Clustering techniques have also been previously used in metaphor processing research in a more traditional sense (i.e., to identify linguistic expressions with a similar or related meaning). Mason (2004) performed WordNet sense clustering to obtain selectional preference classes, and 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. Their core assumption was that all instances of the verb in semantically similar sentences have the same sense, either the literal or the metaphorical one. However, the latter approaches did not investigate how metaphorical associations structure the distributional semantic space, which is what we focus on in this article.

2.2.3 LDA Topic Modeling. Heintz et al. (2013) applied LDA topic modeling (Blei, Ng, and Jordan 2003) to the problem of metaphor identification in experiments with English and Spanish. Their hypothesis was that if a sentence contained both source and target domain vocabulary, it contained a metaphor. The authors focused on the target domain of *governance* and manually compiled a set of source concepts with which *governance* could be associated. They used LDA topics as proxies for source and target concepts: If vocabulary from both source and target topics was present in a sentence, this sentence was tagged as containing a metaphor. The topics were learned from Wikipedia and then aligned to source and target concepts using sets of human-created seed words. When the metaphorical sentences were retrieved, the source topics that are common in the document were excluded. This ensured that the source vocabulary was transferred from a new domain. 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 these data, and output a ranked set of metaphorical examples. They carried out two types of evaluation: (1) top five linguistic examples for each conceptual metaphor were judged by two annotators, yielding an F-score of 0.59 for English ($\kappa = 0.48$); and (2) 250 top-ranked examples in system output were 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.

The method of Heintz et al. (2013) relies on the ideas of the Conceptual Metaphor Theory, in that metaphorical language can be generalized using information about source and target domains. Many supervised classification approaches (e.g., Mohler et al. 2013; Tsvetkov, Mukomel, and Gershman 2013), as well as our own approach, share this intuition. However, our methods are different in their aims. If the method of Heintz et al. (2013) learned information about the internal domain structure from the data (through the use of LDA), our methods aim to learn information about cross-domain mappings, as well as the internal domain structure, from the words' distributional behavior.

In addition, in contrast to most of the systems described in this section, we experiment with minimally supervised and unsupervised techniques that require little or no annotated training data, and thus can be easily adapted to new domains and languages. Unlike most previous approaches, we also experiment with metaphor identification in a general-domain setting.

3. Data Sets and Feature Extraction

Because our approach involves distributional learning from large collections of text, the choice of an appropriate text corpus plays an important role in the experiments and the interpretation of results. We have selected comparably large, wide-coverage corpora in our three languages to train the systems. The corpora were then parsed using a dependency parser and VERB–SUBJECT, VERB–DIRECT_OBJECT, and VERB–INDIRECT_OBJECT relations were extracted from the parser output. Following previous semantic noun and verb clustering experiments (Pantel and Lin 2002; Bergsma, Lin, and Goebel 2008; Sun and Korhonen 2009), we use these grammatical relations (GRs) as features for clustering. The features used for noun clustering consisted of the verb lemmas occurring in VERB–SUBJECT, VERB–DIRECT_OBJECT, and VERB–INDIRECT_OBJECT relations with the nouns in our data set, indexed by relation type. The features used for verb clustering were the noun lemmas, occurring in the above GRs with the verbs in the data set, also indexed by relation type. The feature values were the relative frequencies of the features. For instance, the feature vector for *democracy* in English would contain the following entries: {restore-dobj n_1, establish-dobj n_2, build-dobj n_3, ... , vote_in-iobj n_i, call_for-iobj n_{i+1}, ... , survive-subj n_k, emerge-subj n_{k+1}, ...}, where n is the frequency of the feature.

3.1 English Data

The English verb and noun data sets used for clustering contain the 2,000 most frequent verbs and the 2,000 most frequent nouns in the BNC (Burnard 2007), respectively. The BNC is balanced with respect to topic and genre, which makes it appropriate for the selection of a data set of most common source and target concepts and their linguistic realizations. The features for clustering were, however, extracted from the English Gigaword corpus (Graff et al. 2003), which is more suitable for feature extraction because of its large size. The Gigaword corpus was first parsed using the RASP parser (Briscoe, Carroll, and Watson 2006) and the VERB–SUBJECT, VERB–DIRECT_OBJECT, and VERB–INDIRECT_OBJECT relations were then extracted from the GR output of the parser, from which the feature vectors were formed.

3.2 Spanish Data

The Spanish data were extracted from the Spanish Gigaword corpus (Mendonca et al. 2011). The verb and noun data sets used for clustering consisted of the 2,000 most frequent verbs and 2,000 most frequent nouns in this corpus. The corpus was parsed using the Spanish Malt parser (Nivre et al. 2007; Ballesteros et al. 2010). VERB–SUBJECT, VERB–DIRECT_OBJECT, and VERB–INDIRECT_OBJECT relations were then extracted from the output of the parser and the feature vectors were constructed for all verbs and nouns in the data set in a similar manner to the English system. For example, the feature vector for the noun *democracia* included the following entries: {destruir-dobj n_1, reinstaurar-dobj n_2, proteger-dobj n_3, ... , elegir_a-iobj n_i, comprometer_con-iobj n_{i+1}, ... , florecer-subj n_k, funcionar-subj n_{k+1}, ...}.

3.3 Russian Data

The Russian data were extracted from the RU-WaC corpus (Sharoff 2006), a two-billion-word representative collection of text from the Russian Web. The corpus was parsed using the Malt dependency parser for Russian (Sharoff and Nivre 2011), and the VERB–SUBJECT, VERB–DIRECT_OBJECT, and VERB–INDIRECT_OBJECT relations were extracted to create the feature vectors. Similarly to the English and Spanish experiments, the 2,000 most frequent verbs and 2,000 most frequent nouns, according to the RU-WaC, constituted the verb and noun data sets used for clustering.

4. Semi-Supervised Metaphor Identification Experiments

We first experiment with a flat clustering solution, where metaphorical patterns are learned by means of hard clustering of verbs and nouns at one level of generality.[6] This approach to metaphor identification is based on the hypothesis of clustering by association, which we first introduced in Shutova, Sun, and Korhonen (2010). Our expectation is that clustering by association would allow us to learn numerous new target domains that are associated with the same source domain from the data in a minimally supervised way. Following Shutova, Sun, and Korhonen (2010), we also use clustering techniques to collect source domain vocabulary.

We perform verb and noun clustering using the spectral clustering algorithm, which has proven to be effective in lexical acquisition tasks (Brew and Schulte im Walde 2002; Sun and Korhonen 2009) and is suitable for high-dimensional data (Chen et al. 2006). We experiment with its *unconstrained* and *constrained* versions. The unconstrained algorithm performs clustering (and thus identifies metaphorical patterns) in a fully unsupervised way, relying on the information contained in the data alone. The constrained version uses a small set of example metaphorical mappings as constraints to reinforce clustering by association. We then investigate to what extent adding metaphorical constraints affects the resulting partition of the semantic space as a whole. Further details of these two methods are provided subsequently. Once the clusters have been created in either the unconstrained or constrained setting, the identification of metaphorical expressions is boosted from a small number of linguistic examples—the seed expressions.

The seed expressions in our experiments are verb–subject and verb–direct object metaphors, in which the verb metaphorically describes the noun (e.g., "*mend* marriage"). Note that these are linguistic metaphors; their corresponding metaphorical mappings are not annotated. The seed expressions are then used to establish a link between the verb cluster that contains source domain vocabulary and the noun cluster

[6] Hard clustering produces a partition where every object belongs to one cluster only.

that contains diverse target concepts associated with that source domain. This link then allows the system to identify a large number of new metaphorical expressions in a text corpus. In summary, the system (1) performs noun clustering in order to harvest target concepts associated with the same source domain; (2) creates a source domain verb lexicon by means of verb clustering; (3) uses seed expressions to connect source (verb) and target (noun) clusters between which metaphorical associations hold; and (4) searches the corpus for metaphorical expressions describing the target domain concepts using the verbs from the source domain lexicon.

4.1 Clustering Methods

4.1.1 Spectral Clustering. Spectral clustering partitions objects relying on their similarity matrix. Given a set of data points, the similarity matrix $W \in \mathcal{R}^{N \times N}$ records similarities w_{ij} between all pairs of points. We construct similarity matrices using the **Jensen-Shannon divergence** as a measure. Jensen-Shannon divergence between two feature vectors q_i and q_j is defined as follows:

$$JSD(q_i, q_j) = \frac{1}{2}D(q_i||m) + \frac{1}{2}D(q_j||m) \tag{1}$$

where D is the Kullback-Leibler divergence, and m is the average of the q_i and q_j. We then use the following similarity w_{ij} between i and j as defined in Sun and Korhonen (2009):

$$w_{ij} = e^{-JSD(q_i, q_j)} \tag{2}$$

The similarity matrix W encodes a weighted undirected graph $G := (V, E)$, by providing its adjacency weights. We can think of the points we are going to cluster as the vertices of the graph, and their similarities w_{ij} as connection weights on the edges of the graph. Spectral clustering attempts to find a partitioning of the graph into clusters that are minimally connected to vertices in other clusters, but which are of roughly equal sizes (Shi and Malik 2000). This is important for metaphor identification, as our aim is to identify clusters of target concepts associated with the same source domain on one hand and to ensure that different metaphorical mappings are separated from each other in the overall partition on the other hand. In particular, we use the NJW spectral clustering algorithm introduced by Ng et al. (2002).[7]

In our case, each vertex v_i represents a word indexed by $i \in 1, ..., N$. The weight between vertices v_i and v_j is denoted by $w_{ij} \geq 0$ and represents the similarity or adjacency between v_i and v_j, taken from the adjacency matrix W. If $w_{ij} = 0$, we say vertices v_i and v_j are unconnected. Because G is taken to be undirected, W must be symmetric—this explains our use of Jensen-Shannon divergence rather than the more well-known Kullback-Leibler divergence in constructing our similarity matrix W.[8] We denote the **degree** of a vertex v_i by $d_i := \sum_{j=1}^{N} w_{ij}$. The degree represents the weighted connectivity of v_i to the rest of the graph. Finally, we define the **graph Laplacian** of G as $L := D - W$; the role of the graph Laplacian will become apparent subsequently.

[7] For a comprehensive review of spectral clustering algorithms see Von Luxburg (2007). Our description of spectral clustering here is largely based on this review.

[8] Note that any symmetric matrix with non-negative, real-valued elements can therefore be taken to represent a weighted, undirected graph.

Recall that our goal is to minimize similarities (weights) between clusters while producing clusters of roughly equal sizes. Denote the sum of weights between cluster A and points not in cluster A as $W(A, -A) := \sum_{i \in A, j \notin A} w_{ij}$. The NCUT objective function introduced by Shi and Malik (2000) incorporates a tradeoff between these two objectives as:

$$\mathrm{NCut}(A_1, ..., A_K) := \sum_{k=1}^{K} \frac{W(A_k, -A_k)}{\sum_{v_\ell \in A_k} d_\ell} \tag{3}$$

We can now recast our goal as finding the partitioning $A_1, ... A_K$ that minimizes this objective function. We can achieve some clarity about this objective function by rewriting it using linear algebra. If we define the normalized indicator vectors $h_k := (h_{1k}, ..., h_{Nk})^T$ where we set

$$h_{i,k} := \begin{cases} \frac{1}{\sqrt{\sum_{v_\ell \in A_k} d_\ell}} & \text{if } v_i \in A_k \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

then some straightforward computations reveal that:

$$h_k^T L h_k = \frac{1}{2} \sum_{i \in A_k, j \notin A_k} w_{ij} = \frac{W(A_k, -A_k)}{\sum_{v_\ell \in A_k} d_\ell} \tag{5}$$

Therefore, if we collect the vectors $h_1, ..., h_K$ into a matrix $H = (h_1, ..., h_K)$, then $h_k^T L h_k = (H^T L H)_{kk}$, and minimizing Equation (3) is equivalent to the following minimization problem on the graph Laplacian:

$$\min_H \mathrm{Tr}(H^T L H) \text{ where } H \text{ is subject to constraint 4} \tag{6}$$

If we could find the optimal H, it would be straightforward to find the cluster memberships from H, since h_{ik} is nonzero if and only if v_i is in cluster A_k. Unfortunately, solving this minimization problem is NP hard (Wagner and Wagner 1993; Von Luxburg 2007). However, an approximate solution can be found by relaxing the constraints on the elements of H in constraint 4. Thus, we must relax our optimization problem somewhat. One entailment of constraint 4 is that the matrix $D^{1/2}H$ is a matrix of orthonormal vectors—that is, $(D^{1/2}H)^T(D^{1/2}H) = H^T D H = I$. Ng et al. (2002) proceed by dropping the constraint that h_{ik} be either 0 or $1/\sqrt{\sum_{v_\ell \in A_k} d_\ell}$, but keeping the orthonormality constraint. Thus, they seek to solve the following problem:

$$\min_{H \in \mathcal{R}^{N \times K}} \mathrm{Tr}(H^T L H) \text{ subject to } H^T D H = I \tag{7}$$

By setting $T := D^{1/2}H$, this can be rewritten as

$$\min_{T \in \mathcal{R}^{N \times K}} \mathrm{Tr}(T^T D^{-1/2} L D^{-1/2} T) \text{ subject to } T^T T = I \tag{8}$$

This problem is tractable because it is equivalent to the problem of finding the first K

eigenvectors of $D^{-1/2}LD^{-1/2}$.
Because we have dropped the constraint that $h_{i,k}$ be nonzero if and only if v_i is in cluster A_k from Equation (4), then we can no longer infer the cluster memberships directly from H or T. Instead, Ng et al. (2002) approximately infer cluster memberships by clustering in the eigenspace defined by T using a clustering algorithm such as K-MEANS. The algorithm of Ng et al. (2002) is summarized as Algorithm 1.

4.1.2 Spectral Clustering with Constraints. Constrained clustering methods incorporate prior knowledge about which words belong in the same clusters. In our experiments, we sought methods that were well-behaved when given only positive constraints (i.e., *two words belong in the same cluster*) rather than both positive and negative constraints (i.e., *two words do not belong in the same cluster*). Because we have no hard-and-fast constraints that must be satisfied, but rather subjective information that we believe should influence the constraints, it was also important that our methods not strictly enforce constraints, but rather be capable of weighing the constraints against information available in the similarity matrix over the set of words.

In the constrained spectral clustering algorithm introduced by Ji, Xu, and Zhu (2006), constraints are introduced by a simple modification of the objective function of NCUT. Suppose we have C pairs of constraints indicating that two words belong to the same cluster, and we have N words overall. For each pair c of words i and j that belong to the same cluster, we create an N-dimensional vector $u_c = [u_{c1}, u_{c2}, ..., u_{cN}]^T$ where $u_{ci} = 1$, $u_{cj} = -1$, and the rest of the elements are equal to zero. We then collect these vectors into the $C \times N$ constraint matrix $U^T = [u_1, u_2, ..., u_N]$.

Suppose that we form the matrix H using the constraints on h_{ik} in Equation (4), as before. Then if all of the constraints encoded in U are correctly specified, we have that $UH = 0$ and therefore the spectral norm $\|UH\|^2 = \mathrm{Tr}((UH)^T UH) = 0$. As more and more of the constraints encoded in U are violated by H, $\|UH\|$ will grow. This motivates Ji,

Algorithm 1 NJW algorithm

```
Require: Number $K$ of clusters; similarity matrix $W \in \mathcal{R}^{N \times N}$
 Compute the degree matrix $D$ where $d_{ii} = \sum_{j=1}^{N} w_{ij}$ and $d_{ij} = 0$ if $i \neq j$
 Compute the graph Laplacian $L \leftarrow D - W$
 Compute normalized graph Laplacian $\bar{L} \leftarrow D^{-1/2}LD^{-1/2}$
 Compute the first $K$ eigenvectors $V_1, ..., V_K$ of $D^{-1/2}LD^{-1/2}$
 Let $T \in \mathcal{R}^{N \times K}$ be the matrix containing the normalized eigenvectors $\frac{V_1}{\|V_1\|_2}, ..., \frac{V_K}{\|V_K\|_2}$
 Let $y_i \in \mathcal{R}^K$ be the vector corresponding to the $i^{th}$ row of $T$
 Cluster the points $(y_i)_{i=1,...,N}$ into clusters $A_1, ..., A_K$ using the K-MEANS algorithm
 return $A_1, ..., A_K$
```

Xu, and Zhu (2006) to modify the objective function in Equation (6) by adding a term that penalizes a large norm for UH:

$$\min_{H} \mathrm{Tr}(H^T L H) + \beta \|UH\|^2 \text{ where } H \text{ is subject to constraint 4} \tag{9}$$

Here, β governs how strongly the constraints encoded in U should be enforced. As before, we now relax contraint 4 and set $T = D^{1/2}H$ to yield:

$$\min_{T \in \mathcal{R}^{N \times K}} \mathrm{Tr}(T^T D^{-1/2} L D^{-1/2} T + \beta \|U D^{-1/2} T\|^2) \text{ subject to } T^T T = I \tag{10}$$

Note that $\beta \|UD^{-1/2}T\|^2 = \beta \mathrm{Tr}(T^T D^{-1/2} U^T U D^{-1/2} T)$. Therefore by collecting terms we can rewrite the objective function as:

$$\min_{T \in \mathcal{R}^{N \times K}} \mathrm{Tr}(T^T D^{-1/2} (L + \beta U^T U) D^{-1/2} T) \text{ subject to } T^T T = I \tag{11}$$

Therefore, we can find the optimal T as the first K eigenvectors of $(L + \beta U^T U)$, and we can assign cluster memberships using k-Means in a manner analogous to algorithm NJW. The pseudocode for the JXZ algorithm is shown in Algorithm 2.

4.2 Clustering Experiments

4.2.1 Unconstrained Setting. We first applied the unconstrained version of spectral clustering algorithm to our data. We experimented with different clustering granularities (producing 100, 200, 300, and 400 clusters), examined the obtained clusters, and determined that the number of clusters set to 200 is the optimal setting for both nouns and verbs in our task, across the three languages. This was done by means of qualitative analysis of the clusters as representations of source and target domains—that is, by judging how complete and homogeneous the verb clusters were as lists of potential source domain vocabulary and how many new target domains associated with the same source domain were found correctly in the noun clusters. This analysis was performed on 10 randomly selected clusters taken from different granularity settings and none of the seed expressions were used for it. Examples of clusters generated with this setting are shown in Figures 3 (nouns) and 4 (verbs) for English; Figures 5 (nouns) and 6 (verbs) for Spanish; and Figures 7 (nouns) and 8 (verbs) for Russian. The noun clusters represent

Algorithm 2 JXZ algorithm

```
Require: Number $K$ of clusters; similarity matrix $W \in \mathcal{R}^{N \times N}$; constraint matrix $U \in \mathcal{R}^{N \times C}$; enforcement parameter $\beta$
  Compute the degree matrix $D$ where $d_{ii} = \sum_{j=1}^{N} w_{ij}$ and $d_{ij} = 0$ if $i \neq j$
  Compute the graph Laplacian $L \leftarrow D - W$
  Compute the first $K$ eigenvectors $V_1, ..., V_K$ of $D^{-1/2}(L + \beta U^T U)D^{-1/2}$
  Let $T \in \mathcal{R}^{N \times K}$ be the matrix containing the normalized eigenvectors $\frac{V_1}{\|V_1\|_2}, ..., \frac{V_K}{\|V_K\|_2}$
  Let $y_i \in \mathcal{R}^K$ be the vector corresponding to the $i^{th}$ row of $T$
  Cluster the points $(y_i)_{i=1,...,N}$ into clusters $C_1, ..., C_K$ using the k-Means algorithm
  return $C_1, ..., C_K$
```

Suggested source domain: MECHANISM
Target Cluster: venture partnership alliance network association trust link relationship environment
Suggested source domain: PHYSICAL OBJECT; LIVING BEING; STRUCTURE
Target Cluster: tradition concept doctrine idea principle notion definition theory logic hypothesis interpretation proposition thesis argument refusal
Suggested source domain: STORY; JOURNEY
Target Cluster: politics profession affair ideology philosophy religion competition education
Suggested source domain: LIQUID
Target Cluster: frustration concern excitement anger speculation desire hostility anxiety passion fear curiosity enthusiasm emotion feeling suspicion

Figure 3
Clusters of English nouns (unconstrained setting; the source domain labels in the figure are suggested by the authors for clarity, the system does not assign any labels).

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

Figure 4
Clusters of English verbs.

Suggested source domain: MECHANISM
Target Cluster: avance consenso progreso solución paz acercamiento entendimiento arreglo coincidencia igualdad equilibrio
Target Cluster: relación amistad lazo vínculo conexión nexo vinculación
Suggested source domain: LIVING BEING, ORGANISM, MECHANISM, STRUCTURE, BUILDING
Target Cluster: comunidad país mundo nación africa sector sociedad región europa estados continente asia centroamérica bando planeta latinoamérica
Suggested source domain: STORY, JOURNEY
Target Cluster: tendencia acontecimiento paso curso trayectoria ejemplo pendiente tradición pista evolución
Suggested source domain: CONSTRUCTION, STRUCTURE, BUILDING
Target Cluster: seguridad vida democracia confianza estabilidad salud finanzas credibilidad competitividad

Figure 5
Clusters of Spanish nouns (unconstrained setting; the source domain labels in the figure are suggested by the authors for clarity, the system does not assign any labels).

target concepts associated with the same source concept.[9] The verb clusters contain lists of source domain vocabulary.

4.2.2 Constrained Setting. We then experimented with adding constraints to guide the clustering process. We used two types of constraints: (1) *target–source constraints* (TS) directly corresponding to metaphorical mappings (e.g., *marriage* and *mechanism)*; and (2) *target–target constraints* (TT), where two target concepts were associated with the same

[9] Some suggested source concepts are given in the figures for clarity only. The system does not use or assign those labels.

Source Cluster: distribuir consumir importar ingerir comer fumar comercializar tragar consumar beber recetar
Source Cluster: atropellar chocar volcar colisionar embestir descarrilar arrollar
Source Cluster: secar fluir regar limpiar
Source Cluster: llevar sacar lanzar colocar cargar transportar arrojar tirar echar descargar
Source Cluster: caer subir descender desplomar declinar bajar retroceder progresar repuntar replegar
Source Cluster: inundar llenar abarrotar frecuentar copar colmar atestar saturar vaciar

Figure 6
Clusters of Spanish verbs.

Suggested source domain: construction, structure, building
Target Cluster: снг группировка ислам инфраструктура православие хор клан восстание колония культ социализм пирамида держава индустрия рота оркестр раса кружок заговор
Suggested source domain: mechanism, game, structure, living being, organism
Target Cluster: образ язык бог любовь вещь культура наука искусство бизнес политика природа литература теория стиль секс личность
Suggested source domain: story; journey; battle
Target Cluster: поход сотрудничество танец спор атака беседа карьера переговоры охота битва диалог наступление прогулка
Suggested source domain: liquid
Target Cluster: вопрос проблема тема мысль идея мнение задача чувство интерес желание ощущение необходимость
Target Cluster: боль впечатление радость надежда настроение страх сожаление мечта потребность сомнение эмоция ужас уважение запах
Target Cluster: результат информация ссылка материал данные документ опыт исследование список знание оценка анализ практика

Figure 7
Clusters of Russian nouns (unconstrained setting; the source domain labels in the figure are suggested by the authors for clarity, the system does not assign any labels).

Source Cluster: спуститься спускаться скрываться направляться прятаться направиться бросаться вырваться выбраться устроиться приблизиться двинуться скрыться рваться поселиться оторваться возвратиться
Source Cluster: хлопать вскрыть распахнуть толкнуть стукнуть раскрыться приоткрыть взломать
Source Cluster: разогреть пролить сушить взбить разбавить заправить нагреть остыть протереть выдавить процедить угощать натереть угостить обжарить растворить вонять сливать
Source Cluster: сбросить доставать спрятать повесить выбросить вырезать кинуть подбирать тащить надевать уложить прятать извлечь вынуть выкинуть выбить вставлять
Source Cluster: порвать шить скинуть завязать стирать одевать натянуть сшить

Figure 8
Clusters of Russian verbs

source domain (e.g., *marriage* and *democracy*). The constraints were generated according to the following procedure:

1. **TS constraints:**
   - Select 30 target concepts.
   - For each of the target concepts select a source concept that it is associated with. This results in 30 pairs of TS constraints.
2. **TT constraints:**
   - For each of the resulting 30 TS pairs of concepts, select another target concept associated with the given source.
   - Pair the two target concepts into a TT constraint.

Table 1
Examples of constraints used in English clustering.

TT constraints	**TS constraints**
poverty & inequality	poverty & disease
democracy & friendship	democracy & machine
society & mind	society & organism
education & life	education & journey
politics & marriage	politics & game
country & family	country & building
government & kingdom	government & household
career & change	career & hill
innovation & evolution	innovation & flower
unemployment & panic	unemployment & prison
faith & peace	faith & warmth
violence & passion	violence & fire
mood & love	mood & climate
debt & tension	debt & weight

3. Constraints should satisfy the following criteria:
   - Constraints represent metaphorical mappings that hold in all three languages, as validated by native speakers.
   - Each concept should appear in the set of constraints only once.[10]

We created 30 TS and 30 TT constraint pairs following this procedure. The source and target concepts in the constraints were selected from the lists of 2,000 nouns that we clustered in the three languages. Constraints were selected and validated by the authors (who are native speakers of the respective languages) without taking the output of the unconstrained clustering step into account (i.e., prior to having seen it). The lists of constraints were first created through individual introspection, and then finalized through discussion. Tables 1, 2, and 3 show some examples of TS and TT constraints for the three languages. One pair of constraints (*relationship* & *trade* [TT] and *relationship* & *vehicle* [TS]) was excluded from the set, because *relationship* is usually translated into Spanish and Russian by a plural form (e.g., *relaciones*). We thus used 29 TT constraints and 29 TS constraints in our experiments.

Our expectation is that the TT constraints are better suited to aid metaphor discovery, as the noun clusters tend to naturally contain distinct target domains associated with the same source. The TT constraints are designed to reinforce this principle. However, introducing the TS type of constraint allows us to investigate to what extent explicitly reinforcing the source domain features in clustering allows us to harvest more target domains associated with the source.

We experimented with different constraint enforcement parameter settings ($\beta = 0.25, 1.0, 4.0$) in order to investigate the effect of the constraints on the overall partition of

[10] This applies to both the source and the target concepts. This requirement was imposed to ensure that the constraints are enforced pairwise during clustering.

Table 2
Examples of constraints used in Spanish clustering.

TT constraints	TS constraints
pobreza & desigualdad	pobreza & enfermedad
democracia & amistad	democracia & máquina
sociedad & mente	sociedad & organismo
educación & vida	educación & viaje
política & matrimonio	política & juego
país & familia	país & edificio
gobierno & reino	gobierno & casa
carrera & cambio	carrera & colina
innovación & evolución	innovación & flor
desempleo & pánico	desempleo & prisión
fe & paz	fe & calor
violencia & pasión	violencia & fuego
ánimo & amor	ánimo & clima
deuda & tensión	deuda & peso

Table 3
Examples of constraints used in Russian clustering.

TT constraints	TS constraints
бедность & неравенство	бедность & болезнь
демократия & дружба	демократия & механизм
общество & разум	общество & организм
образование & жизнь	образование & путешествие
политика & брак	политика & игра
страна & семья	страна & постройка
правительство & королевство	правительство & хозяйство
карьера & перемена	карьера & холм
инновация & эволюция	инновация & цветок
безработица & паника	безработица & тюрьма
вера & мир	вера & тепло
насилие & страсть	насилие & огонь
настроение & любовь	настроение & климат
долг & напряжение	долг & вес

the semantic space. Our data analysis has shown that interesting effects of metaphorical constraints were most strongly manifested with $\beta = 4.0$, and we thus used this setting in our further experiments. Examples of clusters generated with the use of constraints in the three languages are shown in Figures 9, 10, and 11. Our analysis of the clusters has confirmed that the use of TT constraints resulted in clusters containing more diverse target concepts associated with the same source. Compare, for instance, the unconstrained and TT constrained clusters in Figure 9. The unconstrained cluster predominantly contains concepts related to *politics*, such as *profession* and *ideology*, albeit also capturing other target domains, such as *religion* and *education*. Adding the constraint MARRIAGE & POLITICS, however, further increases the domain diversity of the cluster, adding such

Unconstrained:
Cluster: **politics** profession affair ideology philosophy religion competition education

TT constraints:
Cluster: fibre **marriage politics** affair career life hope dream religion education economy

TS constraints:
Cluster: field england part card **politics** sport music tape tune guitar trick football organ instrument round match **game** role ball host

Figure 9
Clusters of English nouns: unconstrained and constrained settings.

Unconstrained:
Cluster: dolor impacto miedo repercusión consecuencia escasez efecto **dificultad**

TT constraints:
Cluster: miedo cuidado repercusión epicentro acceso pendiente oportunidad conocimiento **dificultad**

TS constraints:
Cluster: veto bloqueo inmunidad restricción obstáculo **barrera dificultad**

Figure 10
Clusters of Spanish nouns: unconstrained and constrained settings.

Unconstrained:
Cluster: знание способность **красота** усилие умение талант навык точность дар познание мудрость квалификация мастерство

TT constraints:
Cluster: власть **счастье красота** слава честь популярность благо богатство дар авторитет весть

TS constraints:
Cluster: свет звезда солнце **красота** улыбка луна **луч**

Figure 11
Clusters of Russian nouns: unconstrained and constrained settings.

target concepts as *life, hope, dream,* and *economy*. The Spanish TT constrained clustering in Figure 10 shows the wider effects of constrained clustering throughout the whole noun space. Although none of the constraints is explicitly manifested in this cluster, one can see that this cluster nonetheless contains a more diverse set of target concepts associated with the same source, as compared to the original unconstrained cluster (see Figure 10). The TS constraints, as expected, highlighted the source domain features of the target word, resulting in (for example) assigning *politics* to the same cluster as *game* terms, such as *round* and *match* in English (given the TS constraint POLITICS & GAME). These types of constraints are thus less likely to be suitable for metaphor identification, where purely target clusters are desired. These trends were evident across the three languages, as demonstrated by the examples in the respective figures.

4.3 Identification of Metaphorical Expressions

4.3.1 Seed Expressions. Once the clusters have been obtained, we then used a set of seed metaphorical expressions to connect the source and target clusters, thus enabling the system to recognize new metaphorical expressions. The seed expressions in the three languages were extracted from naturally occurring text, manually annotated for linguistic metaphor.

English seed expressions The seed examples used in the English experiments were extracted from the metaphor corpus created by Shutova and Teufel (2010). Their corpus is a subset of the BNC covering a range of genres: fiction; news articles; essays on politics, international relations, and history; and radio broadcast (transcribed speech). As such, the corpus provides a suitable platform for testing the metaphor-processing system on real-world general-domain expressions in contemporary English. We extracted verb–subject and verb–direct object metaphorical expressions from this corpus. All phrases were included unless they fell into one of the following categories:

- Phrases where the subject or object referent is unknown (e.g., containing pronouns such as "in which they [changes] *operated*") or represented by a named entity (e.g., "Then Hillary *leapt* into the conversation").
- Phrases whose metaphorical meaning is realized solely in passive constructions (e.g., "sociologists have been *inclined* to [..]").
- Multi-word metaphors (e.g., "*go on pilgrimage* with Raleigh or *put out to sea* with Tennyson"), because these are beyond the scope of our experiments.

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

Russian and Spanish seed expressions We have collected a set of texts in Russian and Spanish, following the genre distribution of the English corpus of Shutova and Teufel (2010), insofar as possible. Native speakers of Russian and Spanish then annotated linguistic metaphors in these corpora, following the annotation procedures and guidelines of Shutova and Teufel. We then extracted the metaphorical expressions in verb–subject and verb–direct object constructions from these data, according to the same criteria used to create the English seed set. This resulted in 72 seed expressions for Spanish and 85 seed expressions for Russian. The Spanish seed set includes, for instance, the following examples: "*vender* influencia," "*inundar* mercado," "*empapelar* ciudad," "*labrarse* futuro," *contagiar* estado" (verb–direct object constructions); and "violencia *salpicó*," "debate *tropezó*," "alegría *brota*," "historia *gira*," "corazón *saltó*" (subject–verb constructions). The expressions in the seed sets were manually annotated for the corresponding grammatical relations.

4.3.2 Corpus Search. Each individual seed expression implies a connection between a source domain (through the source domain verb; e.g., *mend*) and a target domain (through the target domain noun; e.g., *marriage*). The seed expressions are thus used to connect source and target clusters between which metaphorical associations hold. The system then proceeds to search the respective corpus for source and target domain terms from the connected clusters within a single grammatical relation. Specifically, the system classifies verb–direct object and verb–subject relations in the corpus as metaphorical if the lexical items in the grammatical relation appear in the linked source (verb) and target (noun) clusters. Consider the following example sentence extracted from the BNC for English.

(1) Few would deny that in the nineteenth century change was greatly *accelerated*.

The relevant GRs identified by the parser are presented in Figure 12. The relation between the verb *accelerate* and its semantic object *change* is expressed in the passive voice and is, therefore, tagged by RASP as an `ncsubj` GR. Because this GR contains terminology from associated source (MOTION) and target (CHANGE) domains, it is marked as metaphorical and so is the term *accelerate*, which belongs to the source domain. The search space for metaphor identification was the BNC parsed by RASP for English; the Spanish Gigaword corpus parsed by the Spanish Malt parser for Spanish; and the RuWaC parsed by the Russian Malt parser for Russian. The search was performed similarly in the three languages: The system searched the corpus for the source and target domain vocabulary within a particular grammatical relation (verb–direct object or verb–subject). Some examples of retrieved metaphorical expressions are presented in Figures 13, 14, and 15.

4.4 Evaluation

We applied the UNCONSTRAINED and CONSTRAINED versions of our system to identify metaphor in continuous text in the three languages. Examples of full sentences containing metaphorical expressions as annotated by the UNCONSTRAINED systems are shown in Figures 16, 17, and 18. We evaluated the performance of UNCONSTRAINED and CONSTRAINED methods in the three languages on a random sample of the extracted metaphors against human judgments.

```
(1) change was greatly accelerated
ncsubj head=accelerate+ed_VVN_25 dep=change_NN1_22
aux head=accelerate+ed_VVN_25 dep=be+ed_VBDZ_23
ncmod head=accelerate+ed_VVN_25 dep=greatly_RR_24
passive head=accelerate+ed_VVN_25
```

Figure 12
RASP grammatical relations output for metaphorical expressions.

***cast* doubt** (V–O)
cast fear, *cast* suspicion, *catch* feeling, *catch* suspicion, *catch* enthusiasm, *catch* emotion, *spark* fear, *spark* enthusiasm, *spark* passion, *spark* feeling, *fix* emotion, *shade* emotion, *blink* impulse, *flick* anxiety, *roll* doubt, *dart* hostility ...

campaign *surged* (S–V)
charity *boomed*, effort *dropped*, campaign *shrank*, campaign *soared*, drive *spiraled*, mission *tumbled*, initiative *spiraled*, venture *plunged*, effort *rose*, initiative *soared*, effort *fluctuated*, venture *declined*, effort *dwindled* ...

Figure 13
English metaphorical expressions identified by the system for the seeds "*cast* doubt" and "campaign *surged*."

debate *tropezó* (debate *stumbled*) (S–V)
proceso *empantanó* (get swamped), juicio *empantanó*, proceso *estancó*, debate *estancó*, juicio *prosperó*, contacto *prosperó*, audiencia *prosperó*, proceso *se topó*, juicio *se topó*, proceso *se trabó*, debate *se trabó*, proceso *tropezó*, juicio *tropezó*, contacto *tropezó* ...

***inundar* mercado (to *flood* the market)** (V–O)
abarrotar mercado, *abarrotar* comercio, *atestar* mercado, *colmar* mercado, *colmar* comercio, *copar* mercado, *inundar* comercio, *inundar* negocio, *llenar* mercado, *llenar* comercio, *saturar* mercado, *saturar* venta, *saturar* negocio, *vaciar* negocio, *vaciar* intercambio ...

Figure 14
Spanish metaphorical expressions identified by the system for the seeds "debate *tropezó*" and "*inundar* mercado."

***обойти* закон (*bypass* the law)** (V-O)
перевернуть закон, *обойти* постановление, *перевернуть* пункт, *засечь* норму, *запустить* кодекс, *перевернуть* кодекс, *вносить* запрет, *открыть* законодательство, *вносить* порядок, *приклеить* закон, *растянуть* правило, *засечь* порядок, *растянуть* ограничение, *перевернуть* запрет, *запустить* запрет, *выдавить* пункт, *выдавить* постановление, *сваливать* правило

принцип *отражается* (the principle is *reflected*) (S-V)
диета *основывается*, решение *основывается*, мера *отражается*, правило *отражается*, требование *отражается*, порядок *проявляется*, правило *проявляется*, принцип *проявляется*, условие *проявляется*, договор *сводится*, закон *сводится*, план *сводится*, принцип *сводится*, требование *сводится*, диета *сказывается*, закон *сказывается*, решение *сказывается*

Figure 15
Russian metaphorical expressions identified by the system.

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

Figure 16
Retrieved English sentences.

4.4.1 Baseline. In order to show that our metaphor identification methods generalize well over the seed set and capture diverse target domains (rather than merely synonymous ones), we compared their output with that of a baseline system built upon WordNet. In the baseline system, WordNet synsets represent source and target domains in place of automatically generated clusters. The system thus expands over the seed set by using synonyms of the metaphorical verb and the target domain noun. It then searches the corpus for phrases composed of lexical items belonging to those synsets. For example, given a seed expression "*stir* excitement," the baseline finds phrases such as "*arouse* fervor, *stimulate* agitation, *stir* turmoil," and so forth. The comparison against the WordNet

1. Se espera que el principal mediador se reúna el martes con todos los involucrados en el proceso de paz liberiano, pero es seguro que **los disturbios** ***ensombrecerán*** el proceso.
2. Sigue siendo la falla histórica, religiosa y étnica que puede ***romper*** nuevamente **la estabilidad** regional [..]
3. Desea trasladar las maquiladoras de la zona fronteriza a zonas del interior, con el fin de ***repartir*** **las oportunidades de empleo** más equitativamente.
4. Los precios del café cayeron a principios de la actual década, al ***abarrotarse*** **el mercado** como consecuencia del derrumbe de un sistema de cuotas de exportación.

Figure 17
Retrieved Spanish sentences.

1. Весь 2011 год Кудрин ***зажимал*** **деньги** в бюджете, не пуская их в экономику.
For all of 2011, Kudrin ***plugged up*** **money** in the budget, not letting it into the economy.
2. Именно поэтому не остается от фильма осадка нравоучений, становится только невыносимо тяжело от того, что человек по неосторожности и безответственности может лишиться жизни за секунду, ***разбить*** **судьбы** других ни в чем не повинных людей.
Thus there remains of the film no sediment of moral admonition, it becomes only unbearably hard in that a person through carelessness and irresponsibility can be deprived of life in a second, and ***shatter*** **the fate** of other in no way guilty people.
3. `` **Кризис** ***гуляет*** по стране, люди скорее будут думать о хлебе насущном, чем о зрелищах", - отмечает Вертилецкий.
`` **Crisis** ***strolls*** through the country, people are quicker to think about their daily bread, than about shows," notes Vertiletsky.
4. На турецко-сирийской границе, где долгое время сохраняется напряженная обстановка, снова ***назревает*** острый **конфликт**.
On the Turkish-Syrian border, where the situation has long remained tense, pungent **conflict is** ***ripening*** once again.

Figure 18
Retrieved Russian sentences.

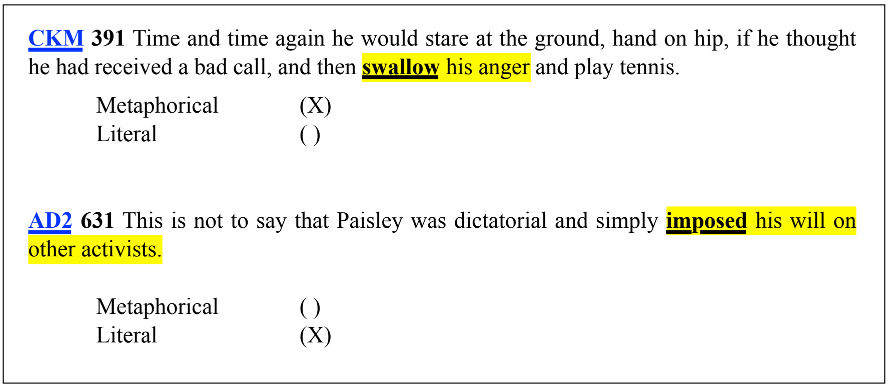
baseline was carried out for the English systems only, because the English WordNet is considerably more comprehensive than the Spanish or the Russian one.

4.4.2 Soliciting Human Judgments. The quality of metaphor identification for the systems and the baseline was evaluated in terms of precision with the aid of human judges. For this purpose, we randomly sampled sentences containing metaphorical expressions as annotated by the UNCONSTRAINED and CONSTRAINED systems and by the baseline (for English) and asked human annotators to decide whether these were metaphorical or not.

Participants Two volunteer annotators per language participated in the experiments.[11] They were all native speakers of the respective languages and held at least a Bachelor's degree.

Materials We randomly sampled 100 sentences from the output of the UNCONSTRAINED, TT CONSTRAINED, and TS CONSTRAINED systems for each language and the WordNet baseline system for English. Each sentence contained a metaphorical expression annotated by the respective system. We then also extracted 100 random

[11] We were limited in resources when recruiting annotators for Russian and Spanish, thus we had to restrict the number of participants to two per language. However, we would like to note that it is generally desirable to recruit multiple annotators for a metaphor annotation task.



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

Metaphorical (X)
Literal ( )

AD2 **631** This is not to say that Paisley was dictatorial and simply **imposed** his will on other activists.

Metaphorical ( )
Literal (X)

Figure 19
Soliciting human judgments: Annotation set-up.

sentences containing verbs in direct object and subject relations from corpora for each language. These examples were used as distractors in the experiments. The subjects were thus presented with a set of 500 sentences for English (UNCONSTRAINED, TT and TS CONSTRAINED, baseline, distractors) and 400 sentences for Russian and Spanish (UNCONSTRAINED, TT and TS CONSTRAINED, distractors). The sentences in the sets were randomized. An example of the sentence annotation format is given in Figure 19.

Task and guidelines The participants were asked to mark which of the expressions were metaphorical in their judgment. They were encouraged to rely on their own intuition of what a metaphor is in the annotation process. However, additional guidance in the form of the following definition of metaphor (Pragglejaz Group 2007) was also provided:

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

Interannotator agreement We assessed the reliability of the annotations in terms of kappa (Siegel and Castellan 1988). The interannotator agreement was measured at $\kappa = 0.62$ ($n = 2, N = 500, k = 2$) in the English experiments (substantial agreement); $\kappa = 0.58$ ($n = 2, N = 400, k = 2$) in the Spanish experiments (moderate agreement); and $\kappa = 0.64$ ($n = 2, N = 400, k = 2$) in the Russian experiments (substantial agreement). The data suggest that the main source of disagreement between the annotators was the presence of highly conventional metaphors (e.g., verbs such as *impose, convey, decline*). According to previous studies (Gibbs 1984; Pragglejaz Group 2007; Shutova and Teufel 2010) such metaphors are deeply ingrained in our everyday use of language and thus are perceived by some annotators as literal expressions.

4.4.3 Results. The system performance was then evaluated against the elicited judgments in terms of precision. The system output was compared with the judgments of each

Table 4
UNCONSTRAINED, CONSTRAINED, and baseline precision in the identification of metaphorical expressions.

System	UNCONSTRAINED	TS CONST	TT CONST	WordNet baseline
English	0.77	0.70	0.76	0.40
Spanish	0.74	0.69	0.72	-
Russian	0.67	0.62	0.73	-

Table 5
English, Russian, and Spanish system coverage (unconstrained setting).

Language	Total seeds	Total expressions identified	Total sentences
English	62	1,512	4,456
Spanish	72	1,538	22,219
Russian	85	1,815	38,703

annotator individually and the average precision across annotators for a given language is reported. The results are presented in Table 4. These results demonstrate that the method is portable across languages, with the UNCONSTRAINED system achieving a high precision of 0.77 in English, 0.74 in Spanish, and 0.67 in Russian. As we expected, TT constraints outperformed the TS constraints in all languages. This is likely to be the result of the explicit emphasis on the source domain features in TS-constrained clustering, which led to a number of literal expressions (containing the source domain noun) being tagged as metaphorical (e.g., *approach a barrier*). The effect of TT constraints is not as pronounced as we expected in English and Spanish. In Russian, however, TT constraints led to a considerable improvement of 6 percentage points in system performance, yielding the highest precision.

The CONSTRAINED and UNCONSTRAINED variants of our method harvested a comparable number of metaphorical expressions. Table 5 shows the number of seeds used in our experiments in each language, the number of unique metaphorical expressions identified by the unconstrained systems for these seeds, and the total number of sentences containing these expressions as retrieved in the respective corpus.[12] These statistics demonstrate that the systems expand considerably over the small seed sets they use as training data and identify a large number of new metaphorical expressions in corpora. It should be noted, however, that the output of the systems exhibits significant overlap in the CONSTRAINED and UNCONSTRAINED settings (e.g., 68% overlap in TS-constrained and unconstrained settings, and 73% in TT-constrained and unconstrained settings in English).

[12] Note that the English BNC is smaller in size than the Spanish Gigaword or the Russian RuWaC, leading to fewer English sentences retrieved.

4.5 Discussion and Error Analysis

We have shown that the method leads to a considerable expansion over the seed set and operates with a high precision—that is, produces high quality annotations—in the three languages. It identifies new metaphorical expressions relying on the patterns of metaphorical use that it learns automatically through clustering. We have conducted a data analysis to compare the UNCONSTRAINED and CONSTRAINED variants of our method and to gain insights about the effects of metaphorical constraints. Although at first glance the performance of the systems appeared not to be strongly influenced by the use of TT constraints (except in the case of Russian), the analysis of the identified expressions revealed interesting qualitative differences. According to our qualitative analysis, the TT constrained clusters exhibited a higher diversity with respect to the target domains they contained in all languages, leading to the system capturing a higher number of new metaphorical patterns, as compared to the unconstrained clusters. As a result, it discovered a more diverse set of metaphorical expressions given the same seeds. Such examples include "*mend* world" (given the seed "*mend* marriage"); "*frame* rule" (given the seed "*glimpse* duty"); or "*lodge* service," "*fuel* life," "*probe* world," "*found* science," or "*fuel* economy" (given the seed "*base* career"). Overall, our analysis has shown that even a small number of metaphorical constraints (such as 29 in our case) has global effects throughout the cluster space, that is, influences the structure of all clusters. The fact that the TT constrained method yielded a performance similar to the unconstrained method in English and Spanish and a considerably better performance in Russian suggests that such effects are desirable for metaphor processing. Another consideration that has arisen from the analysis of the system output is that the TT clustering setting may benefit from a larger cluster size in order to incorporate both similar and diverse target concepts.

The TS constrained clusters exhibit the same trend with respect to cluster diversity. However, the explicit pairing of source and target concepts (that occasionally leads to them being assigned to the same cluster) produces a number of false positives, decreasing the system precision. For instance, in the case of the constraint DIFFICULTY & BARRIER, these two nouns are clustered together. As a result, given the seed "*confront* problem," the system falsely tags expressions such as *approach barrier* or *face barrier* as metaphorical.

The comparison of the English system output to that of a WordNet baseline shows that the clusters in all clustering settings capture diverse concepts, rather than merely the synonymous ones, as in the case of WordNet synsets. The clusters thus provide generalizations over the source and target domains, leading to a wider coverage and acquisition of a diverse set of metaphors. The observed discrepancy in precision between the clustering methods and the baseline (i.e., as high as 37%) can be explained by the fact that a large number of metaphorical senses are included in WordNet. This means that in WordNet synsets, source domain verbs appear together with more abstract terms. For instance, the metaphorical sense of *shape* in the phrase "*shape* opinion" is part of the synset "(determine, shape, mold, influence, regulate)." This results in the low precision of the baseline system, because it tags literal expressions (e.g., *influence opinion*) as metaphorical, assuming that all verbs from the synset belong to the source domain.

System errors were of similar nature across the three languages and had the following key sources: (1) metaphor conventionality and (2) general polysemy. Because a number of metaphorical uses of verbs are highly conventional (such as those in "*hold* views, *adopt* traditions, *tackle* a problem"), such verbs tend to be clustered together with the verbs that would be literal in the same context. For instance, the verb *tackle* is found in a cluster with *solve, resolve, handle, confront, face,* and so on. This results in the system tagging *resolve a problem* as metaphorical if it has seen "*tackle* a problem" as a seed expression. However, the errors of this type do not occur nearly as frequently as in the case of the baseline.

A number of system errors were due to cases of general polysemy and homonymy of both verbs and nouns. For example, the noun *passage* can mean both "the act of passing from one state or place to the next" and "a section of text; particularly a section of medium length," as defined in WordNet. Our method performs hard clustering, that is, it does not distinguish between different word senses. Hence the noun *passage* occurred in only one cluster, containing concepts such as *thought, word, sentence, expression, reference, address, description,* and so on. This cluster models the latter meaning of *passage*. Given the seed phrase "she *blocked* the thought," the system then tags a number of false positives such as *block passage*, *impede passage*, *obstruct passage*, and *speed passage*.

Russian exhibited an interesting difference from English and Spanish in the organization of its word space. This is likely to be due to its rich derivational morphology. In other words, in Russian, more lexical items can be used to refer to the same concept than in English or Spanish, highlighting slightly different aspects of meaning. In English and Spanish, the same meaning differences tend to be expressed at the phrase level rather than at word level. For instance, the English verb *to pour* can be translated into Russian by at least five different verbs: *lit, nalit, slit, otlit, vilit,* roughly meaning *to pour, to pour into, to pour out, to pour only a small amount, to pour all of the liquid out, to pour some of the liquid out,* etc.[13] As a result, some Russian words tend to naturally form highly dense clusters essentially referring to a single concept (as in case of the verbs of *pouring*), while at the same time sharing similar distributional features with other, related but different concepts (such as *sip* or *spill*). This property suggests that it may be necessary to cluster a larger number of Russian nouns or verbs (into the same or lower number of clusters) in order to achieve the cluster coverage and diversity comparable to the English system. With respect to our experiments, this phenomenon has led to the unconstrained clusters containing more near-synonyms (such as the many variations of pouring), and the metaphorical constraints had a stronger effect in diversifying the clusters, thus allowing us to better capture new metaphorical associations.

Although the diversity of the noun clusters is central to the acquisition of metaphorical patterns, it is also worth noting that in many cases the system benefits not only from dissimilar concepts within the noun clusters, but also from dissimilar concepts in the verb clusters. Verb clusters produced automatically relying on contextual features may contain lexical items with distinct, or even opposite meanings (e.g., *throw* and *catch*, *take off* and *land*). However, they tend to belong to the same semantic domain. It is the diversity of verb meanings within the domain cluster that allows the generalization from a limited number of seed expressions to a broader spectrum of previously unseen metaphors, non-synonymous to those in the seed set.

The fact that our approach is seed-dependent is one of its possible limitations, affecting the coverage of the system. Wide coverage is essential for the practical use of the system. In order to obtain full coverage, a large and representative seed set is

[13] Similar examples can be found in other languages with a highly productive derivational morphology, such as German.

necessary. Although it is difficult to capture the whole variety of metaphorical language in a limited set of examples, it is possible to compile a seed set representative of common source-target domain mappings. The learning capabilities of the system can then be used to expand from those to the whole range of conventional metaphorical mappings and expressions. In addition, because the precision of the system was measured on the data set produced by expanding individual seed expressions, we would expect the expansion of new seed expressions to yield a comparable quality of annotations. Incorporating new seed expressions is thus likely to increase the recall of the system without a considerable loss in precision. However, creating seed sets for new languages may not always be practical. We thus further experiment with fully unsupervised metaphor identification techniques.

5. Unsupervised Metaphor Identification Experiments

The focus of our experiments so far has been mainly on metaphorical expressions, and metaphorical associations were modeled implicitly within the system. In addition, both the CONSTRAINED and the UNCONSTRAINED methods relied on a small amount of supervision in the form of seed expressions to identify new metaphorical language. In our next set of experiments, we investigate whether it is possible to learn metaphorical connections between the clusters from the data directly (without the use of metaphorical seeds for supervision) and thus to acquire a large set of explicit metaphorical associations and derive the corresponding metaphorical expressions in a fully unsupervised fashion.

This approach is theoretically grounded in cognitive science findings suggesting that abstract and concrete concepts are organized differently in the human brain (Binder et al. 2005; Crutch and Warrington 2005, 2010; Huang, Lee, and Federmeier 2010; Wiemer-Hastings and Xu 2005; Adorni and Proverbio 2012). According to Crutch and Warrington (2005), these differences emerge from their general patterns of relation with other concepts. In this section, we present a method that learns such different patterns of association of abstract and concrete concepts with other concepts automatically. Our system performs soft hierarchical clustering of nouns to create a network (or a graph) of concepts at multiple levels of generality and to determine the strength of association between the concepts in this graph. We expect that, whereas concrete concepts would tend to naturally organize into a tree-like structure (with more specific terms descending from the more general terms), abstract concepts would exhibit a more complex pattern of association. Consider the example in Figure 20. The figure schematically shows a small portion of the graph describing the concepts of *mechanism* (concrete), *political system*, and *relationship* (abstract) at two levels of generality. One can see from this graph that concrete concepts, such as *bike* or *engine*, tend to be strongly associated with one concept at the higher level in the hierarchy (*mechanism*). In contrast, abstract concepts may have multiple higher-level associates: the literal ones and the metaphorical ones. For instance, the abstract concept of *democracy* is literally associated with the more general concept of *political system*, as well as metaphorically associated with the concept of *mechanism*. Such multiple associations are due to the fact that *political systems* are metaphorically viewed as *mechanisms*; they can *function, break*, they can be *oiled*, and so forth. We often discuss concepts such as *democracy* or *dictatorship* using *mechanism* terminology, and thus a distributional learning approach would learn that they share features with *political systems* (from their literal uses), as well as with *mechanisms* (from their metaphorical uses, as shown next to the respective graph edges in the figure). Our

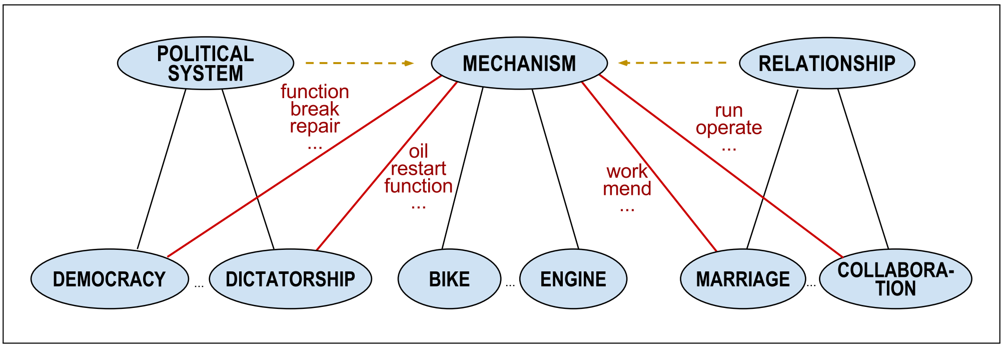


Figure 20
Organization of the hierarchical graph of concepts.

system discovers such association patterns within the graph and uses them to identify metaphorical connections between concepts.

The graph of concepts is built using hierarchical graph factorization clustering (HGFC) (Yu, Yu, and Tresp 2006) of nouns, yielding a network of clusters with different levels of generality. The weights on the edges of the graph indicate the level of association between the clusters (concepts). The system then traverses the graph to find metaphorical associations between clusters using the weights on the edges of the graph. It then generates lists of salient features for the metaphorically connected clusters and searches the corpus for metaphorical expressions describing the target domain concepts, using the verbs from the set of salient features.

5.1 Hierarchical Graph Factorization Clustering

In contrast to flat clustering, which produces a partition at one level of generality, the goal of hierarchical clustering is to organize the objects into a hierarchy of clusters with different granularities. Traditional hierarchical clustering methods widely used in NLP (such as agglomerative clustering [Schulte im Walde and Brew 2001; Stevenson and Joanis 2003; Ferrer 2004; Devereux and Costello 2005]) take decisions about cluster membership at the level of individual clusters when these are merged. As Sun and Korhonen (2011) pointed out, such algorithms suffer from two problems—error propagation and local pairwise merging—because the clustering solution is not globally optimized. In addition, they are designed to perform hard clustering of objects at each level, by successively merging the clusters. This makes them unsuitable to model multi-way associations between concepts within the hierarchy, albeit such association patterns exist in language and reasoning (Crutch and Warrington 2005; Hill, Korhonen, and Bentz 2014). As opposed to this, HGFC allows modeling of multiple relations between concepts simultaneously via a soft clustering solution. It successively derives probabilistic bipartite graphs for every level in the hierarchy. The algorithm delays the decisions about cluster membership of individual words until the overall graph structure has been computed, which allows it to globally optimize the assignment of words to clusters. In addition, HGFC can detect the number of levels and the number of clusters at each level of the hierarchical graph automatically. This is essential for our task as these settings are difficult to pre-define for a general-purpose concept graph.

The algorithm starts from a similarity matrix that encodes similarities between the objects. Given a set of nouns, $V = \{v_n\}_{n=1}^{N}$, we construct their similarity matrix W, using Jensen-Shannon Divergence as a similarity measure (as in the spectral clustering experiments). The matrix W in turn encodes an undirected similarity graph G, where the nouns are mapped to vertices and their similarities represent the weights w_{ij} on the edges between vertices i and j. Such a similarity graph is schematically shown in Figure 21(a). The clustering problem can now be formulated as partitioning of the graph G and deriving the cluster structure from it.

The graph G and the cluster structure can be represented by a bipartite graph $K(V, U)$, where V are the vertices on G and $U = \{u_p\}_{p=1}^{m}$ represent the hidden m clusters. For example, as shown in Figure 21(b), V on G can be grouped into three clusters: u_1, u_2, and u_3. The matrix B denotes the $n \times m$ adjacency matrix, with b_{ip} being the connection weight between the vertex v_i and the cluster u_p. Thus, B represents the connections between clusters at an upper and lower level of clustering. In order to derive the clustering structure, we first need to compute B from the original similarity matrix. The similarities w_{ij} in W can be interpreted as the probabilities of direct transition between v_i and v_j: $w_{ij} = p(v_i, v_j)$. The bipartite graph K also induces a similarity (W') between v_i and v_j, with all the paths from v_i to v_j going through vertices in U. This means that the similarities w'_{ij} are to be computed via the weights $b_{ip} = p(v_i, u_p)$.

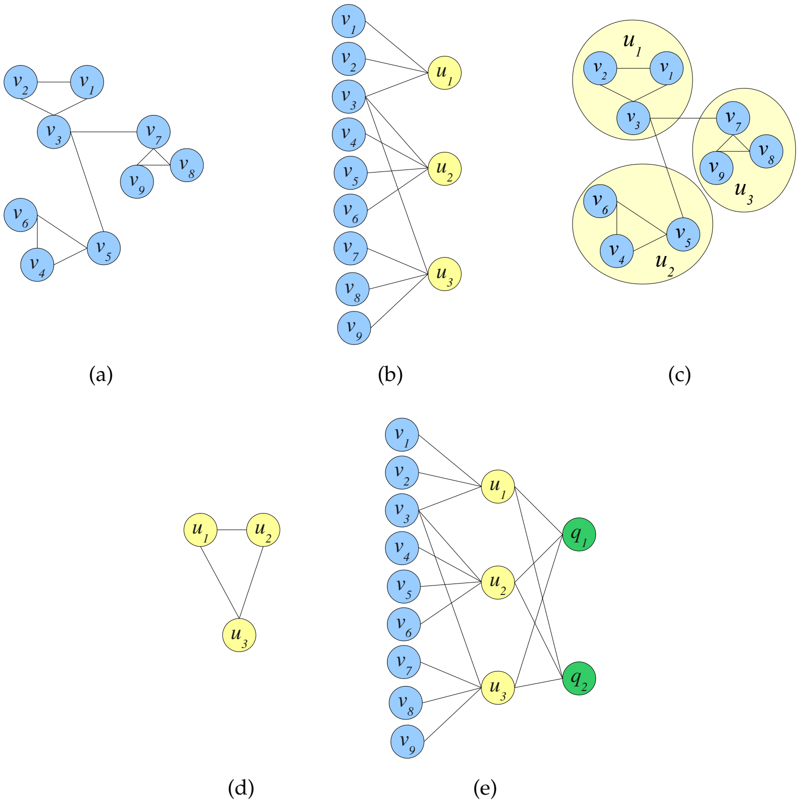


Figure 21
(a) An undirected graph G representing the similarity matrix. (b) The bipartite graph showing three clusters on G. (c) The induced clusters U. (d) The new graph G_1 over clusters U. (e) The new bipartite graph over G_1.

$$p(v_i, v_j) = p(v_i)p(v_j|v_i) = p(v_i)\sum_p p(u_p|v_i)p(v_j|u_p) = \\ p(v_i)\sum_p \frac{p(v_i, u_p)p(u_p, v_j)}{p(v_i)p(u_p)} = \sum_p \frac{p(v_i, u_p)p(u_p, v_j)}{p(u_p)} = \sum_p \frac{b_{ip}b_{jp}}{\lambda_p} \tag{12}$$

where $\lambda_i = \sum_{i=1}^{n} b_{ip}$ is the degree of vertex u_p. The new similarity matrix W' can thus be derived as follows:

$$W' : w'_{ij} = \sum_{p=1}^{m} \frac{b_{ip}b_{jp}}{\lambda_p} = (B\Lambda^{-1}B^T)_{ij} \tag{13}$$

where $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_m)$. B can then be found by minimizing the divergence distance (ζ) between the similarity matrices W and W'.

$$\min_{H,\Lambda} \zeta(W, H\Lambda H^T), s.t. \sum_{i=1}^{n} h_{ip} = 1 \tag{14}$$

We remove the coupling between B and Λ by setting $H = B\Lambda^{-1}$. Following Yu, Yu, and Tresp (2006) we define ζ as

$$\zeta(X, Y) = \sum_{ij}(x_{ij} \log \frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij}) \tag{15}$$

Yu, Yu, and Tresp (2006) showed that this cost function is non-increasing under the update rule.[14]

$$\tilde{h}_{ip} \propto h_{ip} \sum_j \frac{w_{ij}}{(H\Lambda H^T)_{ij}} \lambda_p h_{jp} \text{ s.t.} \sum_i \tilde{h}_{ip} = 1 \tag{16}$$

$$\tilde{\lambda}_p \propto \lambda_p \sum_j \frac{w_{ij}}{(H\Lambda H^T)_{ij}} h_{ip} h_{jp} \text{ s.t.} \sum_p \tilde{\lambda}_p = \sum_{ij} w_{ij} \tag{17}$$

We optimized this cost function by alternately updating h and λ.

A flat clustering algorithm can be induced by computing B and assigning a lower level node to the parent node that has the largest connection weight. The number of clusters at any level can be determined by only counting the number of non-empty nodes (namely, the nodes that have at least one lower level node associated). To create a hierarchical graph we need to repeat this process to successively add levels of clusters to the graph. To create a bipartite graph for the next level, we first need to compute a

[14] See Yu, Yu, and Tresp (2006) for the full proof.

new similarity matrix for the clusters U. The similarity between clusters $p(u_p, u_q)$ can be induced from B, as follows:

$$p(u_p, u_q) = p(u_p)p(u_p|u_q) = (B^T D^{-1} B)_{pq} \tag{18}$$

$$D = \mathrm{diag}(d_1, ..., d_n) \text{ where } d_i = \sum_{p=1}^{m} b_{ip}$$

We can then construct a new graph G_1 (Figure 21(d)) with the clusters U as vertices, and the cluster similarities $p(u_p, u_q)$ as the connection weights. The clustering algorithm can now be applied again (Figure 21(e)). This process can go on iteratively, leading to a hierarchical graph.

The number of levels (L) and the number of clusters (m_ℓ) are detected automatically, using the method of Sun and Korhonen (2011). Clustering starts with an initial setting of number of clusters (m_0) for the first level. In our experiment, we set the value of m_0 to 800. For the subsequent levels, m_ℓ is set to the number of non-empty clusters (bipartite graph nodes) on the parent level – 1. The matrix B is initialized randomly. We found that the actual initialization values have little impact on the final result. The rows in B are normalized after the initialization so the values in each row add up to one.

For a word v_i, the probability of assigning it to cluster $x_p^{(\ell)} \in X_\ell$ at level ℓ is given by

$$\begin{aligned} p(x_p^{(\ell)}|v_i) &= \sum_{X_{\ell-1}} ... \sum_{x^{(1)} \in X_1} p(x_p^{(\ell)}|x^{(\ell-1)}) ... p(x^{(1)}|v_i) \\ &= (D_1^{(-1)} B_1 D_2^{-1} B_2 ... D_\ell^{-1} B_\ell)_{ip} \end{aligned} \tag{19}$$

m_ℓ can then be determined as the number of clusters with at least one member noun according to Equation (19). Because of the random walk property of the graph, m_ℓ is non-increasing for higher levels (Sun and Korhonen 2011). The algorithm can thus terminate when all nouns are assigned to one cluster. We run 1,000 iterations of updates of h and λ (Equations (16) and (17)) for each two adjacent levels.

The whole algorithm can be summarized as follows.

Algorithm 3 HGFC algorithm

```
Require: $N$ nouns $V$, initial number of clusters $m_1$
 Compute the similarity matrix $W_0$ from $V$
 Build the graph $G_0$ from $W_0$, $\ell \leftarrow 1$
while $m_\ell > 1$ do
    Factorize $G_{\ell-1}$ to obtain bipartite graph $K_\ell$ with adjacency matrix $B_\ell$ (eqs. 16, 17)
    Build a graph $G_\ell$ with similarity matrix $W_\ell = B_\ell^T D_\ell^{-1} B_\ell$ according to equation 18
    $\ell \leftarrow \ell + 1$ ; $m_\ell \leftarrow m_{\ell-1} - 1$
end while
return $B_\ell, B_{\ell-1} ... B_1$
```

The resulting graph is composed of a set of bipartite graphs defined by $B_\ell, B_{\ell-1}, ..., B_1$. A bipartite graph has a similar structure to the one shown in Figure 20. For a given noun, we can rank the clusters at any level according to the soft assignment probability (Equation (19)). The clusters that have no member noun were hidden from the ranking because they do not explicitly represent any concept. However, these clusters are still part of the organization of the conceptual space within the model and they contribute to the probability for the clusters at upper levels (Equation (19)). We call the view of the hierarchical graph where these empty clusters are hidden an **explicit graph**.

5.2 Identification of Metaphorical Associations

Once we obtain the explicit graph of concepts, we can now identify metaphorical associations based on the weights connecting the clusters at different levels. Taking a single noun (e.g., *fire*) as input, the system computes the probability of its cluster membership for each cluster at each level, using the weights on the edges of the graph (Equation (19)). We expect the cluster membership probabilities to indicate the level of association of the input noun with the clusters. The system can then rank the clusters at each level based on these probabilities. We chose level 3 as the optimal level of generality for our experiments, based on our qualitative analysis of the graph.[15] The system selects six top-ranked clusters from this level (we expect an average source concept to have no more than five typical target associates) and excludes the literal cluster containing the input concept (e.g., *fire flame blaze*). The remaining clusters represent the target concepts associated with the input source concept.

Example output for the input concepts of *fire* and *disease* in English is shown in Figure 22. One can see from the figure that each of the noun-to-cluster mappings represents a new conceptual metaphor (e.g., EMOTION is FIRE, VIOLENCE is FIRE, CRIME is a DISEASE). These mappings are exemplified in language by a number of metaphorical expressions (e.g., "His anger will *burn* him," "violence *flared* again," "it's time they found a *cure* for corruption"). Figures 23 and 24 show metaphorical associations identified by the Spanish and Russian systems for the same source concepts. As we can see from the figures, FEELINGS tend to be associated with FIRE in all three languages. Unsurprisingly, however, many of the identified metaphors differ across languages. For instance, VICTORY, SUCCESS, and LOOKS are viewed as FIRE in Russian, whereas IMMIGRANTS and PRISONERS have a stronger association with FIRE in English and Spanish, according to the systems. All of the languages exhibit CRIME IS A DISEASE metaphor, with Russian and Spanish also generalizing it to VIOLENCE IS A DISEASE. Interestingly, throughout our data set, Spanish data tends to exhibit more negative metaphors about CORPORATIONS, as it is demonstrated by the DISEASE example in Figure 23. Although we do not claim that this output is exhaustively representative of all conceptual metaphors present in a particular culture, we believe that these examples showcase some interesting differences in the use of metaphor across languages that can be discovered via large-scale statistical processing.

5.3 Identification of Metaphorical Expressions

After extracting the source–target domain mappings, we now move on to the identification of the corresponding metaphorical expressions. The system does this by harvesting the salient features that lead to the input noun being strongly associated with the extracted clusters. The salient features are selected by ranking the features according to the joint probability of the feature (f) occurring both with the input

[15] However, the level of granularity can be adapted depending on the task and application in mind.

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
TARGET 5: patrol militia squad warplane peacekeeper

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
TARGET 5: wind storm flood rain weather

Figure 22
Metaphorical associations discovered by the English system.

SOURCE: **fuego (fire)**
TARGET 1: esfuerzo negocio tarea debate operación operativo ofensiva gira acción actividad trabajo juicio campaña gestión labor proceso negociación
TARGET 2: quiebra indignación ira perjuicio pánico caos alarma
TARGET 3: rehén refugiado preso prisionero detenido inmigrante
TARGET 4: soberanía derecho independencia libertad autonomía
TARGET 5: referencia sustitución exilio lengua reemplazo

SOURCE: **enfermedad (disease)**
TARGET 1: calentamiento inmigración impunidad
TARGET 2: desaceleración brote fenómeno epidemia sequía violencia mal recesión escasez contaminación
TARGET 3: petrolero fabricante gigante firma aerolínea
TARGET 4: mafia
TARGET 5: hamas milicia serbio talibán

Figure 23
Metaphorical associations discovered by the Spanish system.

source noun (w) and the target cluster (c). Under a simplified independence assumption, $p(w, c|f) = p(w|f) \times p(c|f)$. $p(w|f)$ and $p(c|f)$ are calculated as the ratio of the frequency of the feature f to the total frequency of the input noun and the cluster, respectively. The features ranked higher are expected to represent the source domain vocabulary that can be used to metaphorically describe the target concepts. Example features (verbs and their grammatical relations) extracted for the source domain noun *fire* and the *violence* cluster in English are shown in Figure 25.

We then refined the lists of features by means of selectional preference (SP) filtering. Many features that co-occur with the source noun and the target cluster may be general, that is, they can describe many different domains rather than being characteristic of the source domain. For example, the verb *start*, which is a common feature for both the *fire* and the *violence* cluster (e.g., *start a war*, *start a fire*) also co-occurs with many other arguments in a large corpus. We use SPs to quantify how well the extracted features describe the source domain (e.g., *fire*) by measuring how characteristic the domain word is as an argument of the verb. This allows us to filter out non-characteristic verbs, such as *start* in our example. We extracted nominal argument distributions of the verbs in

SOURCE: огонь (fire)
TARGET 1: облик (looks)
TARGET 2: победа успех (victory, success)
TARGET 3: душа страдание сердце дух (soul, suffering, heart, spirit)
TARGET 4: страна мир жизнь россия (country, world, life, russia)
TARGET 5: множество масса ряд (multitude, crowd, range)

SOURCE: болезнь (disease)
TARGET 1: готовность соответствие зло добро (evil, kindness, readiness)
TARGET 2: убийство насилие атака подвиг поступок преступление ошибка грех нападение (murder, crime, assault, mistake, sin etc.)
TARGET 3: депрессия усталость напряжение нагрузка стресс приступ оргазм (depression, tiredness, stress etc.)
TARGET 4: сражение война битва гонка (battle, war, race)
TARGET 5: аспект симптом нарушение тенденция феномен проявление (aspect, trend, phenomenon, violation, symptom)

Figure 24
Metaphorical associations discovered by the Russian system.

rage-ncsubj *engulf*-ncsubj *erupt*-ncsubj *burn*-ncsubj *light*-dobj *consume*-ncsubj *flare*-ncsubj *sweep*-ncsubj *spark*-dobj *battle*-dobj *gut*-idobj *smolder*-ncsubj *ignite*-dobj *destroy*-idobj *spread*-ncsubj *damage*-idobj *light*-ncsubj *ravage*-ncsubj *crackle*-ncsubj *open*-dobj *fuel*-dobj *spray*-idobj *roar*-ncsubj *perish*-idobj *destroy*-ncsubj *wound*-idobj *start*-dobj *ignite*-ncsubj *injure*-idobj *fight*-dobj *rock*-ncsubj *retaliate*-idobj *devastate*-idobj *blaze*-ncsubj *ravage*-idobj *rip*-ncsubj *burn*-idobj *spark*-ncsubj *warm*-idobj *suppress*-dobj *rekindle*-dobj ...

Figure 25
Salient features for the *fire* and the *violence* cluster.

our feature lists for VERB–SUBJECT, VERB–DIRECT_OBJECT, and VERB–INDIRECT_OBJECT relations. We used the algorithm of Sun and Korhonen (2009) to create SP classes and the measure of Resnik (1993) to quantify how well a particular argument class fits the verb. Sun and Korhonen (2009) create SP classes by distributional clustering of nouns with lexico-syntactic features (i.e., the verbs they co-occur with in a large corpus and their corresponding grammatical relations). Resnik measures selectional preference strength $S_R(v)$ of a predicate as a Kullback-Leibler distance between two distributions: the prior probability of the noun class $P(c)$ and the conditional probability of the noun class given the verb $P(c|v)$.

$$S_R(v) = D(P(c|v)||P(c)) = \sum_c P(c|v) \log \frac{P(c|v)}{P(c)} \tag{20}$$

In order to quantify how well a particular argument class fits the verb, Resnik defines selectional association as

$$A_R(v, c) = \frac{1}{S_R(v)} P(c|v) \log \frac{P(c|v)}{P(c)} \tag{21}$$

We rank the nominal arguments of the verbs in our feature lists using their selectional association with the verb, and then only retain the features whose top five arguments contain the source concept. For example, the verb *start*, which is a common feature for both *fire* and the *violence* cluster, would be filtered out in this way because its top five argument classes do not contain *fire* or any of the nouns in the *violence* cluster.

In contrast, the verbs *flare* or *blaze* would be retained as descriptive source domain vocabulary.

Similarly to the spectral clustering experiments, we then search the parsed corpus for grammatical relations, in which the nouns from the target domain cluster appear with the verbs from the source domain vocabulary (e.g., "war *blazed*" (subj), "to *fuel* violence" (dobj) for the mapping VIOLENCE is FIRE in English). The system thus annotates metaphorical expressions in text, as well as the corresponding conceptual metaphors, as shown in Figure 26. Metaphorical expressions identified by the Spanish and Russian systems are shown in Figures 27 and 28, respectively.

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 (Dobj), violations *spread* (Subj) ...

Figure 26
Identified metaphorical expressions for the mappings FEELING IS FIRE and CRIME IS A DISEASE in English.

SENTIDO ES FUEGO (FEELING IS FIRE)

bombardear con indignación, *estallar* de indignación, *reavivar* indignación, *detonar* indignación, indignación *estalla*, *consumido* por pánico, *golpear* por pánico, *sacudir* por pánico, *contener* pánico, *desatar* pánico, pánico *golpea*, *consumido* por ira, *estallar* de ira, *abarcado* a ira, ira *destruya*, ira *propaga*, *encender* ira, *atizar* ira, *detonar* ira ...

CRIMEN ES ENFERMEDAD (CRIME IS A DISEASE)

tratar mafia, *erradicar* mafia, *detectar* mafia, *eliminar* mafia, *luchar contra* mafia, *impedir* mafia, *señalar* mafia, mafia *propaga*, mafia *mata*, mafia *desarrolla*, *padecer de* mafia, *debilitar por* mafia, *contaminar con* mafia ...

Figure 27
Identified metaphorical expressions for the mappings FEELING IS FIRE and CRIME IS A DISEASE in Spanish.

ЧУВСТВА -- ОГОНЬ (feeling is fire)

потушить страдания, *погасить* страдания, душа *пылает*, душа *полыхает*, душа *горит*, *зажигать* сердце, сердце *пылает*, *сжечь* сердце, сердце *зажглось*, сердце *вспыхнуло*, *разжечь* дух, дух *пылает*, *зажечь* дух

ПРЕСТУПНОСТЬ -- БОЛЕЗНЬ (crime is a disease)

выявить преступление, преступление *заразило*, *обнаружить* преступление, *провоцировать* преступление, *вызывать* убийства, *искоренить* убийства, *симулировать* убийство, *предупреждать* убийство, *излечить* насилие, *перенести* насилие, *распознать* насилие, *исцелять* грехи, *заболеть* грехом, *излечивать* грехи, *вылечить* грехи, *болеть* грехом

Figure 28
Identified metaphorical expressions for the mappings FEELING IS FIRE and CRIME IS A DISEASE in Russian.

5.4 Evaluation

Because there is no large and comprehensive gold standard of metaphorical mappings available, we evaluated the quality of metaphorical mappings and metaphorical expressions identified by the system against human judgments. We conducted two types of evaluation: (1) precision-oriented, for both metaphorical mappings and metaphorical expressions; and (2) recall-oriented, for metaphorical expressions. In the first setting, the human judges were presented with a random sample of system-produced metaphorical mappings and metaphorical expressions, and asked to mark the ones they considered valid as correct. In the second setting, 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.

5.4.1 Baselines. We compared the system performance with that of two baseline systems: an unsupervised agglomerative clustering baseline (AGG) for the three languages and a supervised baseline built upon Wordnet (WN) for English.

AGG: We constructed the agglomerative clustering baseline using SciPy implementation (Oliphant 2007) of Ward's linkage method (Ward 1963). The output tree was cut according to the number of levels and the number of clusters of the explicit graph detected by HGFC. The resulting tree was then converted into a graph by adding connections from each cluster to all the clusters one level above. We computed the connection weights as cluster distances measured using Jensen-Shannon Divergence between the cluster centroids. This graph was then used in place of the HGFC graph in the metaphor identification experiments.

WN: In the WN baseline, the WordNet hierarchy was used as the underlying graph of concepts to which the metaphor extraction method was applied. Given a source concept, the system extracted all its sense-1 hypernyms two levels above and subsequently all of their sister terms. The hypernyms themselves were considered to represent the literal sense of the source noun and were therefore removed. The sister terms were kept as potential target domains.

5.4.2 Evaluation of Metaphorical Associations. To create our data set, we extracted 10 common source concepts that map to multiple targets from the Master Metaphor List (Lakoff, Espenson, and Schwartz 1991) and linguistic analyses of metaphor (Lakoff and Johnson 1980; Shutova and Teufel 2010). These included FIRE, CHILD, SPEED, WAR, DISEASE, BREAKDOWN, CONSTRUCTION, VEHICLE, SYSTEM, BUSINESS. We then translated them into Spanish and Russian. Each of the systems and the baselines identified 50 source–target domain mappings for the given source domains. This resulted in a set of 150 conceptual metaphors for English (HGFC,AGG,WN), 100 for Spanish (HGFC,AGG), and 100 for Russian (HGFC,AGG). Each of these conceptual mappings represents a number of submappings since all the target concepts are clusters or synsets. These were then evaluated against human judgments in two different experimental settings.

Setting 1

Task and guidelines The judges were presented with a set of conceptual metaphors identified by the three systems, randomized. They were asked to annotate the mappings they considered valid as correct. In all our experiments, the judges were encouraged to rely on their own intuition of metaphor, but they also reviewed the metaphor annotation guidelines of Shutova and Teufel (2010) at the beginning of the experiment.

Participants Two judges per language, who were native speakers of English, Russian, and Spanish participated in this experiment. All of them held at least a bachelor's degree.

Interannotator agreement The agreement on this task was measured at $\kappa = 0.60$ ($n = 2, N = 150, k = 2$) for English, $\kappa = 0.59$ ($n = 2, N = 100, k = 2$) for Spanish, and $\kappa = 0.55$ ($n = 2, N = 100, k = 2$) for Russian. The main differences in the annotators' judgments stem from the fact that some metaphorical associations are less obvious and common than others, and thus need more context (or imaginative effort) to establish. Such examples where the judges disagreed included metaphorical mappings such as INTENSITY is SPEED, GOAL is a CHILD, COLLECTION is a SYSTEM, and ILLNESS is a BREAKDOWN.

Results The system performance was then evaluated against these judgments in terms of precision (P), i.e., the proportion of the valid metaphorical mappings among those identified. We calculated system precision (in all experiments) as an average over both annotations. The results across the three languages are presented in Table 6.

Setting 2
To measure recall, R, of the systems we asked two annotators per language (native speakers with a background in metaphor, different from Setting 1) to write down up to five target concepts they strongly associated with each of the 10 source concepts. Their annotations were then aggregated into a single metaphor association gold standard, including all of the mappings listed by the annotators. The gold standard consisted of 63 mappings for English, 70 mappings for Spanish, and 68 mappings for Russian. The recall of the systems was measured against this gold standard. The results are shown in Table 6.

5.4.3 Evaluation of Metaphorical Expressions. For each of the identified conceptual metaphors, the systems extracted a number of metaphorical expressions from the corpus. For the purposes of this evaluation, we selected the top 50 features from the ranked feature list (as described in Section 5.3) and searched the corpus for expressions where the verbs from the feature list co-occurred with the nouns from the target cluster. Figure 29 shows example sentences annotated by HGFC for English. The identification of metaphorical expressions was also evaluated against human judgments.

Materials The judges were presented with a set of randomly sampled sentences containing metaphorical expressions as annotated by the systems and by the baselines (200 each). This resulted in a data set of 600 sentences for English (HGFC, AGG, WN), 400 sentences for Spanish (HGFC, AGG), and 400 sentences for Russian (HGFC, AGG). The order of the presented sentences was randomized.

Table 6
HGFC and baseline performance in the identification of metaphorical associations.

System	AGG		WN		HGFC	
	Precision	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
English	0.36	0.11	0.29	0.03	0.69	0.61
Spanish	0.23	0.12	-	-	0.59	0.54
Russian	0.28	0.09	-	-	0.62	0.42

Task and guidelines The judges were asked to mark the expressions that were metaphorical in their judgment as correct, following the same guidelines as in the spectral clustering evaluation.

Participants Two judges per language, who were native speakers of English, Russian, and Spanish, participated in this experiment. All of them held at least a bachelor's degree.

Interannotator agreement Their agreement on the task was measured at $\kappa = 0.56$ ($n = 2, N = 600, k = 2$) for English, $\kappa = 0.52$ ($n = 2, N = 400, k = 2$) for Spanish, and $\kappa = 0.55$ ($n = 2, N = 400, k = 2$) for Russian.

Results The system performance was measured against these annotations in terms of an average precision across judges. The results are presented in Table 7. HGFC outperforms both AGG and WN, yielding a precision of 0.65 in English, 0.54 in Spanish, and 0.59 in Russian.

5.5 Discussion and Error Analysis

As expected, HGFC outperforms both AGG and WN baselines in all evaluation settings. AGG has been previously shown to be less accurate than HGFC in the verb clustering task (Sun and Korhonen 2011). Our analysis of the noun clusters indicated that HGFC tends to produce more pure and complete clusters than AGG. Another important reason AGG fails is that it by definition organizes all concepts into a tree and optimizes its solution locally, taking into account a small number of clusters at a time. However, being able to discover connections between more distant domains and optimizing globally over all concepts is crucial for metaphor identification. This makes AGG less suitable for the task, as demonstrated by our results. However, AGG identified a number of interesting mappings missed by HGFC (e.g. CAREER IS A CHILD, LANGUAGE IS A SYSTEM, CORRUPTION IS A VEHICLE, EMPIRE IS A CONSTRUCTION), as well as a number of mappings in

EG0 275 In the 1930s the words *means test* was a curse, ***fuelling*** **the resistance** against it both among the unemployed and some of its administrators.
CRX 1054 These problems would be serious enough even if the rehabilitative approach were demonstrably successful in ***curing*** **crime**.
HL3 1206 [..] he would strive to ***accelerate*** **progress** towards the economic integration of the Caribbean.
HXJ 121 [..] it is likely that some **industries will** ***flourish*** in certain countries as the **market** ***widens***.
CEM 2622 The attack in Bautzen, Germany, came as racial **violence** ***flared*** again.

Figure 29
Metaphors tagged by the English HGFC system (in bold).

Table 7
HGFC and baseline precision in the identification of metaphorical expressions.

System	AGG	WN	HGFC
English	0.47	0.12	0.65
Spanish	0.38	-	0.54
Russian	0.40	-	0.59

common with HGFC (e.g. DEBATE IS A WAR, DESTRUCTION IS A DISEASE). The fact that both HGFC and AGG identified valid metaphorical mappings across languages confirms our hypothesis that clustering techniques are well suited to detect metaphorical patterns in a distributional word space.

The WN system also identified a few interesting metaphorical mappings (e.g., COGNITION IS FIRE, EDUCATION IS CONSTRUCTION), but its output is largely dominated by the concepts similar to the source noun and contains some unrelated concepts. The comparison of HGFC to WN shows that HGFC identifies meaningful properties and relations of abstract concepts that cannot be captured in a tree-like classification (even an accurate, manually created one such as WordNet). The latter is more appropriate for concrete concepts, and a more flexible representation is needed to model abstract concepts. The fact that both baselines identified some valid metaphorical associations, relying on less suitable conceptual graphs, suggests that our way of traversing the graph is a viable approach to identification of metaphorical associations in principle.

HGFC identifies valid metaphorical associations for a range of source concepts. One of them (CRIME IS A DISEASE, or CRIME IS A VIRUS) happened to have been already validated in behavioral experiments with English speakers (Thibodeau and Boroditsky 2011). The most frequent type of error of HGFC across the three languages is the presence of target clusters similar or closely related to the source noun. For instance, the source noun CHILD tends to be linked to other "human" clusters across languages—for example, the *parent* cluster for English; the *student*, *resident*, and *worker* clusters in Spanish, and the *crowd*, *journalist*, and *emperor* clusters in Russian. The clusters from the same domain can, however, be filtered out if their nouns frequently occur in the same documents with the source noun (in a large corpus), that is, by topical similarity. The latter is less likely to be the case for the metaphorically associated nouns. However, we leave such an experiment to future work.

The system errors in the identification of metaphorical expressions stem from multiple word senses of the salient features or the source and target sharing some physical properties (e.g., one can "*die from* crime" and "*die from* a disease," an error that manifested itself in all three languages). Some identified expressions invoke a chain of mappings (e.g., ABUSE IS A DISEASE, DISEASE IS AN ENEMY for "*combat* abuse"); however, such chains are not yet incorporated into the system. In some cases, the same salient feature could be used metaphorically both in the source and target domain (e.g., "to *open* fire" vs. "to *open* one's heart" in Russian). In this example the expression is correctly tagged as metaphorical, although representing a different conceptual metaphor than FEELING IS FIRE. The performance of AGG in the identification of metaphorical expressions is higher than in the identification of metaphorical associations, because it outputs only few expressions for the incorrect associations. In contrast, WN tagged a large number of literal expressions due to the incorrect prior identification of the underlying associations.

The performance of the Russian and Spanish systems is slightly lower than that of the English system. This is likely to be due to errors from the data preprocessing step (i.e., parsing). The quality of parser output in English is likely to be higher than in Russian or Spanish, for which fewer parsers exist. Another important difference lies in the corpora used. Whereas the English and Spanish systems have been trained on English and Spanish Gigaword corpora (containing data extracted from news sources), the Russian system has been trained on RuWaC, which is a Web corpus containing a greater amount of noisy text (including misspellings, slang, etc.) The difference in corpora is also likely to have an impact on the mappings identified—that is, different target domains and different metaphorical mappings may be prevalent in different types of data. However, because our goal is to test the capability of clustering techniques to identify metaphorical associations and expressions in principle, the specific types of metaphors identified from different corpora (e.g., the domains covered) are less relevant.

Importantly, our results show that the method is portable across languages. This is an encouraging result, particularly because HGFC is unsupervised, making metaphor processing technology available to a large number of languages for which metaphor-annotated data sets and lexical resources do not exist.

6. Cross-Linguistic Analysis and Metaphor Variation

By automatically discovering metaphors in a data-driven way, our methods allow us to investigate and compare the semantic spaces of different languages and gain insights for cross-linguistic research on metaphor. The contrastive study of differences in metaphor is important for several reasons. Understanding how metaphor varies across languages could provide clues about the roles of metaphor and cognition in structuring each other (Kövecses 2004). Contrastive differences in metaphor also have implications for second-language learning (Barcelona 2001), and thus a systematic understanding of variation of metaphor across languages would benefit educational applications. From an engineering perspective, metaphor poses a challenge for machine translation systems (Zhou, Yang, and Huang 2007; Shutova, Teufel, and Korhonen 2013), and can even be difficult for human translators (Schäffner 2004).

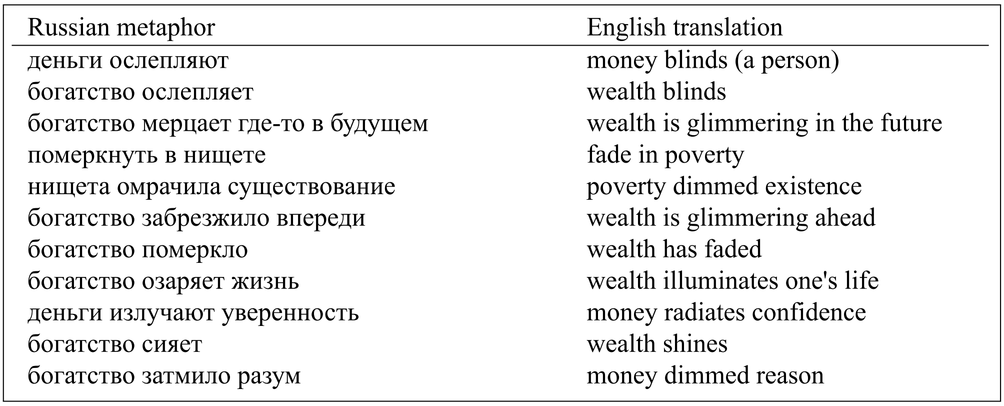
Although some aspects of the way that metaphor structures language may be widely shared and near-universal (Kövecses 2004), there are significant differences in how conventionalized and pervasive different metaphors are in different languages and cultures. The earliest analyses of cross-lingual metaphorical differences were essentially qualitative.[16] In these studies, the authors typically produce examples of metaphors that they argue are routine and widely used in one language, but unconventionalized or unattested in another language. Languages that have been studied in such a way include Spanish (Barcelona 2001), Chinese (Yu 1998), Japanese (Matsuki 1995), and Zulu (Taylor and Mbense 1998). One drawback of these studies is that they rely on the judgment of the authors, who may not be representative of the speakers of the language at large. They also do not allow for subtler differences in metaphor use across languages to be exposed. One possibility for addressing these shortcomings involves manually searching corpora in two languages and counting all instances of a metaphorical mapping. This is the approach taken by Charteris-Black and Ennis (2001) with respect to financial metaphors in English and Spanish. They find several metaphors that are much more common in one language than in the other. However, the process of manually identifying instances is time-consuming and expensive, limiting the size of corpora and the scope of metaphors that can be analyzed in a given time frame. As a result, it can be difficult to draw broad conclusions from these studies.

Our systems present a step towards a large-scale data-driven analysis of linguistic variation in the use of metaphor. In order to investigate whether statistically learned patterns of metaphor can capture such variation, we conducted an analysis of the metaphors identified by our systems in the three languages. We ran the HGFC systems with a larger set of source domains taken from the literature on metaphor and conducted a

[16] See Kövecses (2004) for a review.

qualitative analysis of the resulting metaphorical mappings to identify the similarities and the differences across languages. As one might expect, the majority of metaphorical mappings identified by the systems are present across languages. For instance, VIOLENCE and FEELINGS are associated with FIRE in all three languages, DEBATE or ARGUMENT are associated with WAR, CRIME is universally associated with DISEASE, MONEY with LIQUID, and so on. However, although the instances of a conceptual metaphor may be present in all three languages, interestingly, it is often the case that the same conceptual metaphor is lexicalized differently in different languages. For instance, although FEELINGS are generally associated with LIQUIDS in both English and Russian, the expression "*stir* excitement" is English-specific and cannot be used in Russian. At the same time, the expression "*mixed* feelings" (another instantiation of the same conceptual metaphor) is common in both languages. Our systems allow us to trace such variation through the different metaphorical expressions that they identify for the same or similar conceptual metaphors.

Importantly, besides the linguistic variation our methods are also able to capture and generalize conceptual differences in metaphorical use in the three languages. For instance, they exposed some interesting cross-linguistic differences pertaining to the target domains of *business* and *finance*. The Spanish conceptual metaphor output manifested rather negative metaphors about business, market, and commerce: BUSINESS was typically associated with BOMB, FIRE, WAR, DISEASE, and ENEMY. Although it is the case that BUSINESS is typically discussed in terms of a WAR or a RACE in English and Russian, the other four Spanish metaphors are uncommon. Russian, in fact, has rather positive metaphors for the related concepts of MONEY and WEALTH, which are strongly associated with SUN, LIGHT, STAR, and FOOD, possibly indicating that money is viewed primarily as a way to improve one's own life. An example of the linguistic instantiations of the Russian MONEY is LIGHT metaphor and their corresponding word-for-word English translations is shown in Figure 30. We have validated that the word-for-word English translations of the Russian expressions in the Figure are not typically used in English by searching the BNC, where none of the expressions were found. In contrast, in English, MONEY is frequently discussed as a WEAPON, that is, a means to achieve a goal or win a struggle (which is directly related to BUSINESS IS A WAR metaphor). At the same time, the English data exhibit positive metaphors for POWER and INFLUENCE, which are viewed as LIGHT, SUN, or WING. In Russian, on the contrary,



Russian metaphor	English translation
деньги ослепляют	money blinds (a person)
богатство ослепляет	wealth blinds
богатство мерцает где-то в будущем	wealth is glimmering in the future
померкнуть в нищете	fade in poverty
нищета омрачила существование	poverty dimmed existence
богатство забрезжило впереди	wealth is glimmering ahead
богатство померкло	wealth has faded
богатство озаряет жизнь	wealth illuminates one's life
деньги излучают уверенность	money radiates confidence
богатство сияет	wealth shines
богатство затмило разум	money dimmed reason

Figure 30
Linguistic instantiations of MONEY IS LIGHT metaphor in Russian.

POWER is associated with BOMB and BULLET, perhaps linking it to the concepts of physical strength and domination. The concepts of FREEDOM and INDEPENDENCE were also associated with a WING, WEAPON, and STRENGTH in the Russian data, however. English and Spanish data also exhibited interesting differences with respect to the topic of immigration. According to the system output, in English IMMIGRANTS tend to be viewed as FIRE or ENEMIES, possibly indicating danger. In Spanish, on the other hand, IMMIGRANTS and, more specifically, undocumented people have a stronger association with ANIMALS, which is likely a reference to them as victims, being treated like animals.

Although these differences may be a direct result of the contemporary socio-economic context and political rhetoric, and are likely to change over time, other conceptual differences have a deeper grounding in our culture and way of life. For instance, the concept of BIRTH tends to be strongly associated with LIGHT in Spanish and BATTLE in Russian, each metaphor highlighting a different aspect of birth. The differences that stem from highly conventional metaphors seem to be even more deeply entrenched in the conceptual system of the speakers of a language. For instance, our analysis of system-produced data revealed systematic differences in discussing quantity and intensity in the three languages. Let us consider, for instance, the concept of *heat*. In English, heat intensity is typically measured on a vertical scale; for example, it is common to say "*low* heat" and "*high* heat." In Russian, heat intensity is rather thought of in terms of strength; for example, one would say "*strong* heat" or "*weak* fire." As opposed to this, Spanish speakers talk about heat in terms of its speed; for example, "fuego *lento*" (literally "*slow* fire") refers to "*low* heat" (on the stove). This metaphor also appears to generalize to other phenomena whose level or quantity can be assessed (e.g., INTELLIGENCE is also discussed in terms of SPEED in Spanish, HEIGHT in English, and STRENGTH in Russian). Such a systematic variation provides new insights for the study of cognition of quantity, intensity, and scale. Statistical methods provide a tool to expose such variation through automatic analysis of large quantities of linguistic data.

More generally, such systematic cross-linguistic differences in the use of metaphor have significance beyond language and can be associated with contrastive behavioral patterns across the different linguistic communities (Casasanto and Boroditsky 2008; Fuhrman et al. 2011). Psychologists Thibodeau and Boroditsky (2011) investigated how the metaphors we use affect our decision-making. They presented two groups of human subjects 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. As a result, the first group tended to opt for preventive measures in tackling crime (e.g., stronger social policies), whereas the second group converged on punishment- or restraint-oriented measures. According to the researchers, their results demonstrate that metaphors have profound influence on how we conceptualize and act with respect to societal issues. Although Thibodeau and Boroditsky's study did not investigate cross-linguistic contrasts in the use of metaphor, it still suggests that metaphor-induced differences in decision-making may manifest themselves across communities. Applying data-driven methods such as ours to investigate variation in the use of metaphor across (linguistic) communities would allow this research to be scaled-up, using statistical patterns learned from linguistic data to inform experimental psychology.

7. Conclusions and Future Directions

We have presented three methods for metaphor identification that acquire metaphorical patterns from distributional properties of concepts. All of the methods

(UNCONSTRAINED, CONSTRAINED, HGFC) are based on distributional word clustering using lexico-syntactic features. The methods are minimally supervised and unsupervised and, as our experiments have shown, they can be successfully ported across languages. Despite requiring little supervision, their performance is competitive even in comparison to fully supervised systems.[17] In addition, the methods identify a large number of new metaphorical expressions in corpora (e.g., given the English seed "*accelerate* change," the UNCONSTRAINED method identifies as many as 113 new, different metaphors in the BNC), enabling large-scale cross-linguistic analyses of metaphor.

Our experimental results have demonstrated that lexico-syntactic features are effective for clustering and metaphor identification in all three languages. However, we have also identified important differences in the structure of the semantic spaces across languages. For instance, in Russian, a morphologically rich language, the semantic space is structured differently from English or Spanish. Because of its highly productive derivational morphology, Russian exhibits a higher number of near-synonyms (often originating from the same stem) for both verbs and nouns. This has an impact on clustering, in that (1) more nouns or verbs need to be clustered in order to represent a concept with sufficient coverage and (2) the clusters need to be larger, often containing tight subclusters of derivational word forms. While playing a role in metaphor identification, this finding may also have implications for other multilingual NLP tasks beyond metaphor research.

Importantly, our results confirm the hypothesis that metaphor and cross-domain vocabulary projection are naturally encoded in the distributional semantic spaces in all three languages. As a result, metaphorical mappings could be learned from distributional properties of concepts using clustering techniques. The differences in performance across languages are mainly explained by the differences in the quality of the data and pre-processing tools available for them. However, both our quantitative results and the analysis of the system output confirm that all systems successfully discover metaphorical patterns from distributional information.

We have investigated different kinds of supervision: learning from a small set of metaphorical expressions, metaphorical mappings, and without supervision. Although both minimally supervised (UNCONSTRAINED, CONSTRAINED) and unsupervised (HGFC) methods successfully discover new metaphorical patterns from the data, our results indicate that minimally supervised methods achieve a higher precision. The use of annotated metaphorical mappings for supervision at the clustering stage does not significantly alter the performance of the system, because their patterns are already to a certain extent encoded in the data and can be learned. However, metaphorical expressions are a good starting point in learning metaphorical generalizations in conjunction with clustering techniques.

Despite its comparatively lower performance, we believe that HGFC may prove to be a practically useful tool for NLP applications. Because it does not require any metaphor annotation, it can be easily applied to a new language (including low resource languages) for which a large enough corpus and a shallow syntactic parser are available. In addition, whereas the semi-supervised CONSTRAINED and UNCONSTRAINED methods discover metaphorical expressions somewhat related to the seeds, the range of metaphors discovered by HGFC is unrestricted and thus considerably wider. Since the two types of methods differ in their precision vs. their coverage, one may also consider a

[17] The precision typically reported for supervised metaphor identification is in the range of 0.56–0.78, with the highest performing systems frequently evaluated within a limited domain (Shutova 2015).

combination of these methods when designing a metaphor processing component for a real-world application—or, depending on the needs of the application, one may choose a more suitable one.

In the future, the models need to be extended to identify not only verb–subject and verb–object metaphors, but also metaphorical expressions in other syntactic constructions (e.g., adjectival or nominal metaphors). Previous distributional clustering and lexical acquisition research has shown that it is possible to model the meanings of a range of word classes using similar techniques (Hatzivassiloglou and McKeown 1993; Boleda Torrent and Alonso i Alemany 2003; Brockmann and Lapata 2003; Zapirain, Agirre, and Màrquez 2009). We thus expect our methods to be equally applicable to metaphorical uses of other word classes and syntactic constructions. For spectral clustering systems, such an extension would require incorporation of adjectival and nominal modifier features in clustering, clustering adjectives, and adding seed expressions representing a variety of syntactic constructions. The extension of HGFC would be more straightforward, only requiring ranking additional adjectival and nominal features that the metaphorically associated clusters in the graph share.

The results of our HGFC experiments also offer support to the cognitive science findings on the differences in organization of abstract and concrete concepts in the human brain (Crutch and Warrington 2005; Wiemer-Hastings and Xu 2005; Huang, Lee, and Federmeier 2010; Adorni and Proverbio 2012). Specifically, our experiments have shown that abstract concepts exhibit both within-domain and cross-domain association patterns (i.e., the literal ones and the metaphorical ones) and that the respective patterns can be successfully learned from linguistic data via the words' distributional properties. The metaphorical patterns that the system is able to acquire (for different languages or different data sets) can in turn be used to guide further cognitive science and psychology research on metaphor and concept representation more generally. In addition, we believe that the presented techniques may have applications in NLP beyond metaphor processing and would impact a number of tasks in computational semantics that model the properties of and relations between concepts in a distributional space.

Acknowledgments

We would like to thank our anonymous reviewers for their most insightful comments. Ekaterina Shutova's research is supported by the Leverhulme Trust Early Career Fellowship.

References

Adorni, Roberta and Alice Mado Proverbio. 2012. The neural manifestation of the word concreteness effect: An electrical neuroimaging study. *Neuropsychologia*, 50(5):880–891.

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.

Ballesteros, Miguel, Jesús Herrera, Virginia Francisco, and Pablo Gervás. 2010. A feasibility study on low level techniques for improving parsing accuracy for Spanish using MaltParser. In *Proceedings of the 6th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications*, pages 39–48, Athens.

Barcelona, Antonio. 2001. On the systematic contrastive analysis of conceptual metaphors: Case studies and proposed methodology. In Martin Pütz, Susanne Niemeier, René Dirven (editors), *Applied Cognitive Linguistics II: Language Pedagogy*. Mouton-De Gruyter, Berlin, pages 117–146.

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

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.

Beigman Klebanov, Beata, Ben Leong, Michael Heilman, and Michael Flor. 2014. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17, Baltimore, MD.

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

Binder, Jeffrey R., Chris F. Westbury, Kristen A. McKiernan, Edward T. Possing, and David A. Medler. 2005. Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, 17(6):905–917.

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.

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

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

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

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

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

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

Burnard, Lou. 2007. *Reference Guide for the British National Corpus (XML Edition).*

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.

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

Casasanto, Daniel and Lera Boroditsky. 2008. Time in the mind: Using space to think about time. *Cognition*, 106(2):579–593.

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.

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

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

Crutch, Sebastian J. and Elizabeth K. Warrington. 2005. Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3):615–627.

Crutch, Sebastian J. and Elizabeth K. Warrington. 2010. The differential dependence of abstract and concrete words upon associative and similarity-based information: Complementary semantic interference and facilitation effects. *Cognitive Neuropsychology*, 27(1):46–71.

Devereux, Barry and Fintan Costello. 2005. Propane stoves and gas lamps: How the concept hierarchy influences the interpretation of noun–noun compounds. In *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*, pages 1–6, Streso.

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. 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.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Ferrer, Eva E. 2004. Towards a semantic classification of Spanish verbs based on subcategorisation information. In *Proceedings of the ACL 2004 Workshop on Student Research*, pages 13–19, Barcelona.

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

Fuhrman, Orly, Kelly McCormick, Eva Chen, Heidi Jiang, Dingfang Shu, Shuaimei Mao, and Lera Boroditsky. 2011. How linguistic and cultural forces shape conceptions of time: English and Mandarin time in 3D. *Cognitive Science*, 35:1305–1328.

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, NY.

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.

Graff, David, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. Linguistic Data Consortium, Philadelphia.

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.

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

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.

Hill, Felix, Anna Korhonen, and Christian Bentz. 2014. A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive Science*, 38(1):162–177.

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.

Huang, Hsu-Wen, Chia-Lin Lee, and Kara D. Federmeier. 2010. Imagine that! ERPs provide evidence for distinct hemispheric contributions to the processing of concrete and abstract concepts. *NeuroImage*, 49(1):1116–1123.

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.

Ji, Xiang, Wei Xu, and Shenghuo Zhu. 2006. Document clustering with prior knowledge. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 405–412, Seattle, WA.

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 and 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.

Kövecses, Zoltán. 2004. Introduction: Cultural variation in metaphor. *European Journal of English Studies*, 8:263–274.

Kovecses, Zoltan. 2005. *Metaphor in Culture: Universality and Variation*. Cambridge University Press.

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

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

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.

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. 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, Los Angeles, CA.

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.

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.

Martin, James. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press, 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.

Matsuki, Keiko. 1995. Metaphors of anger in Japanese. In John Taylor and Robert MacLaury, editors, *Language and the Cognitive Construal of the World*. Gruyter, Berlin.

Mendonca, Angelo, Daniel Jaquette, David Graff, and Denise DiPersio. 2011. Spanish Gigaword Third Edition. Linguistic Data Consortium, Philadelphia.

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.

Mohler, Michael, Bryan Rink, David Bracewell, and Marc Tomlinson. 2014. A novel distributional approach to multilingual conceptual metaphor recognition. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1752–1763, Dublin.

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.

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.

Ng, Andrew Y., Michael I. Jordan, Yair Weiss et al. 2002. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 2:849–856.

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 2003 International Conference on Information and Knowledge Engineering (IKE 03)*, pages 412–416, Las Vegas, NV.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Oliphant, Travis E. 2007. Python for scientific computing. *Computing in Science and Engineering*, 9:10–20.

Pantel, Patrick and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22:1–39.

Resnik, Philip. 1993. *Selection and Information: A Class-based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

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

Schäffner, Christina. 2004. Metaphor and translation: Some implications of a cognitive approach. *Journal of Pragmatics*, 36:1253–1269.

Schulte im Walde, Sabine and Chris Brew. 2001. Inducing German semantic verb classes from purely syntactic subcategorisation information. In *ACL '02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 223–230, Morristown, NJ.

Sharoff, Serge. 2006. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working Papers on the Web as Corpus*, pages 657–670, Moscow.

Sharoff, Serge and Joakim Nivre. 2011. The proper place of men and machines in language technology processing Russian without any linguistic knowledge. In *Dialogue 2011, Russian Conference on Computational Linguistics*, pages 591–605, Moscow.

Shi, J. and J. Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

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. 2015. Design and Evaluation of Metaphor Processing Systems. *Computational Linguistics*, 41(4):579–623.

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 COLING 2010*, pages 1002–1010, 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.

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

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.

Stevenson, Suzanne and Eric Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of HLT-NAACL 2003*, pages 71–78, Edmonton.

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.

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

Sun, Lin and Anna Korhonen. 2011. Hierarchical verb clustering using graph factorization. In *Proceedings of EMNLP*, pages 1023–1033, Edinburgh.

Taylor, John and Thandi Mbense. 1998. *Red Dogs and Rotten Mealies: How Zulus Talk About Anger*, volume Speaking of Emotions. Gruyter, Berlin.

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, 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.

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.

Von Luxburg, Ulrike. 2007. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wagner, Dorothea and Frank Wagner. 1993. Between min cut and graph bisection. Volume 711 of *Lecture Notes in Computer Science*. Springer, pages 744–750.

Ward, Joe H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Wiemer-Hastings, Katja and Xu Xu. 2005. Content differences for abstract and concrete concepts. *Cognitive Science*, 29(5):719–736.

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.

Yu, Kai, Shipeng Yu, and Volker Tresp. 2006. Soft clustering on graphs. In *Proceedings of Advances in Neural Information Processing Systems*, 18, Vancouver.

Yu, Ning. 1998. *The Contemporary Theory of Metahpor in Chinese: A Perspective from Chinese*. John Benjamins, Amsterdam.

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

Zhou, Chang-Le, Yun Yang, and Xiao-Xi Huang. 2007. Computational mechanisms for metaphor in languages: A survey. *Journal of Computer Science and Technology*, 22:308–319.