UNIVERSITY OF CALIFORNIA, SAN DIEGO

Computational and Cognitive Semantics: Three Investigations into the Statistics and Structure of Meaning in Language

A Dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Cognitive Science

by

Elkin Darío Gutiérrez

Committee in charge:

Professor Benjamin K. Bergen, Chair
Professor Seana Coulson
Professor Virginia de Sa
Professor Andrew Kehler
Professor Zhuowen Tu

2016

The Dissertation of Elkin Darío Gutiérrez is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

____

Chair

University of California, San Diego

2016

DEDICATION

This dissertation is hereby dedicated.

EPIGRAPH

Waaahhh! Laaalalalaaa!

—My son

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Large portions of the research presented in this thesis were performed in collaboration with Ben Bergen, Gerard de Melo, Luca Gilardi, Patricia Lichtenstein, and Tyler Marghetis, and Ekaterina Shutova; this would would absolutely not be possible without their fantastic ideas. Thanks also to Marta Kutas, Jaime Alexandre, Andrew Schork, Charles Elkan, and the members of my dissertation committee for invaluable advice. I especially thank Ben Bergen for being a fantastic advisor: always ready not only with interesting new ideas and careful insights, but also with hilarious zingers and helpful parenting advice. I'd also like to thank him for his patience and understanding through all these years. Additionally, I would like to thank the anonymous reviewers at TACL and ACL for their helpful insights and comments on drafts of the work in this thesis.

Chapter 2, in part, is a reprint of the material as it appears in Gutierrez, E.D.; Shutova, Ekaterina; Marghetis, Tyler; Bergen, Benjamin. "Literal and Metaphorical Senses in Compositional Distributional Semantic Models", Proceedings of the Association for Computational Linguistics, 2016. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, is a reprint of the material as it appears in Gutierrez, E.D.; Levy, Roger; Bergen, Benjamin. "Finding Non-Arbitrary Form-Meaning Systematicity Using String-Metric Learning for Kernel Regression", Proceedings of the Association for Computational Linguistics, 2016. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, is a reprint of the material as it appears in Gutierrez, E.D.; Shutova, Ekaterina; Lichtenstein, Patricia; de Melo, Gerard; Gilardi, Luca.

"Detecting cross-cultural differences using a probabilistic topic model", Transactions of the Association for Computational Linguistics, vol. 4, 2016. The dissertation author was the primary investigator and author of this paper.

VITA

2004	B. A. in Mathematics and Economics , Columbia College, Columbia University
2010	M. S. in Cognitive Science, University of California, San Diego
2016	Ph. D. in Cognitive Science, University of California at San Diego

PUBLICATIONS

E. Darío Gutiérrez, Ekaterina Shutova, Tyler Marghetis, Benjamin Bergen, “Literal and Metaphorical Senses in Compositional Distributional Semantic Models”, *Proceedings of the Association for Computational Linguistics*, To appear, 2016.

E. Darío Gutiérrez, Roger Levy, Benjamin Bergen, “Finding Non-Arbitrary Form-Meaning Systematicity Using String-Metric Learning for Kernel Regression”, *Proceedings of the Association for Computational Linguistics*, To appear, 2016.

E. Darío Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, Luca Gilardi, “Detecting cross-cultural differences using a probabilistic topic model”, *Transactions of the Association for Computational Linguistics*, **4**: 47-60, 2016.

Ekaterina Shutova, Lin Sun, E.D. Guti’errez, et al. (2016). Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning *Computational Linguistics*. To Appear.

Chi-Hua Chen, Mark Fiecas, E.D. Gutiérrez, et al. (2014). Genetic topography of brain morphology. *Proceedings of the National Academy of Sciences (PNAS).* **11**: 17089-17094.

Chi-Hua Chen, E.D. Gutirrez, W.K. Thompson, et al. (2012). Hierarchical genetic organization of human cortical surface area. *Science*. **335**: 1634-1636.

Bruno Galantucci, Carrie A. Theisen, E.D. Guti’errez, Christian Kroos, Theo Rhodes (2012). The diffusion of novel signs beyond the dyad. *Language Sciences* **34**: 583-590.

ABSTRACT OF THE DISSERTATION

Computational and Cognitive Semantics: Three Investigations into the Statistics and Structure of Meaning in Language

by

Elkin Darío Gutiérrez

Doctor of Philosophy in Cognitive Science

University of California, San Diego, 2016

Professor Benjamin K. Bergen, Chair

This dissertation presents the results of three projects at the intersection of computational semantics and cognitive semantics. The theme underlying all three projects is that closer coordination between cognitive semantics and computational semantics can lead to more accurate computational semantic tools, and can also provide tools that can help address questions in cognitive semantics.

Chapter 1

Introduction

This dissertation presents the results of three independent projects. The common thread across these projects is that they reside at the intersection of computational semantics and cognitive semantics, drawing upon the theoretical resources and techniques of both fields. This is an especially fruitful intersection at which to work, because these fields face similar challenges, yet have communicated relatively little. This presents opportunities on two fronts.

On the one hand, many problems in cognitive semantics relate to large-scale (language-wide or cross-linguistic) phenomena. Exhaustive, systematic corpus analysis is one method that can be used to place hypotheses about such phenomena on a solid footing. Unfortunately, such analysis is usually prohibitively expensive in terms of both time and money. The plethora of new machine learning techniques that have emerged in recent years have the potential to reshape corpus analysis via automation.

Meanwhile, computational semantics has generally focused on straightforwardly applying new machine-learning techniques, with relatively little emphasis on existing theory in cognitive semantics. Indeed, much of the substantial progress made in computational semantics in the past 15 years rests largely on a single idea from cognitive semantics, the distributional hypothesis [Har54]. While it is impressive that a single, 60-year-old concept could lead to so much practical progress, it raises the question of how much more rapidly the field could advance if armed with the full arsenal of empirical results from cognitive semantics.

1.1 A Brief Introduction to Distributional Semantics

The distributional hypothesis asserts that the meaning of a word is reflected in how the word is used in context: Words that have similar meanings are used in similar contexts. As mentioned above, the importance of this hypothesis to computational semantics lies in that it opens a window of empirical inquiry into linguistic meaning through the powerful machinery of statistics and probability theory: We can compile statistics of how a word is distributed across many documents. Using these statistics we can build a semantic vector for the word, a numerical representation of the approximate meaning of the word. Ideally, the closer that the semantic vectors for two words are, the closer that their meanings are, and *vice versa*.

Figure 1.1 illustrates how the distributional hypothesis can be used to build a distributional semantic model. Here we have a toy semantic space with three axes: each one represents a context in which a word can appear. The word *cup* occurs often near the words *wine* and *drink*, so in this toy distributional semantic model (or *DSM* for short), it is represented by a semantic vector that is a mixture of these two corresponding context axes. Conversely, *cup* doesn't often occur in the context of the

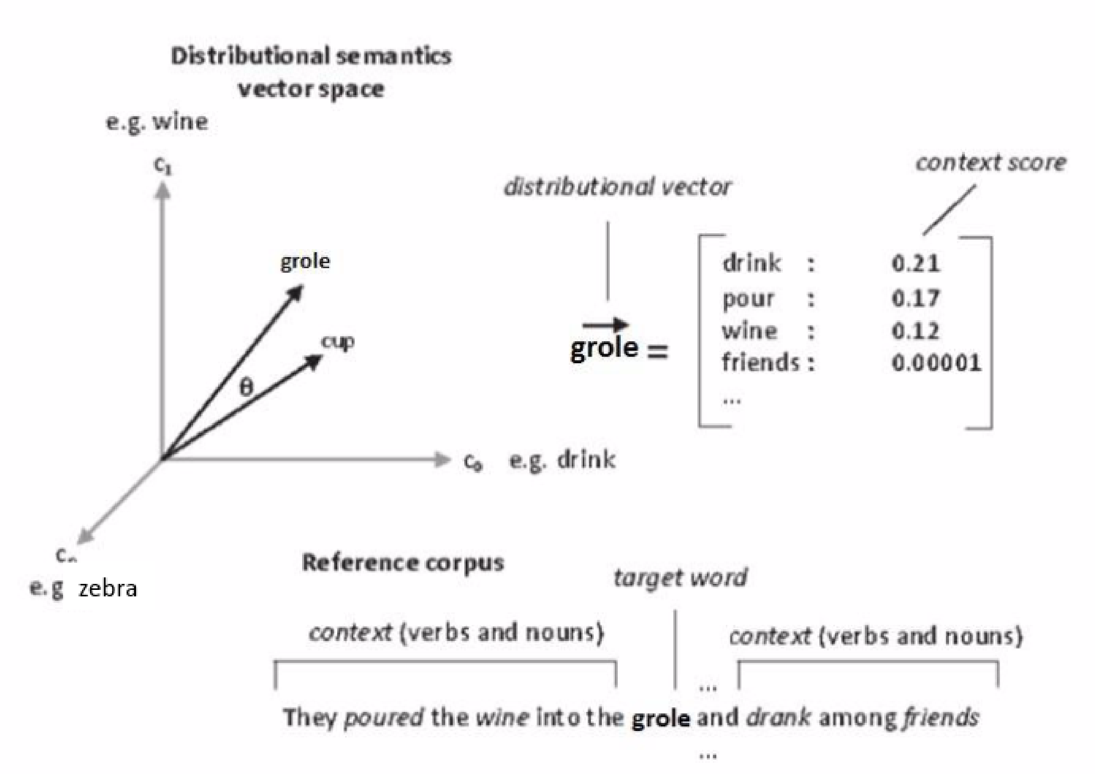

Figure 1.1: Illustration of the use of a word's distributional statistics to create semantic vectors for the word.

word *zebra*, so the magnitude of the *cup*'s semantic vector in the *zebra* context axis would be near zero. The DSM could be updated with a sentence including a totally novel word that is not present in its original training corpus. Suppose the novel word is *grole*, and the phrase is *They poured the wine into the grole and drank among friends*. Then we might reasonably guess that *grole* has a meaning (and a semantic vector) that is close to that of cup in semantic space, because its is used in similar contexts. Indeed, a *grole* looks like a cup in that you can drink out of it, but it has multiple spouts for sharing. Note that the sorts of context axes in the figure are not usually used directly, since they result in a very high-dimensional space. Usually, dimensionality reduction techniques are used to obtain a dense, compact vector representation of each word. There are many dimensionality reduction techniques that have been proposed;

the major categories are explicit matrix factorization/spectral techniques [DDL+90], Bayesian generative models [BNJ03], and neural-network [MSC+13] approaches; see Turney and Pantel [TP10] for a thorough survey. DSMs also differ in how to define the context of a word. For instance, Latent Semantic Analysis [DDL+90, LD97], one of the first matrix factorization DSMs, defines each document as a context—a word occurs in a document context if it is instantiated in that document. Meanwhile, HAL [LBA95], a later matrix factorization-based technique, defined contexts as words; HAL counts a word as occurring within a given context if it occurred within a window of N words of the context word (for some integer N defined by the user). While most DSM approaches ignore syntax (and are known as *bag-of-words models*), some models even integrate syntactic information, for instance by taking account of the syntactic relation between a term and its context [EP08], or, as described in chapter 2, by learning vector representations optimized over syntactic trees [SPH+11].

1.2 Outline of the Rest of This Thesis

The projects described in this dissertation provide a small illustration of the potential of working at the intersection of computational and cognitive semantics.

Chapter 3, which was completed in collaboration with Benjamin Bergen and Roger Levy, concerns the relationship between the form and meaning of words in the English lexicon. While the relationship between form and meaning is obvious at the level of morphemes (e.g., *glow* and *glowing* share both the form (morpheme) and meaning of *glow*), form-meaning systematicity is more controversial at the sub-morphemic level. Arbitrariness of the sign—the notion that the forms of words are unrelated to their meanings at the sub-morphemic level—is an underlying assumption of many linguistic theories. Two lines of research have recently challenged this assumption, but they produce differing characterizations of non-arbitrariness in language. Behavioral and corpus studies have confirmed the validity of localized form-meaning patterns manifested in limited subsets of the lexicon. Meanwhile, global (lexicon-wide) statistical analyses instead find diffuse form-meaning systematicity across the lexicon as a whole. The approach in this chapter bridges the gap with an approach that can detect both local and global form-meaning systematicity in language. In the kernel regression formulation we introduce, form-meaning relationships can be used to predict words' distributional semantic vectors from their forms. Furthermore, we introduce a novel metric learning algorithm that can learn weighted edit distances that minimize kernel regression error. Our results suggest that the English lexicon exhibits far more global form-meaning systematicity than previously discovered, and that much of this systematicity is focused in localized form-meaning patterns.

Chapter 4, which was co-written with Ekaterina Shutova, Gerard de Melo, and Patricia Lichtenstein, looks at how meaning interacts with opinions and the discourse level, and how this differs across languages. Understanding cross-cultural differences has important implications for world affairs and many aspects of the life of society. Yet, the majority of text-mining methods to date focus on the analysis of monolingual texts. In contrast, we present a statistical model that simultaneously learns a set of common topics from multilingual, non-parallel data and automatically discovers the differences in perspectives on these topics across linguistic communities. We perform a behavioural evaluation of a subset of the differences identified by our model in English and Spanish to investigate their psychological validity.

Chapter 2 formalizes the idea of conceptual metaphors as transformations (i.e., mappings) within the context of distributional semantic (i.e., vector-space) models. Metaphorical expressions are pervasive in natural language and pose a substantial challenge for computational semantics. The inherent compositionality of metaphor makes it an important test case for compositional distributional semantic models (CDSMs). The work presented here is the first investigation of whether metaphorical composition warrants a distinct treatment in the CDSM framework. We propose a method to learn metaphors as linear transformations in a vector space and find that, across a variety of semantic domains, explicitly modeling metaphor improves the resulting semantic representations. We then use these representations in a metaphor identification task, achieving high performance.

Chapter 2

Literal and Metaphorical Senses in Compositional Distributional Semantic Models

2.1 Introduction

An extensive body of behavioral and corpus-linguistic studies suggests that metaphors are pervasive in everyday language [Cam03, SDH$^+$10] and play an important role in how humans define and understand the world. According to Conceptual Metaphor Theory (CMT) [LJ81], individual metaphorical expressions, or *linguistic metaphors* (LMs), are instantiations of broader generalizations referred to as *conceptual metaphors* (CMs). For example, the phrases *half-baked idea*, *food for thought*, and *spoon-fed information* are LMs that instantiate the CM IDEAS ARE FOOD. These phrases reflect a mapping from the *source domain* of FOOD to the *target domain* of IDEAS [Lak89]. Two central claims of the CMT are that this mapping is systematic, in the sense that it consists of a fixed set of ontological correspondences, such as *thinking is preparing*, *communication is feeding*, *understanding is digestion*; and that this mapping can be productively extended to produce novel LMs that obey these correspondences.

Recent years have seen the rise of statistical techniques for metaphor detection. Several of these techniques leverage distributional statistics and vector-space models of meaning to classify utterances as literal or metaphorical [Uts06, SSK10, HSJ$^+$13, TBG$^+$14]. An important insight of these studies is that metaphorical meaning is not merely a property of individual words, but rather arises through cross-domain composition. The meaning of *sweet*, for instance, is not intrinsically metaphorical. Yet this word may exhibit a range of metaphorical meanings—e.g., *sweet dreams, sweet person, sweet victory*–that are created through the interplay of source and target domains. If metaphor is compositional, how do we represent it, and how can we use it in a compositional framework for meaning?

Compositional distributional semantic models (CDSMs) provide a compact model of compositionality that produces vector representations of phrases while avoiding the sparsity and storage issues associated with storing vectors for each phrase in a language explicitly. One of the most popular CDSM frameworks [BZ10, Gue10, CSC10] represents nouns as vectors, adjectives as matrices that act on the noun vectors, and transitive verbs as third-order tensors that act on noun or noun phrase vectors. The meaning of a phrase is then derived by composing these lexical representations. The vast majority of such models build a single representation for all senses of a word, collapsing distinct senses together. One exception is the work of Kartsaklis and Sadrzadeh [KS$^+$13], who investigated homonymy, in which lexical items have identical form but unrelated meanings (e.g., *bank*). They found that deriving verb tensors from all instances of a homonymous form (as compared to training a separate tensor for each distinct sense) loses information and degrades the resultant phrase vector representations. To the best of our knowledge, there has not yet been a study of regular polysemy (i.e. metaphorical or metonymic sense distinctions) in the context of compositional distributional semantics. Yet, due to systematicity in metaphorical cross-domain mappings, there are likely to be systematic contextual sense distinctions that can be captured by a CDSM, improving the resulting semantic representations.

In this paper, we investigate whether metaphor, as a case of regular polysemy, warrants distinct treatment under a compositional distributional semantic framework. We propose a new approach to CDSMs, in which metaphorical meanings are distinct but structurally related to literal meanings. We then extend the generalizability of our approach by proposing a method to automatically learn metaphorical mappings as linear transformations in a CDSM. We focus on modeling adjective senses and evaluate our methods on a new data set of 8592 adjective-noun pairs annotated for metaphoricity, which we will make publicly available. Finally, we apply our models to classify unseen adjective-noun (AN) phrases as literal or metaphorical and obtain state-of-the-art performance in the metaphor identification task.

2.2 Background & Related Work

Metaphors as Morphisms. The idea of metaphor as a systematic mapping has been formalized in the framework of category theory [Gog99, KF91]. In category theory, morphisms are transformations from one object to another that preserve some essential structure of the original object. Category theory provides a general formalism for analyzing relationships as morphisms in a wide range of systems (see Spivak [Spi14]). Category theory has been used to formalize the CM hypothesis with applications to user interfaces, poetry, and information visualization [KF91, GH10, GH05]. Although these formal treatments of metaphors as morphisms are rigorous and well-formalized, they have been applied at a relatively limited scale. This is because this work does not suggest a straightforward and data-driven way to quantify semantic domains or morphisms, but rather focuses on the transformations and relations between semantic domains and morphisms, assuming some appropriate quantification has already been established. In contrast, our methods can learn representations of source-target domain mappings from corpus data, and so are inherently more scalable.

Compositional DSMs. Similar issues arose in modeling compositional semantics. Formal semantics has dealt with compositional meaning for decades, by using mathematical structures from abstract algebra, logic, and category theory [Mon70, Par94, Lam99]. However, formal semantics requires manual crafting of features. The central insight of CDSMs is to model the composition of words as algebraic operations on their vector representations, as provided by a conventional DSM [ML08]. Guevara [Gue10] and Baroni and Zamparelli [BZ10] were the first to treat adjectives and verbs differently from nouns. In their models, adjectives are represented by matrices that act on noun vectors. Adjective matrices can be learned using regression techniques. Other CDSMs have also been proposed and successfully applied to tasks such as sentiment analysis and paraphrase [SPH$^+$11, SHMN12, TDSM13, Tur13].

Handling Polysemy in CDSMs. Several researchers argue that terms with ambiguous senses can be handled by DSMs without any recourse to additional disambiguation steps, as long as contextual information is available [BVCM12, EP10, PL02, Sch98, TDSM13]. Baroni et al. [BBZ14] conjecture that CDSMs might largely avoid problems handling adjectives with multiple senses because the matrices for adjectives implicitly incorporate contextual information. However, they do draw a distinction between two ways in which the meaning of a term can vary. Continuous *polysemy*—the subtle and continuous variations in meaning resulting from the different contexts in which a word appears—is relatively tractable, in their opinion. This contrasts with discrete *homonymy*—the association of a single term with completely independent meanings (e.g., *light house* vs. *light work*). Baroni et al. concede that homonymy is more difficult to handle in CDSMs. Unfortunately, they do not propose a definite way to determine whether any given variation in meaning is polysemy or homonymy, and offer no account of regular polysemy (i.e., metaphor and metonymy) or whether it would pose similar problems as homonymy for CDSMs.

To handle the problematic case of homonymy, Kartsaklis and Sadrzadeh [KSP13] adapt a clustering technique to disambiguate the senses of verbs, and then train separate tensors for each sense, using the previously mentioned CDSM framework of Coecke et al. [CSC10]. They found that prior disambiguation resulted in semantic similarity measures that correlated more closely with human judgments.

In principle, metaphor, as a type of regular polysemy, is different from the sort of semantic ambiguity described above. General ambiguity or vagueness in meaning (e.g. *bright light* vs *bright color*) is generally context-dependent in an unsystematic manner. In contrast, in regular polysemy meaning transfer happens in a systematic way (e.g. *bright light* vs. *bright idea*), which can be explicitly modeled within a CDSM. The above CDSMs provide no account of such systematic polysemy, which is the gap this paper aims to fill.

Computational Work on Metaphor. There is now an extensive literature on statistical approaches to metaphor detection. The investigated methods include clustering [BS06, SSK10, LS10]; topic modeling [BLM09, LRS10, HGS+13]; topical structure and imageability analysis [SBT+13]; semantic similarity graphs [SL09], and feature-based classifiers [GBNC06, LS09, TNAC11, Dun13a, Dun13b, HSJ+13, MBHT13, NAC+13, TMG13, TBG+14]. We refer readers to the survey by Shutova [Shu15] for a more thorough review.

Most relevant to the present work are approaches that attempt to identify whether adjective-noun phrases are metaphorical or literal. Krishnakumaran and Zhu [KZ07] use AN co-occurrence counts and WordNet hyponym/hypernym relations for this task. If the noun and its hyponyms/hypernyms do not occur frequently with the given adjective, then the AN phrase is labeled as metaphorical. Krishnakumaran and Zhu's system achieves a precision of 0.67. Turney et al. [TNAC11] classify verb and adjective phrases based on their level of concreteness or abstractness in relation to the noun they appear with. They learn concreteness rankings for words automatically (starting from a set of examples) and then search for expressions where a concrete adjective or verb is used with an abstract noun (e.g., *dark humor* is tagged as a metaphor; *dark hair* is not). They measure performance on a set of 100 phrases involving one of five adjectives, attaining an average accuracy of 0.79. Tsvetkov et al. [TBG+14] train a random-forest classifier using several features, including abstractness and imageability rankings, WordNet supersenses, and DSM vectors. They report an accuracy of 0.81 on the Turney et al. [TNAC11] AN phrase set. They also introduce a new set of 200 AN phrases, on which they measure an F-score of 0.85.

2.3 Experimental Data

Corpus. We trained our DSMs from a corpus of 4.58 billion tokens. Our corpus construction procedure is modeled on that of Baroni and Zamparelli [BZ10]. The corpus consisted of a 2011 dump of English Wikipedia, the UKWaC [BBFZ09], the BNC [BNC07], and the English Gigaword corpus [GKCM03]. The corpus was tokenized, lemmatized, and POS-tagged using the NLTK toolkit [BL04] for Python.

Metaphor Annotations. We created an annotated dataset of 8592 AN phrases (3991 literal, 4601 metaphorical). Our choice of adjectives was inspired by the test set of Tsvetkov et al. [TBG$^+$14], though our annotated dataset is considerably larger. We focused on 23 adjectives that can have both metaphorical and literal senses, and which function as source-domain words in relatively productive CMs: TEMPERATURE (*cold*, *heated*, *icy*, *warm*), LIGHT (*bright*, *brilliant*, *dim*), TEXTURE (*rough*, *smooth*, *soft*); SUBSTANCE (*dense*, *heavy*, *solid*), CLARITY (*clean*, *clear*, *murky*), TASTE (*bitter*, *sour*, *sweet*), STRENGTH (*strong*, *weak*), and DEPTH (*deep*, *shallow*). We extracted all AN phrases involving these adjectives that occur in our corpus at least 10 times. We filtered out all phrases that require wider context to establish their meaning or metaphoricity—e.g., *bright side*, *weak point*.

The remaining phrases were annotated using a procedure based on Shutova et al. [SSK10]. Annotators were encouraged to rely on their own intuition of metaphor, but were provided with the following guidance:

- For each phrase, establish the meaning of the adjective in the context of the phrase.
- Try to imagine a more basic meaning of this adjective in other contexts. Basic

meanings tend to be: more concrete; related to embodied actions, perceptions, or sensations; more precise; historically older/more “original”.

- If you can establish a basic meaning distinct from the meaning of the adjective in this context, it is likely to be used metaphorically.

If requested, a randomly sampled sentence from the corpus that contained the phrase in question was also provided. The annotation was performed by one of the authors. The author’s annotations were compared against those of a university graduate native English-speaking volunteer who was not involved in the research, on a sample of 500 phrases. Interannotator reliability [Coh60, FCE69] was $\kappa = 0.80$ ($SE = .02$). Our annotated data set is publicly available at http://bit.ly/1R5Yhn1.

2.4 Representing Metaphorical Senses in a Compositional DSM

In this section we test whether separate treatment of literal and metaphorical senses is justified in a CDSM framework. In that case, training adjective matrix representations on literal and metaphorical subsets separately may result in systematically improved phrase vector representations, despite each matrix making use of fewer training examples.

2.4.1 Method

Our goal is to learn accurate vector representations for unseen adjective-noun (AN) phrases, where adjectives can take on metaphorical or literal senses. Our models build off the CDSM framework of Baroni and Zamparelli [BZ10], as extended by Li et al. [LBD14]. Each adjective a is treated as a linear map from nouns to AN phrases:

$$\mathbf{p} = \mathbf{A}_a \mathbf{n},$$

where $\mathbf{p}$ is a vector for the phrase, $\mathbf{n}$ is a vector for the noun, and $\mathbf{A}_a$ is a matrix for the adjective.

Contextual Variation Model. The traditional representations do not account for the differences in meaning of an adjective in literal vs metaphorical phrases. Their assumption is that the contextual variations in meaning that are encoded by literal and metaphorical senses may be subtle enough that they can be handled by a single catch-all matrix per adjective, $\mathbf{A}_{\text{BOTH}(a)}$. In this model, every phrase i can be represented by

$$\mathbf{p}_i = \mathbf{A}_{\text{BOTH}(a)} \mathbf{n}_i \tag{2.1}$$

regardless of whether a is used metaphorically or literally in i. This model has the advantage of simplicity and requires no information about whether an adjective is being used literally or metaphorically. In fact, to our knowledge, all previous literature has handled metaphor in this way.

Discrete Polysemy Model Alternatively, the metaphorical and literal senses of an adjective may be distinct enough that averaging the two senses together in a single adjective matrix produces representations that are not well-suited for either metaphorical or literal phrases. Thus, the literal-metaphorical distinction could be problematic for CDSMs in the way that Baroni et al. [BBZ14] suggested that homonyms are. Just as Kartsaklis and Sadrzadeh [KS$^+$13] solve this problem by representing each sense of a homonym by a different adjective matrix, we represent literal and metaphorical senses by different adjective matrices. Each literal phrase i is represented by

$$\mathbf{p}_i = \mathbf{A}_{\mathrm{LIT}(a)}\mathbf{n}_i, \tag{2.2}$$

where $\mathbf{A}_{\mathrm{LIT}(a)}$ is the literal matrix for adjective a. Likewise, a metaphorical phrase is represented by

$$\mathbf{p}_i = \mathbf{A}_{\mathrm{MET}(a)}\mathbf{n}_i, \tag{2.3}$$

where $\mathbf{A}_{\mathrm{MET}(a)}$ is the metaphorical matrix for a.

Learning. Given a data set of noun and phrase vectors $\mathcal{D}(a) = \{(\mathbf{n}_i, \mathbf{p}_i)\}_{i=1}^{N}$ for AN phrases involving adjective a extracted using a conventional DSM, our goal is to learn $\mathbf{A}_{\mathcal{D}(a)}$. This can be treated as an optimization problem, of learning an estimate $\hat{\mathbf{A}}_{\mathcal{D}(a)}$ that minimizes a specified loss function. In the case of the squared error loss, $L(\mathbf{A}_{\mathcal{D}(a)}) = \sum_{i\in\mathcal{D}(a)} \|\mathbf{p}_i - \mathbf{A}_{\mathcal{D}(a)}\mathbf{n}_i\|_2^2$, the optimal solution can be found precisely using ordinary least-squares regression. However, this may result in overfitting because of the large number of parameters relative to the number of samples (i.e., phrases). Regularization parameters $\lambda = (\lambda_1, \lambda_2)$ can be introduced to keep $\hat{\mathbf{A}}_{\mathcal{D}(a)}$ small:

$$\sum_{i\in\mathcal{D}(a)} \|\mathbf{p}_i - \hat{\mathbf{A}}_{\mathcal{D}(a)}\mathbf{n}_i\|_2^2 + R(\lambda; \hat{\mathbf{A}}_{\mathcal{D}(a)}),$$

where $R(\lambda; \hat{\mathbf{A}}_{\mathcal{D}}) = \lambda_1\|\hat{\mathbf{A}}_{\mathcal{D}}\|_1 + \lambda_2\|\hat{\mathbf{A}}_{\mathcal{D}}\|_2$. This approach, known as elastic-net regression [ZH05], produces better adjective matrices than unregularized regression [LBD14]. Note that the same procedure can be used to learn the adjective representations in both the Contextual Variation model and the Discrete Polysemy model by varying what phrases are included in the training set $\mathcal{D}(a)$. In the Contextual Variation model $\mathcal{D}(a)$ includes both metaphorical and literal phrases, while in the Discrete Polysemy model it includes only metaphorical phrases when learning $\hat{\mathbf{A}}_{\textsc{met}(a)}$ and testing on metaphorical phrases (and only literal phrases when learning $\hat{\mathbf{A}}_{\textsc{lit}(a)}$ and testing on literal phrases).

2.4.2 Experimental Setup

Extracting Noun & Phrase Vectors. Our approach for constructing term vector representations is similar to that of Dinu et al. [DPB13]. We first selected the 10K most frequent nouns, adjectives, and verbs to serve as context terms. We then constructed a co-occurrence matrix that recorded term-context co-occurrence within a symmetric 5-word context window of the 50K most frequent POS-tagged terms in the corpus. We then used these co-occurrences to compute the positive pointwise mutual information (PPMI) between every pair of terms, and collected these into a term-term matrix. Next, we reduced the dimensionality of this matrix to 100 dimensions using singular-value decomposition. Additionally, we computed "ground truth" distributional vectors for all the annotated AN phrases in our data set by treating the phrases as single terms and computing their PPMI with the 50K single-word terms, and then projecting them onto the same 100-dimensional basis.

Training Adjective Matrices. For each adjective a that we are testing, we split the phrases involving that adjective into two subsets, the literal (LIT) subset and the metaphorical (MET) subset. We then split the subsets into 10 folds, so that we do not train and test any matrices on the same phrases. For each fold k, we train three adjective matrices: $\hat{\mathbf{A}}_{\textsc{met}(a)}$ using all phrases from the MET set not in fold k; $\hat{\mathbf{A}}_{\textsc{lit}(a)}$ using all phrases from the LIT set not in fold k; and $\hat{\mathbf{A}}_{\textsc{both}(a)}$ using all the phrases from either subset not in fold k. Within each fold, we use nested cross-validation as outlined in Li et al. [LBD14] to determine the regularization parameters for each regression problem.

2.4.3 Evaluating Vector Representations

Evaluation. Our goal is to produce a vector prediction of each phrase that will be close to its ground truth distributional vector. Phrase vectors directly extracted from the corpus by treating the phrase as a single term are the gold standard for predicting human judgment and producing paraphrases [DPB13], so we use these as our ground truth. The quality of the vector prediction for phrase i is measured using the cosine distance between the phrase's ground truth vector $\mathbf{p}_i$ and the vector prediction $\hat{\mathbf{p}}_i$:

$$err(\hat{\mathbf{p}}_i) = 1 - \cos(\hat{\mathbf{p}}_i, \mathbf{p}_i).$$

We then analyze the benefit of training on a reduced subset by calculating a "subset improvement" (SI) score for the MET and LIT subsets of each adjective a. We define the SI for each subset $\mathcal{D}(a) \in \{\text{LIT}(a), \textsc{met}(a)\}$ as:

$$SI(\mathcal{D}(a)) = 1 - \frac{\sum_{i \in \mathcal{D}(a)} err(\hat{\mathbf{A}}_{\mathcal{D}(a)}\mathbf{n}_i)}{\sum_{i \in \mathcal{D}(a)} err(\hat{\mathbf{A}}_{\textsc{both}(a)}\mathbf{n}_i)}$$

Positive values of SI thus indicate improved performance when trained on a reduced subset compared to the full set of phrases. For example $SI_{\textsc{lit}(a)} = 5\%$ tells us that predicting the phrase vectors for LIT phrases of adjective a using the LIT matrix

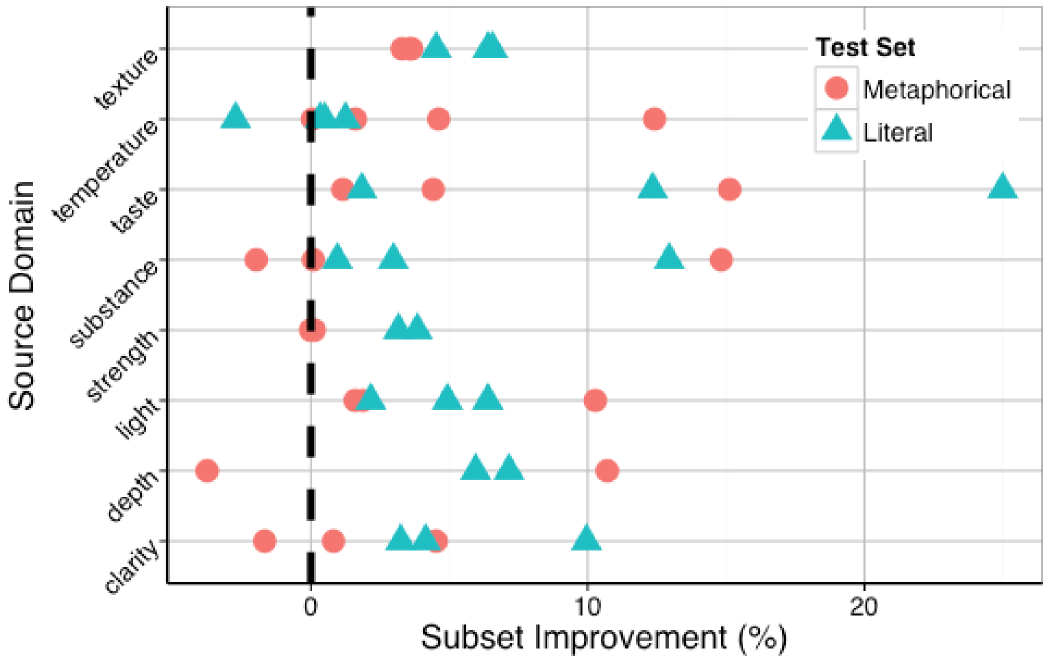

Figure 2.1: Reduction in error from training on targeted subset (MET/LIT) rather than on all phrases.

resulted in a 5% reduction in mean cosine error compared to predicting the phrase vectors using the BOTH matrix.

Results. The results are summarized in Fig. 2.1. Each point indicates the SI for a single adjective and for a single subset. Adjectives are grouped by source domain along the y-axis. Overall, almost every item shows a subset improvement; and, for every source domain, the majority of adjectives show a subset improvement.

We analyzed per-adjective SI by fitting a linear mixed-effects model, with a fixed intercept, a fixed effect of test subset (MET vs. LIT), a random effect of source domain, and the maximal converging random effects structure (uncorrelated random intercepts and slopes) [BLST13]. Training on a targeted subset improved performance by $4.4\% \pm 0.009(SE)$ ($p = .002$). There was no evidence that this differed by test subset (i.e., metaphorical vs. literal senses, $p = .35$). The positive SI from training on a targeted subset suggests that metaphorical and literal uses of the same adjective are semantically distinct.

2.4.4 Metaphor Classification

Method. The results of the previous section suggest a straightforward classification rule: classify unseen phrase i involving adjective a as metaphorical if

$$\cos(\mathbf{p}_i, \hat{\mathbf{A}}_{\text{MET}(a)}\mathbf{n}_i) < \cos(\hat{\mathbf{A}}_{\text{LIT}(a)}\mathbf{n}_i).$$

Otherwise, we classify it as literal.

Evaluation. We test this method on our data set of 8593 annotated AN phrases using 10-fold cross validation. It is possible that our method's classification performance is not due to the compositional aspect of the model, but rather to some semantic coherence property among the nouns in the AN phrases that we are testing. To control for this possibility, we compare the performance of our method against four baselines. The first baseline, NOUN-NN, measures the cosine distance between the vector for the noun of the AN phrase being tested and the noun vectors of the nouns participating in an AN phrase in the training folds. The test phrase is then assigned the label of the AN phrase whose noun vector is nearest. PHRASE-NN proceeds similarly, but using the ground-truth phrase vectors for the test phrase and the training phrases. The test phrase is then assigned the label of the AN phrase whose vector is nearest. The baseline NOUN-CENT first computes the centroid of the noun vectors of the training phrases that are literal, and the centroid of the noun vectors of the training phrases that are metaphorical. It then assigns the test phrase the label of the centroid whose cosine distance from the test phrase's noun vector is smallest. PHRASE-CENT, proceeds similarly, but using phrase vectors. We measure performance against the manual annotations.

Results. Our classification method achieved a held-out F-score of 0.817, recall of 0.793, precision of 0.842, and accuracy of 0.809. These results were superior to those of the baselines (Table 2.1). These results are competitive with the state of the art and demonstrate the importance of compositionality in metaphor identification.

2.5 Metaphors as Linear Transformations

One of the principal claims of the CM hypothesis is that CMs are productive: A CM (i.e., mapping) can generate endless new LMs (i.e., linguistic expressions). Cases where the LMs involve an adjective that has already been used metaphorically and for which we have annotated metaphorical and literal examples can be handled by the methods of §2.4, but when the novel LM involves an adjective that has only

Table 2.1: Performance of the method of §2.4.4 (MET-LIT) against various baselines.

Method	F-score	Precision	Recall	Accuracy
MET-LIT	0.817	0.842	0.793	0.809
NOUN-NN	0.709	0.748	0.675	0.703
PHRASE-NN	0.590	0.640	0.547	0.592
NOUN-CENT	0.717	0.741	0.695	0.706
PHRASE-CENT	0.629	0.574	0.695	0.559

been observed in literal usage, we need a more elaborate model. According to the CM hypothesis, an adjective's metaphorical meaning is a result of the action of a source-to-target CM mapping on the adjective's literal sense. If so, then given an appropriate representation of this mapping it should be possible to infer the metaphorical sense of an adjective without ever seeing metaphorical exemplars—that is, using only the adjective's literal sense. Our next experiments seek to determine whether it is possible to represent and learn CM mappings as linear maps in distributional vector space.

2.5.1 Model

We model each CM mapping $\mathcal{M}$ from source to target domain as a linear transformation $\mathbf{C}_{\mathcal{M}}$:

$$\mathbf{A}_{\text{MET}(a)}\mathbf{n}_i \approx \mathbf{C}_{\mathcal{M}}\mathbf{A}_{\text{LIT}(a)}\mathbf{n}_i \tag{2.4}$$

We can apply a two-step regression to learn $\mathbf{C}_{\mathcal{M}}$. First we apply elastic-net regression to learn the literal adjective matrix $\hat{\mathbf{A}}_{\text{LIT}(a)}$ as in §2.4.2. Then we can substitute this estimate into Eq. (2.4), and apply elastic-net regression to learn the $\hat{\mathbf{C}}_{\mathcal{M}}$ that minimizes the regularized squared error loss:

$$\sum_{a \in \mathcal{M}} \sum_{i \in \mathcal{D}(a)} \|\mathbf{p}_i - \hat{\mathbf{C}}_{\mathcal{M}}\hat{\mathbf{A}}_{\text{LIT}(a_i)}\mathbf{n}_i\|_2^2 + R(\lambda; \hat{\mathbf{C}}_{\mathcal{M}}).$$

To learn $C_{\mathcal{M}}$ in this regression problem, we can pool together and train on phrases from many different adjectives that participate in $\mathcal{M}$.

2.5.2 Experimental Setup

We used a cross-validation scheme where we treated each adjective in a source domain as a fold in training the domain's metaphor transformation matrix. The nested cross-validation procedure we use to set regularization parameters λ and evaluate performance requires at least 3 adjectives in a source domain, so we evaluate on the 6 source domain classes containing at least 3 adjectives. The total number of phrases for these 19 adjectives is 6987 (3659 metaphorical, 3328 literal).

2.5.3 Evaluating Vector Representations

Evaluation. We wish to test whether CM mappings learned from one set of adjectives are transferable to new adjectives for which metaphorical phrases are unseen. As in §2.4, models were evaluated using cosine error compared to the ground truth phrase vector representation. Since our goal is to improve the vector representation of metaphorical phrases given no metaphorical annotations, we measure performance on the MET phrase subset for each adjective. We compare the performance of the transformed LIT matrix $\mathbf{C}_{\mathcal{M}}\mathbf{A}_{\text{LIT}(a)}$ against the performance of the original LIT matrix $\mathbf{A}_{\text{LIT}(a)}$ by defining the metaphor transformation improvement (MTI) as:

$$MTI(a) = 1 - \frac{\sum_{i \in \text{MET}} err(\mathbf{C}_{\mathcal{M}}\hat{\mathbf{A}}_{\text{LIT}(a)})}{\sum_{i \in \text{MET}} err(\hat{\mathbf{A}}_{\text{LIT}(a)})}.$$

Results. Per-adjective MTI was analyzed with a linear mixed-effects model, with a fixed intercept, a random effect of source domain, and random intercepts. Transforming the LIT matrix using the CM mapping matrix improved performance by 11.5% ± $0.023(SE)$ ($p < .001$). On average, performance improved for 18 of 19 adjectives and

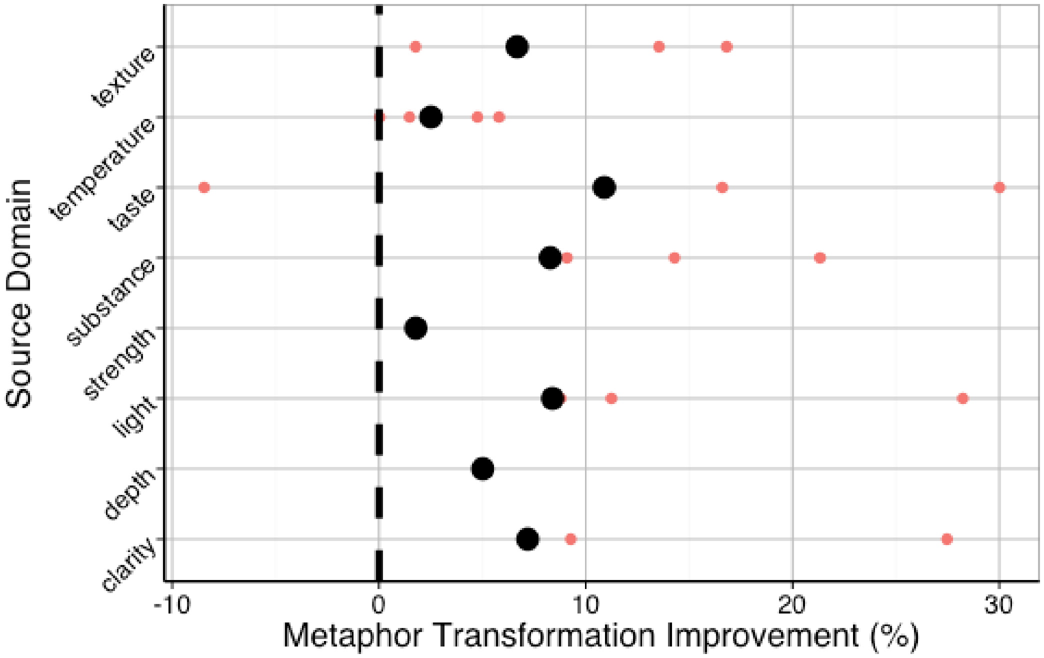

Figure 2.2: Reduction in error from transforming LIT matrix using metaphorical mapping. Mean change was positive for every domain (large black), and for all but one adjective (small red).

for every source domain ($p = .03$, binomial test; Fig. 2.2). Thus, mapping structure is indeed shared across adjectives participating in the same CM.

2.5.4 Metaphor Classification

Method. Once again our results suggest a procedure for metaphor classification. This procedure can classify phrases involving adjectives without seeing any metaphorical annotations. For any unseen phrase i involving an adjective a_i, we classify the phrase as metaphorical if

$$\cos(\mathbf{p}_i, \hat{\mathbf{C}}_{\mathcal{M}}\hat{\mathbf{A}}_{\text{LIT}(a_i)}\mathbf{n}_i) < \cos(\mathbf{p}_i, \hat{\mathbf{A}}_{\text{LIT}(a_i)}\mathbf{n}_i)$$

Otherwise, we classify it as literal. We used the same procedure as in §2.4.2 to learn $\hat{\mathbf{A}}_{\textsc{lit}(a_i)}$.

Results. Our method achieved an F-score of 0.793 on the classification of phrases involving unseen adjectives. On this same set of phrases, the method of §2.4.4 achieved an F-score of 0.838. Once again, the performance of our method was superior to the performance of the baselines (Table 2.2; the MET-LIT figures in Table 2.2 differ slightly from those in Table 2.1 because only 19 of 23 adjectives are tested). For comparison, we also include the classification performance using the MET-LIT method of §2.4.4. While MET-LIT slightly outperforms TRANS-LIT, the latter has the benefit of not needing annotations for metaphorical phrases for the test adjective. Hence, our approach is generalizable to cases where such annotations are unavailable with only slight performance reduction.

2.6 Discussion

Overall, our results show that taking metaphor into account has the potential to improve CDSMs and expand their domain of applicability. The findings of §2.4 suggest that collapsing across metaphorical and literal uses may hurt accuracy of vector representations in CDSMs. While the method in §2.4 depends on explicit annotations of metaphorical and literal senses, the method in §2.5 provides a way to generalize these representations to adjectives for which metaphorical training data is unavailable, by showing that metaphorical mappings are transferable across adjectives from the same source domain. Note that an accurate matrix representation of the literal sense of each adjective is still required in the experimental setup of §2.5. This particular choice of setup allowed a proof of concept of the hypothesis that metaphors function as cross-domain transformations, but in principle it would be desirable to learn transformations from a general BOTH matrix representation for any adjective in a source domain to its MET matrix representation. This would enable improved vector representations of metaphorical AN phrases without annotation for unseen adjectives.

The success of our models on the metaphor classification tasks demonstrates that there is information about metaphoricity of a phrase inherent in the composition of the meanings of its components. Notably, our results show that this metaphorical compositionality can be captured from corpus-derived distributional statistics. We also noticed some trends at the level of individual phrases. In particular, classification performance and vector accuracy tended to be lower for metaphorical phrases whose nouns are distributionally similar to nouns that tend to participate in literal phrases (e.g., *reception* is similar to *foyer* and *refreshment* in our corpus; *warm reception* is metaphorical while *warm foyer* is literal). Another area where classification accuracy is low is in phrases with low corpus occurrence frequency. The ground truth vectors for these phrases exhibit high sample variance and sparsity. Many such phrases sound paradoxical (e.g., *bitter sweetness*).

Our results could also inform debates within cognitive science. First, cognitive scientists debate whether words that are used both literally and figuratively (e.g., *long road*, *long meeting*) are best understood as having a single, abstract meaning that varies with context or two distinct but related meanings. For instance, some argue that domains like space, time, and number operate over a shared, generalized magnitude system, yet others maintain that our mental representation of time and number is distinct from our mental representation of space, yet inherited metaphorically from it

[WMM15]. Our results suggest that figurative and literal senses involve quite different patterns of use. This is statistical evidence that adjectives that are used metaphorically have distinct related senses, not a single abstract sense.

Second, the Conceptual Metaphor Theory account hypothesizes that LMs are an outgrowth of metaphorical thought, which is in turn an outgrowth of embodied experiences that conflate source and target domains—experience structures thought, and thought structures language [Lak93]. However, recent critics have argued for the opposite causal direction: Linguistic regularities may drive the mental mapping between source and target domains [HL13, Cas14, HL14]. Our results show that, at least for AN pairs, the semantic structure of a source domain and its mapping to a metaphorical target domain are available in the distributional statistics of language itself. There may be no need, therefore, to invoke embodied experience to explain the prevalence of metaphorical thought in adult language users. A lifetime of experience with literal and metaphorical language may suffice.

2.7 Conclusion

We have shown that modeling metaphor explicitly within a CDSM can improve the resulting vector representations. According to our results, the systematicity of metaphor can be exploited to learn linear transformations that represent the action of metaphorical mappings across many different adjectives in the same semantic domain. Our classification results suggest that the compositional distributional semantics of a phrase can inform classification of the phrase for metaphoricity.

Beyond improvements to the applications we presented, the principles underlying our methods also show potential for other tasks. For instance, the LIT and MET adjective matrices and the CM mapping matrix learned with our methods could be applied to improve automated paraphrasing of AN phrases. Our work is also directly extendable to other syntactic constructions. In the CDSM framework we apply, verbs would be represented as third-order tensors. Tractable and efficient methods for estimating these verb tensors are now available [FPC15]. It may also be possible to extend the coverage of our system by using automated word-sense disambiguation to bootstrap annotations and therefore construct LIT and MET matrices in a minimally supervised fashion [KSP13]. Finally, it would be interesting to investigate modeling metaphorical mappings as nonlinear mappings within the deep learning framework.

Chapter 2, in part, is a reprint of the material as it appears in Gutierrez, E.D.; Shutova, Ekaterina; Marghetis, Tyler; Bergen, Benjamin. "Literal and Metaphorical Senses in Compositional Distributional Semantic Models", Proceedings of the Association for Computational Linguistics, 2016. The dissertation author was the primary investigator and author of this paper.

Table 2.2: Performance of method of §2.5.4 (TRANS-LIT) against method of §2.4.4 (MET-LIT) and various baselines.

Method	F-score	Precision	Recall	Accuracy
TRANS-LIT	0.793	0.716	0.819	0.804
MET-LIT	0.838	0.856	0820	0.833
NOUN-NN	0.692	0.732	0.655	0.693
PHRASE-NN	0.575	0.625	0.532	0.587
NOUN-CENT	0.703	0.722	0.685	0.696
PHRASE-CENT	0.610	0.552	0.681	0.542

Chapter 3

Finding Non-Arbitrary Form-Meaning Systematicity Using String-Metric Learning for Kernel Regression

3.1 Introduction

Arbitrariness of the sign refers to the notion that the phonetic/orthographic forms of words have no relationship to their meanings [dS16]. It is a foundational assumption of many theories of language comprehension, production, acquisition, and evolution. For instance, Hockett's [Hoc60] influential enumeration of the design features of human language gives arbitrariness a central role in enabling the combination and recombination of phonemic units to create new words. Gasser [Gas04] uses simulations to show that for large vocabularies, arbitrary form-meaning mappings may provide an advantage in acquisition. Meanwhile, modular theories of language comprehension rely upon the duality of patterning to support the independence of the phonetic and semantic aspects of language comprehension [LRM99]. Quantifying the extent to which the arbitrariness principle actually holds is important for understanding how language works.

Language researchers have long noted exceptions to arbitrariness. Most of these are patterns that occur in some relatively localized subset of the lexicon. These patterns are sub-morphemic because, unlike conventional morphemes, they cannot combine reliably to produce new words. *Phonaesthemes* [Fir30] are one example. A phonaestheme is a phonetic cluster that recurs in many words that have related meanings. One notable phonaestheme is the onset *gl-*, which occurs at the beginning of at least 38 English words relating to vision: *glow*, *glint*, *glaze*, *gleam*, etc. [Ber04]. At least 46 candidate phonaesthemes have been posited in the linguistics literature, according to a list compiled by Hutchins [Hut98]. *Iconicity* is another violation of arbitrariness that can lead to non-arbitrary local regularities. Iconicity occurs when the form of a word is transparently motivated by some perceptual aspect of its referent. When several referents share perceptual features, their associated word tokens may tend to be similar as well. For instance, Ohala [Oha84] conjectures that vowels with high acoustic frequency tend to associate with smaller items while vowels with low acoustic frequency tend to associate with larger items, due to the experiential association between vocalizer size and frequency. Iconicity is also manifested in sets of onomatopoeic words that echo similar sounds (e.g., *clink*, *clank*). Although these exceptions to non-arbitrariness differ, in each case, specific form-meaning relationships emerge in a subset of the lexicon. We will refer to all such specific localized form-meaning patterns as *phonosemantic sets*.

In recent decades, behavioral and corpus studies have empirically confirmed the psychological reality and statistical reliability of many phonosemantic sets that had previously been identified by intuition and observation. Various candidate phonaesthemes have significant effects on reaction times during language processing tasks [Hut98, Mag98, Ber04]. Sagi and Otis [SO08] test the statistical significance of the 46 candidates in Hutchins's [Hut98] list, and find that 27 of the 46 exhibit more within-category distributional semantic coherence than would be expected by chance. These results have been replicated using other corpora and distributional semantic models [AFS13]. Klink [Kli00] shows that sound-symbolic attributes such as those proposed by Ohala [Oha84] are associated with human judgments about nonwords' semantic attributes, such as smallness or beauty. Using a statistical corpus analysis and WordNet semantic features, Monaghan et al. [MLC14] look at a similar hypothesis space of sound-symbolic phonological and semantic attributes, and reach similar conclusions.

While these localized studies support the existence of some islands of non-arbitrariness in language, their results do not address how pervasive non-arbitrariness is at the global level—that is, in the lexicon of a language as a whole. After all, some seemingly non-arbitrary local patterns can be expected to emerge merely by chance. How can we measure whether local phonosemantic patterning translates into global *phonosemantic systematicity*–that is, strong, non-negligible lexicon-wide non-arbitrariness? Shillcock et al. [SKMB01] introduce the idea of measuring phonosemantic systematicity by analyzing the correlation between phonological edit distances and distributional semantic distances. In a lexicon of monomorphemic and monosyllabic English words, they find a small but statistically significant correlation between these two distance measures. Monaghan et al. [MSCK14] elaborate on this methodology, showing that the statistical effect is robust to different choices of form-distance and semantic-distance metrics. They also look at the effect of leaving out each word in the lexicon on the overall correlation measure; from this, they derive a phonosemantic systematicity measure for each word. Interestingly, they find that systematicity is diffusely distributed across the words in English in a pattern indistinguishable from random chance. Hence, they conclude that "systematicity in the vocabulary is not a consequence of small clusters of sound symbolism." This line of work provides a proof-of-concept that it is possible to detect the phonosemantic systematicity of a language, and confirms that English exhibits significant phonosemantic systematicity.

Broadly speaking, both the localized tests of individual phonosemantic sets and the global analyses of phonosemantic systematicity challenge the arbitrariness of the sign. However, they attribute responsibility for non-arbitrariness differently. The local methods reveal dozens of specific phonosemantic sets that have strong, measurable behavioral effects and statistical signatures in corpora. Meanwhile, the global methods find small and diffuse systematicity. How can we reconcile this discrepancy?

Original Contributions. We attempt to bridge the gap with a new approach that builds off of previous lexicon-wide analyses, making two innovations. The first addresses the concern that the lexicon-wide methods currently in use may not be well suited to finding local regularities such as phonosemantic sets, because they make the assumption that systematicity exists only in the form of a global correlation between distances in form-space and distances in meaning-space. Instead, we model the problem using kernel regression, a nonparametric regression model. Crucially, in kernel regression the prediction for a point is based on the predictions of neighboring points; this enables us to conduct a global analysis while still capturing local, neighborhood effects. As in previous work, we represent word-forms by their orthographic strings, and word-meanings by their semantic vector representations as produced by a distributional semantic vector space model. The goal of the regression is then to learn a mapping from string-valued predictor variables to vector-valued target variables that minimizes regression error in the vector space. Conveniently, our model allows us to produce predictions of the semantic vectors associated with both words and nonwords.

Previous work may also underestimate systematicity in that it weights all edits (substitutions, insertions, and deletions) equally in determining edit distance. A priori, there is no reason to believe this is the case—indeed, the work on individual phonosemantic sets suggests that some orthographic/phonetic attributes are more important than others for non-arbitrariness. To address this, we introduce String-Metric Learning for Kernel Regression (SMLKR), a metric-learning algorithm that is able to learn a weighted edit distance metric that minimizes the prediction error in kernel regression.

We find that SMLKR enables us to recover more systematicity from a lexicon of monomorphemic English words than reported in previous global analyses. Using SMLKR, we propose a new measure of per-word phonosemantic systematicity. Our analyses using this systematicity measure indicate that specific phonosemantic sets do contribute significantly to the global phonosemantic systematicity of English, in keeping with previous local-level analyses. Finally, we evaluate our systematicity measure against human judgments, and find that it accords with raters' intuitions about what makes a word's form well suited to its meaning.

3.2 Background & Related Work

3.2.1 Previous Approaches to Finding Lexicon-Wide Systematicity

Measuring Form, Meaning, and Systematicity. To our knowledge, all previous lexicon-level analyses of phonosemantic systematicity have used variations of the method of Shillcock et al. [SKMB01]. The inputs for this method are form-meaning tuples $(\mathbf{y}_i, s_i)$ for each word i in the lexicon, where $\mathbf{y}_i$ is the vector representation of the word in a distributional semantic model, and s_i is the string representation of the word (phonological, phonemic, or orthographic). Semantic distances are measured as cosine distances between the vectors of each pair of words. Shillcock et al. [SKMB01] and Monaghan et al. [MSCK14] measure form-distances in terms of edit distance between each pair of strings. In addition Monaghan et al. [MSCK14] and Tamariz [Tam06] study distance measures based on a selected set binary phonological features, with similar results. Phonosemantic systematicity is then measured as the correlation between all the pairwise semantic distances and all the pairwise string distances.

Hypothesis Testing. In this line line of work, statistical significance of the results is assessed using the Mantel test, a permutation test of the correlation between two sets of pairwise distances [Man67]. The test involves randomly shuffling the assignments of semantic vectors to word-strings in the lexicon. We can think of each form-meaning shuffle as a member of the set of all possible lexicons. Next, the correlation between the semantic distances and the string distances is computed under each reassignment. An empirical p-value for the true lexicon is then derived by performing many shufflings, and comparing the correlation coefficients measured under the shuffles to the correlation coefficient measured in the true lexicon. Under the null hypothesis that form-meaning assignments are arbitrary, the probability of observing a form-meaning correlation of at least the magnitude actually observed in the true lexicon is asymptotically equal to the proportion of reassignments that produce greater correlations than the true lexicon.

Previous Findings. Shillcock et al. [SKMB01] find a statistically significant correlation between semantic and phonological edit distances in a lexicon of the 1733 most frequent monosyllabic monomorphemic words in the BNC. Tamariz [Tam08] extends these results to Spanish data, looking only at words with one of three consonant-vowel (CV) structures (CVCV, CVCCV, and CVCVCV). [SKMB01], Monaghan et al. [MSCK14] derive a list of 5138 monomorphemic monosyllabic words and a list of 5604 monomorphemic polysyllabic from the CELEX database [BPG96], and find significant form-meaning correlations in both.

3.2.2 Kernel Regression

In contrast to previous studies, we study form-meaning systematicity using a kernel regression framework. Kernel regression is a nonparametric supervised learning technique that is able to learn highly nonlinear relationships between predictor variables and target variables. Rather than assuming any particular parametric relationship between the predictor and target variables, kernel regression assumes only that the value of the target variable is a smooth function of the value of the predictors. In other words, given a new point in predictor space, the value of the target at that point can reasonably be estimated by the value of the targets at points that are nearby in the predictor space. In this way, kernel regression is analogous to an exemplar model. We performed kernel regression on our lexicon using the Nadaraya-Watson estimator [Nad64]. Given a data set $\mathcal{D}$ of vector-valued predictor variables $\{\mathbf{x}_i\}_{i=1}^N$, and targets $\{\mathbf{y}_i\}_{i=1}^N$, the Nadaraya-Watson estimator of the target for sample i is

$$\hat{\mathbf{y}}_i = \hat{\mathbf{y}}(\mathbf{x}_i) = \frac{\sum_{j \neq i} k_{ij} \mathbf{y}_j}{\sum_{j \neq i} k_{ij}}, \tag{3.1}$$

where k_{ij} is the *kernel* between point i and point j. A commonly used kernel is the exponential kernel:

$$k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-d(\mathbf{x}_i, \mathbf{x}_j)/h),$$

where $d(\cdot, \cdot)$ is a distance metric and h is a *bandwidth* that determines the radius of the effective neighborhood around each point that contributes to its estimate. For our purposes we use the Levenshtein string edit distance metric [Lev66]. The Levenshtein edit distance between two strings is the minimum number of edits needed to transform one string into the other, where an edit is defined as the insertion, deletion, or substitution of a single character. Using this edit distance and semantic vectors derived from a distributional semantic model, the Nadaraya-Watson estimator can estimate the position in the semantic vector space for each word in the lexicon. The exponential edit distance kernel has been useful for modeling behavior in many tasks involving word similarity and neighborhood effects; see, for example the Generalized Context Model [Nos86], which has been applied to word identification, recognition, and categorization, to inflectional morphology, and to artificial grammar learning [BH01].

3.2.3 Metric Learning for Kernel Regression

In kernel regression, the bandwidth h of the kernel function must be fine-tuned by testing out many different bandwidths. Moreover, for many tasks there is no reason to assume that all of the dimensions of a vector-valued predictor are equally important. This is problematic for conventional kernel regression, as the quality of its predictions is wholly reliant on the appropriateness of the given distance metric.

Weinberger and Tesauro [WT07] introduce metric learning for kernel regression (MLKR), an algorithm that can learn a task-specific Mahalanobis (i.e., weighted Euclidean) distance metric over a real-vector-valued predictor space, in which small distances between two vectors imply similar target values. They note that this metric induces a kernel function whose parameters are set entirely from the data. Specifically, MLKR can learn a weight matrix W for a Mahalanobis metric that optimizes the leave-one out mean squared error of kernel regression (MSE), defined as:

$$\mathcal{L}(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\hat{\mathbf{y}}_i, \mathbf{y}_i) = \frac{1}{N}\sum_{i=1}^{N}\|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2,$$

where $\hat{\mathbf{y}}_i$ is estimated using $\hat{\mathbf{y}}_j$ for all $i \neq j$, as in Eq. 3.1.

In MLKR, the weighted distance metric is learned using stochastic gradient descent. As an added benefit, MLKR is implicitly able to learn an appropriate kernel bandwidth.

3.3 String-Metric Learning for Kernel Regression (SMLKR)

Our novel contribution is an extension of MLKR to situations where the predictor variables are not real-valued vectors, but strings, and the distance metric we wish to learn is a weighted Levenshtein edit distance. Vector-valued representations of the strings themselves would only approximately preserve edit distance. Fortunately, it turns out that we do not need vector-valued representations of the strings at all. Define the *minimum edit-distance path* as the smallest-length sequence of edits that is needed to transform one string into another. Observe that the weighted edit distance between two strings s_i and s_j can be represented as the weighted sum of all the edits that must take place to transform one string into the other along the minimum edit-distance path [BHS12]. In turn, these edits can be represented by a vector $\boldsymbol{\nu}_{ij}$ constructed as in Fig 3.1, while the weights can be represented by a vector $\mathbf{w} = (w_1, ..., w_M)^T$:

$$d_{WL}(s_i, s_j) = \sum_{m=1}^{M} w_m \nu_{ijm} = \mathbf{w}^T \boldsymbol{\nu}_{ij}.$$

Each entry of $\boldsymbol{\nu}_{ij}$ corresponds to a particular type of edit operation (e.g., substitution of character a for character b). The value assigned to each entry is the count of the total number of times that the corresponding edit operation must be applied to achieve transformation of string i to string j along the minimum edit-distance path.

We note that $\boldsymbol{\nu}_{ij}$ does not admit a unique representation, since there are multiple ways to transform one string to another in the same number of edits, using different edit operations. However, we adopt the convention that some class of edit operations always takes priority over another—e.g., that deletions always occur before

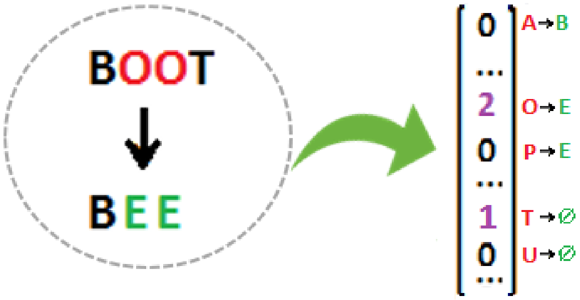

Figure 3.1: Each element in $\boldsymbol{\nu}_{ij}$ (the vector at left) represents a type of edit. The entry ν_{ijm} represents the number of edits of type m that occur as string s_i (*boot*) is transformed into string s_j (*bee*).

substitutions. This then enables us to specify $\boldsymbol{\nu}_{ij}$ uniquely. We also adopt the convention that the weights for edit operations are symmetric—e.g., that the weight for substituting character a for character b is the same as the weight for substituting character b for character a, so we represent every such pair of edit operations by a single entry in $\boldsymbol{\nu}_{ij}$.

As in MLKR, our goal is to minimize the leave-one-out MSE,[1] where $k_{ij} = e^{-\mathbf{w}^T \boldsymbol{\nu}_{ij}}$. The gradient of the regression error for MSE is

$$\frac{\partial \mathcal{L}(\mathcal{D})}{\partial \mathbf{w}} = \frac{2}{N} \sum_{i=1}^{N} (\mathbf{y}_i - \hat{\mathbf{y}}_i) \frac{\partial \hat{\mathbf{y}}_i}{\partial \mathbf{w}}$$

where

$$\frac{\partial \hat{\mathbf{y}}_i}{\partial \mathbf{w}} = \frac{\sum_{j \neq i} (\mathbf{y}_j - \hat{\mathbf{y}}_i)^T k_{ij} \boldsymbol{\nu}_{ij}}{\sum_{j \neq i} k_{ij}}.$$

Using this exact gradient, we can find the edit weights that minimize the loss function. We wish to constrain the weights to be nonnegative, since weighted edit distance only

[1] We attained similar results minimizing mean cosine error. The gradient for mean cosine error is

$$\frac{\partial \mathcal{L}(\mathcal{D})}{\partial \mathbf{w}} = \frac{1}{N} \sum_{i=1}^{N} \frac{(\|\hat{\mathbf{y}}_i\| \mathbf{y}_i - \mathcal{L}(\mathbf{y}_i, \hat{\mathbf{y}}_i) \hat{\mathbf{y}}_i)}{\|\hat{\mathbf{y}}_i\|^2} \frac{\partial \hat{\mathbf{y}}_i}{\partial \mathbf{w}}.$$

makes sense with nonnegative weights. Thus, to minimize the loss we use the limited-memory BroydenFletcherGoldfarbShanno algorithm for box constraints (L-BFGS-B) [BLNZ95], a quasi-Newton method that allows bounded optimization. We made a Python implementation of SMLKR available at `http://bit.ly/22t1Jgx`.

3.4 Experimental Setup

3.4.1 Data

Lexicon. A principal concern is the possibility that our models may detect morphemes rather than sub-morphemic units. To minimize this concern, we adopted an approach similar to that of Shillcock et al. [SKMB01], of training our model only on monomorphemic words. Monomorphemic words were selected by cross-referencing the morphemic analyses contained in the CELEX lexical database [BPG96] with the morphemic analyses contained in the etymologies of the Oxford English Dictionary Online (`http://www.oed.com`). Then, we went through the filtered list and removed any remaining polymorphemic words as well as place names, demonyms, spelling variants, and proper nouns. Finally, words that were not among the 40,000 most frequent non-filler word types in the corpus were excluded. The final lexicon was composed of 4,949 word types.

Corpus and Semantic Model. The corpus we used to train our semantic model is a concatenation of the UKWaC, BNC, and Wikipedia corpora [FZBB08, BNC07, PGK+11]. We trained our vector-space model on this corpus using the Word2Vec [MSC+13], as instantiated in the Gensim package [ŘS10] for Python using default parameters. We produced 100-dimensional word-embedding vectors using the Skip-

Gram algorithm of Word2Vec and normalized the 100-dimensional vector for each word so that its Euclidean norm was equal to 1.

3.4.2 Training

We trained SMLKR on the 100-dimensional Word2Vec embeddings using L-BGFS-B, and placing non-negativity constraints on the weights $\mathbf{w}$. We let SMLKR run until convergence, as determined by the following criterion:

$$\frac{|\mathcal{L}^{(k-1)} - \mathcal{L}^{(k)}|}{\max(|\mathcal{L}^{(k-1)}|, |\mathcal{L}^{(k)}|)} = \epsilon$$

where $\mathcal{L}^{(k)}$ is the loss at the k^{th} iteration of learning, and we set $\epsilon = 2 \times 10^{-8}$. We randomly initialized the L-BGFS-B algorithm 10 times to avoid poor local minima, and kept the solution with the lowest loss.

3.4.3 Other Approaches

We tried several other, less successful approaches to finding systematicity in the lexicon. The results of these experiments are presented in appendix A.

3.5 Experiments

3.5.1 Model Analysis

Weighted Edit Distance Reveals More Non-Arbitrariness. We first assessed whether the structure found by kernel regression could arise merely by arbitrary, random pairings of form and meaning (i.e., strings and semantic vectors). We adopt a

Monte Carlo testing procedure similar to the Mantel test of §3.2.1. We first randomly shuffled the assignment of the semantic vectors of all the words in the lexicon. We then trained SMLKR on the shuffled lexicon just as we did on the true lexicon. We measured the mean squared error of the SMLKR prediction. Out of 1000 reassignments, none produced a prediction error as small as the prediction error in the true lexicon (i.e., empirical p-value of $p < .001$).

For comparison, we analyzed our corpus using the correlation method of Monaghan et al. [MSCK14]. In our implementation, we measured the correlation between the pairwise cosine distances produced by Word2Vec and pairwise orthographic edit distances for all pairs of words in our lexicon. The correlation between the Word2Vec semantic distances and the orthographic edit distances in our corpus was $r = 0.0194$, similar to the correlation reported by Monaghan et al. of $r = 0.016$ between the phoneme edit distances and the semantic distances in the monomorphemic English lexicon. We also looked at the correlation between the weighted edit distances produced by SMLKR and the Word2Vec semantic distances. The correlation between these distances was $r = 0.0464$; thus, the weighted edit distance captures more than 5.7 times as much variance as the unweighted edit distance. Further, using the estimated semantic vectors produced by the SMLKR model, we can actually produce new estimates of the semantic distances between the words. The correlation between these estimated semantic distances and the true semantic distances was $r = 0.1028$, revealing much more systematicity than revealed by the simple linear correlation method. The Mantel test with 1,000 permutations produced significant empirical p-values for all correlations ($p < .001$).

Systematicity Not Evenly Distributed Across Lexicon. What could be accounting for the higher degree of systematicity detected with SMLKR? Applying a more expressive model could result in a better fit simply because incidental but inconsequential patterns are being captured. Conversely, SMLKR could be finding phonosemantic sets which the correlation method of Monaghan et al. [MSCK14] is unable to detect. We investigated further by determining what was driving the better fit produced by SMLKR. Monaghan et al. measure per-word systematicity as the change in the lexicon-wide form-meaning correlation that results from removing the word from the lexicon. The more the correlation decreases from removing the word, the more systematic the word is, according to this measure. They compared the distribution of this systematicity measure across the words in the lexicon to the distribution of systematicity in lexicons with randomly shuffled form-meaning assignments, and found that the null hypothesis that the distributions were identical could not be rejected. From this, they conclude that the observed systematicity of the lexicon is not a consequence only of small pockets of sound symbolism, but is rather a feature of the mappings from sound to meaning across the lexicon as a whole. However, it is possible that their methods may not be sensitive enough to find localized phonosemantic sets.

We developed our own measure of per-word systematicity by measuring the per-word regression error of the SMLKR model. We presume words with lower regression errors to be more systematic. A list of the words with the lowest per-word regression error in our corpus can be found in Table 3.1. Notably, many of these words, such as *fluff*, *flutter*, and *flick*, exhibit word beginnings or word endings that have been previously identified as phonaesthemes [Hut98, OS08]. Others exhibit regular onomatopoeia, such as *clang* and *croak*.

We decided to investigate the distribution of systematicity across two-letter word-beginnings in our lexicon using a permutation test. The goal of the permutation test is to estimate a p-value for the likelihood that each set of words sharing a word beginning would exhibit the mean regression error it exhibits, if systematicity is randomly distributed across the lexicon. For each set $\mathcal{S}_\omega$ of words with word-beginning ω, we measured the mean SMLKR regression error of the words in $\mathcal{S}_\omega$. To get an empirical p-value for each $\mathcal{S}_\omega$ with cardinality greater than 5 (i.e., more than 5 word tokens), we randomly chose 10^5 sets of words in the lexicon with the same cardinality, and measured the mean SMLKR regression error for each of these random sets. If r of the randomly assembled sets had a lower mean regression error than $\mathcal{S}_\omega$ did, we assign an empirical p-value of $\frac{r}{10^5}$ to $\mathcal{S}_\omega$. A histogram of empirical p-values is in Fig. 3.2. From the figure, it seems clear that the p-values are not uniformly distributed; instead, an inordinate number of word-beginnings exhibit mean errors that are unlikely to occur if error is distributed arbitrarily across word-beginnings.

We can confirm this observation statistically. On the assumption that systematicity is arbitrarily distributed across word-beginnings, the empirical p-values of the permutation test should approximately conform to a $\text{Unif}(0, 1)$ distribution. We can test this hypothesis using a χ^2 test on the negative logarithms of the p-values [Fis32]. Using this test, we reject the hypothesis that the p-values are uniformly distributed with $p < .0001$ ($\chi^2_{156} = 707.8$). The particular word-beginnings with statistically significant empirical p-values ($p < .05$ after Benjamini-Hochberg [BH95] correction for multiple comparisons) are in Table 3.2. Eight of these ten features are among the 18 two-letter onsets posited to be phonaesthemes by Hutchins [Hut98].

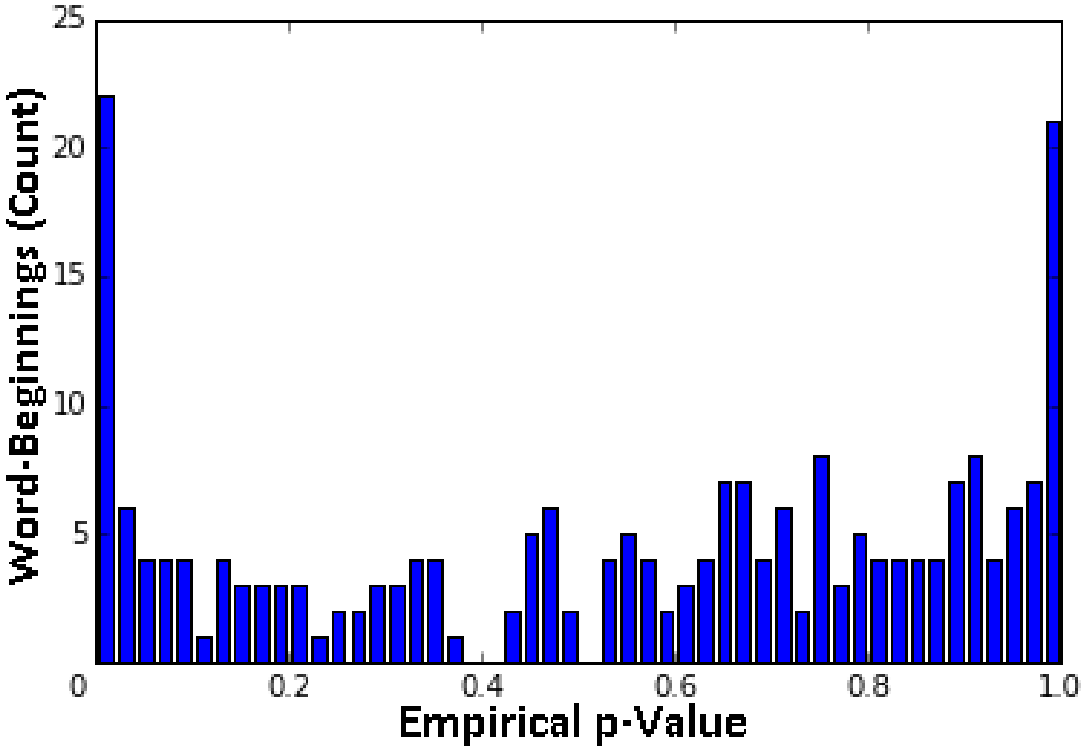

Figure 3.2: Histogram showing distribution of systematicity across two-letter word-beginnings, as measured by permutation-test empirical p-value.

For comparison, Otis and Sagi [OS08] identified eight of Hutchins's 18 two-letter word-beginning candidate phonaesthemes (and 12 two-letter word-beginnings overall) as statistically significant, though they restricted their hypothesis space to only 50 pre-specified word-beginnings and word-endings. We are able to identify just as many candidate phonaesthemes, but with a much less restricted hypothesis space of candidates (225 rather than the 50 in Otis and Sagi's analysis) and with a general model not specifically attuned to finding phonaesthemes in particular, but rather systematicity in general.

3.5.2 Behavioral Evaluation of Systematicity Measure

We empirically tested whether the systematicity measure based on SMLKR regression error accords with naïve human judgments about how well-suited a word's form is to its meaning (its "phonosemantic feeling") [Ste02]. We recruited 60 native English-speaking participants through Mechanical Turk, and asked them to judge the phonosemantic feeling of the 60 words in Table 3.1 on a sliding scale from 1 to 5.[2] We used Cronbach's α to measure inter-annotator reliability at $\alpha = 0.96$, indicating a high degree of inter-annotator reliability [Cro51, Geo00]. The results showed that the words in the SMLKR list were rated higher for phonosemantic feeling than the words in the Correlation and Random lists. We fit a parametric linear mixed-effects model to the phonosemantic feeling judgments [BDB08], as implemented in the `lme4` library for R. As fixed effects, we entered the list identity (SMLKR, Correlation, Random), the word length, and the log frequency of the word in our corpus. Our random effects structure included a random intercept for word, and random subject slopes for all fixed effects, with all correlations allowed (a "maximal" random-effects structure [BLST13]). Including list identity in the maximal mixed-effects model significantly improved model fit ($\chi^2_{11} = 126.08$, $p < 10^{-6}$). Post-hoc analysis revealed that the SMLKR list elicited average suitability judgments that were 0.49 points higher than the Random list ($p < 10^{-6}$) and 0.59 points higher than the Correlation list ($p < 10^{-6}$). Post-hoc analysis did not find a significant difference in suitability judgments between the Random and Correlation lists ($p > .16$).[3]

[2] Participants were given the following guidance: "Your job is to decide how well-suited each word is to what it means. This is known as the 'phonosemantic feeling.' Basically, most people feel like some of the words in their native language sound right, given what they mean." Full instructions and experiment available at `http://goo.gl/Z6Lzlp`

[3] Post hoc analyses were produced by comparing the items in only two of the lists at a time, and fitting the same mixed-effects model as above.

3.6 Conclusion

In this paper, we proposed SMLKR, a novel algorithm that can learn weighted string edit distances that minimize kernel regression error. We succeed in applying this algorithm to the problem of finding form-meaning systematicity in the monomorphemic English lexicon. Our algorithm offers improved global predictions of word-meaning given word-form at the lexicon-wide level. We show that this improvement seems related to localized pockets of form-meaning systematicity such as those previously uncovered in behavioral and corpus analyses. Unlike previous lexicon-wide analyses, we find that form-meaning systematicity is not randomly distributed throughout the English lexicon. Moreover, the measure of systematicity that we compute using SMLKR accords significantly with human raters' judgments about form-meaning correspondences in English.

Future work may investigate to what extent the SMLKR model can predict human intuitions about form-meaning systematicity in language. We do not know, for instance, if our model can predict human semantic judgments of novel words that have never been encountered. This is a question that has received attention in the market research literature, where new brand names are tested for the emotions they elicit [Kli00]. We would also like to investigate the degree to which our statistical model predicts the behavioral effects of phonosemantic systematicity during human semantic processing that have been reported in the psycholinguistics literature. Our model makes precise quantitative predictions that should allow us to address these questions.

While developing our model on preliminary versions of the monomorphemic lexicon, we noticed that the model detected high degrees of systematicity in words with suffixes such as *-ate* and *-tet* (e.g., *quintet*, *quartet*). We removed such words in the final analysis since they are polymorphemic, but this observation suggests that our algorithm may have applications in unsupervised morpheme discovery.

Finally, we would like to test our model using other representations of word-form and word-meaning. We chose to use orthographic rather than phonetic representations of words because of the variance in pronunciation present in the dialects of English that are manifested in our corpus. However, it would be interesting to verify our results in a phonological setting, perhaps using a monodialectal corpus. Moreover, previous local-level analyses suggest that systematicity seems to be concentrated in word-beginnings and word-endings. Thus, it may be worthwhile to augment the representation of edit distance in our model by making it context-sensitive. Future work could also test whether a more interpretable meaning-space representation such as that provided by binary WordNet feature vectors reveals patterns of systematicity not found using a distributional semantic space.

Chapter 3, in part, is a reprint of the material as it appears in Gutierrez, E.D.; Levy, Roger; Bergen, Benjamin. “Finding Non-Arbitrary Form-Meaning Systematicity Using String-Metric Learning for Kernel Regression”, Proceedings of the Association for Computational Linguistics, 2016. The dissertation author was the primary investigator and author of this paper.

Table 3.1: *Left*: Most systematic words according to SMLKR. *Center*: Most systematic words according to the leave-one-out correlation method proposed by Monaghan et al. [MSCK14]. *Right*: Randomly generated list for comparison.

SMLKR	**Correlation**	**Random**
gurgle	emu	tunic
tingle	nexus	decay
hoop	asylum	skirmish
chink	ethic	scroll
swirl	odd	silk
ladle	slime	prom
flick	snare	knob
wobble	scarlet	havoc
tangle	deem	irate
knuckle	balustrade	veer
glitter	envoy	wear
twig	scrape	phone
fluff	essay	surgeon
rasp	ambit	hiccup
quill	echo	bowel
flutter	onus	sack
whirl	exam	lens
croak	pirouette	hovel
squeal	kohl	challenge
clang	chandelier	box

Table 3.2: Word-beginnings with mean errors lower than predicted by random distribution of errors across lexicon. **Bold** are among the phonaesthemes identified by Hutchins [Hut98]. *Italics* were identified by Otis and Sagi [OS08].

Onset	***p*-value**
fl-	$< 1 \times 10^{-4}$
sn-	$< 1 \times 10^{-4}$
sw-	$< 1 \times 10^{-4}$
tw-	$< 1 \times 10^{-4}$
gl-	1×10^{-3}
sl-	1×10^{-3}
bu-	1×10^{-3}
mu-	2×10^{-3}
wh-	2×10^{-3}
sc-/sk-	3×10^{-3}

Chapter 4

Detecting Cross-Cultural Differences Using a Multilingual Topic Model

4.1 Introduction

Recent years have seen a growing interest in text-mining applications aimed at uncovering public opinions and social trends [FRMQ07, MCQ08, GB11, PP11]. They rest on the assumption that the language we use is indicative of our underlying worldviews. Research in cognitive and sociolinguistics suggests that linguistic variation across communities systematically reflects differences in their cultural and moral models and goes beyond lexicon and grammar [KÖ4, LW12]. Cross-cultural differences manifest themselves in text in a multitude of ways, most prominently through the use of explicit opinion vocabulary with respect to a certain topic (e.g. "policies that *benefit* the poor"), idiomatic and metaphorical language (e.g. "the company is *spinning its* *wheels*") and other types of figurative language, such as irony or sarcasm.

The connection between language, culture and reasoning remains one of the central research questions in psychology. Thibodeau and Boroditsky [TB11] investigated how metaphors 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, while the first group tended to opt for preventive measures (e.g. stronger social policies), the second group converged on punishment- or restraint-oriented measures. According to Thibodeau and Boroditsky, their results demonstrate that metaphors have profound influence on how we conceptualize and act with respect to societal issues. This suggests that in order to gain a full understanding of social trends across populations, one needs to identify subtle but systematic linguistic differences that stem from the groups' cultural backgrounds, expressed both literally and figuratively. Performing such an analysis by hand is labor-intensive and often impractical, particularly in a multilingual setting where expertise in all of the languages of interest may be rare.

With the rise of blogging and social media, applying text-mining techniques to aid political and social science has become an active research area in natural language processing (NLP) [GS13]. NLP techniques have been successfully used for a number of tasks in political science, including automatically estimating the influence of particular politicians in the US senate [FRMQ07], identifying lexical features that differentiate political rhetoric of opposing parties [MCQ08], predicting voting patterns of politicians based on their use of language [GB11], and predicting political affiliation of Twitter users [PP11]. Fang et al. [FSSY12] addressed the problem of automatically detecting and visualising the contrasting perspectives on a set of topics attested in multiple distinct corpora. While successful in their tasks, all of these approaches focused on monolingual data and did not reach beyond literal language. In contrast, we present a method that detects fine-grained cross-cultural differences from multilingual data, where such differences abound, expressed both literally and figuratively. Our method brings together opinion mining and cross-lingual topic modelling techniques for this purpose. Previous approaches to cross-lingual topic modelling [BGB09, JD10] addressed the problem of mining common topics from multilingual corpora. We present a model that learns such common topics, while simultaneously identifying lexical features that are indicative of the underlying differences in perspectives on these topics by speakers of English, Spanish and Russian. These differences are mined from multilingual, non-parallel datasets of Twitter and news data. In contrast to previous work, our model does not merely output a list of monolingual lexical features for manual comparison, but also automatically infers multilingual contrasts.

Our system (1) uses word-document co-occur-rence data as input, where the words are labeled as *topic* words or *perspective* words; (2) finds the highest-likelihood dictionary between topic words in the two languages given the co-occurrence data; (3) finds cross-lingual topics specified by distributions over topic-words and perspective-words; and (4) automatically detects differences in perspective-word distributions in the two languages. We perform a behavioural evaluation of a subset of the differences identified by the model and demonstrate their psychological validity. Our data and dictionaries are available from the first author upon request.

4.2 Related work

View detection. Identifying different viewpoints is related to the well-studied area of subjectivity detection, which aims at exposing opinion, evaluation, and speculation in text [WWB$^+$04]. There is a large literature on identifying such opinions and attributing it to specific people [ARW11, AJDDR12]. In our work, we are less interested in such explicit local forms of subjectivity, instead aiming at detecting more general contrasts across linguistic communities.

Another line of research has focused on inferring author attributes such as gender, age [GY09], location [JKPT07], or political affiliation [PP11]. Such studies make use of of latent aspects of language, including syntactic style, discourse characteristics, as well as lexical choice. The models used for this are typically binary classifiers trained in a fully supervised fashion such as SVMs. In contrast, in our task, we automatically infer the topic distributions and find topic-specific contrasts.

Probabilistic topic models. Probabilistic topic models have proven useful for a variety of semantic tasks, such as selectional-preference induction [OS10, REE10], sentiment analysis [BGR10] and studying the evolution of concepts and ideas [HJM08]. The goal of a topic model is to characterize observed data in terms of a much smaller set of unobserved, semantically coherent topics. A particularly popular probabilistic topic model is Latent Dirichlet Allocation (LDA) [BNJ03]. Under its assumptions, each document has a unique mix of topics, and each topic is a distribution over terms in the vocabulary. A topic is chosen for every word token according to the topic mix of the document to which it belongs, and then the word's identity is drawn from the corresponding topic's distribution.

Handling multilingual corpora. LDA is designed for monolingual text and thus it lacks the structure necessary to model cross-lingually valid topics. While topic models can be trained individually on two languages and then the acquired topics can be matched, the correspondences between the topics for the two terms will be highly unstable. To address this, Boyd-Graber and Blei [BGB09] (MuTo) and Jagarlamudi and Daumé [JD10] (JointLDA) introduced the notion of cross-lingually valid concepts associated with different terms in different languages, using bilingual dictionaries to model topics across languages. Based on a model by Haghighi et al. [HLBKK08], MuTo is capable of learning translations–i.e., matching between terms in the different languages being compared. The Polylingual Topic Model of Mimno et al. [MWN$^+$09] is another approach to finding topics in multilingual corpora, but it requires tuples composed of comparable documents in each language of the corpus.

Topic models for view detection. LDA also assumes that the distribution of each topic is fixed across all documents in a corpus. Therefore, a topic associated with, e.g., *war* will have the same distribution over the lexicon regardless of whether the document was taken from a pro-war editorial or an anti-war speech. However, in reality we may expect a single topic to exhibit systematic and predictable variations in its distribution based on authorship.

The cross-collection LDA model of Paul and Girju [PG09] addresses this by specifically aiming to expose viewpoint differences across different document collections. Ahmed and Xing [AX10] proposed a similar model for detecting ideological differences. Fang et al.'s [FSSY12] Cross-Perspective Topic (CPT) model breaks up the terms in the vocabulary into topic terms and perspective terms with different generative processes, and differentiates between different collections of documents within the

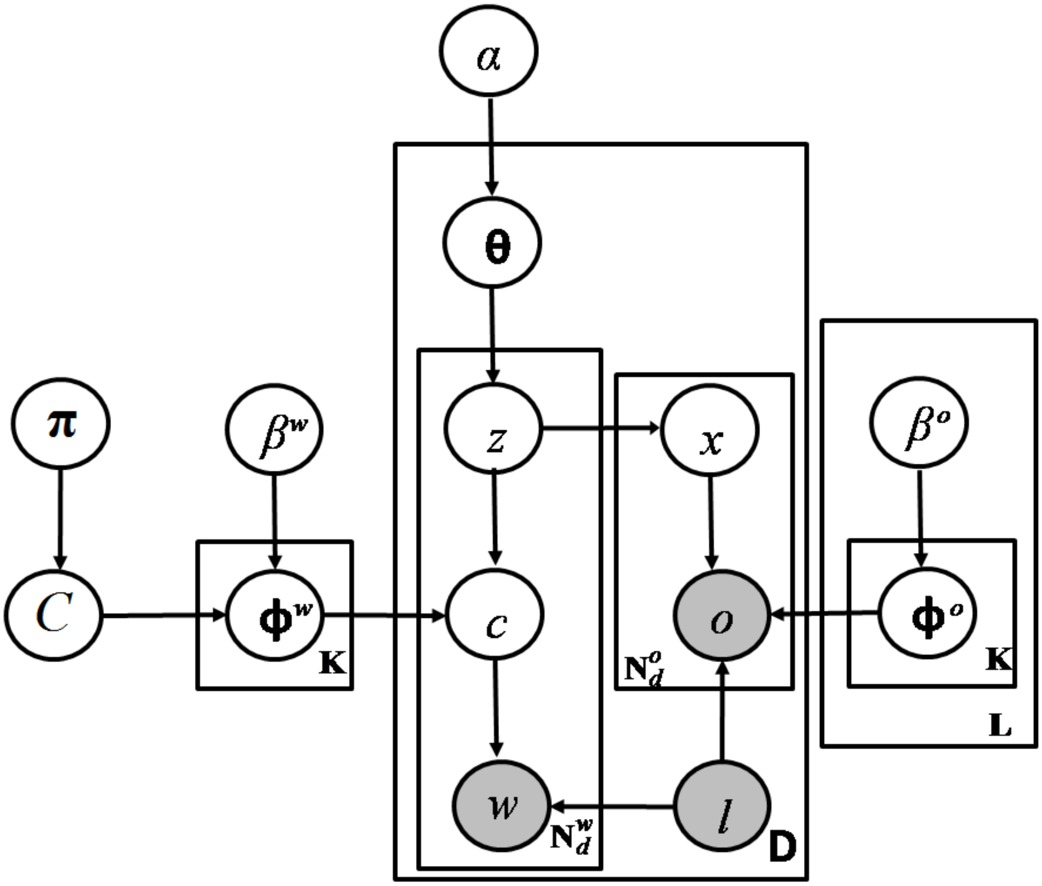

Figure 4.1: Basic generative model.

corpus. The topic terms are assumed to be generated as in LDA. However, the distribution of perspective terms in a document is taken to be dependent on both the topic mixture of the document as well as the collection from which the document is drawn.

Recent works proposed models for specific types of data. Qiu and Jiang [QJ13] use user identities and interactions in threaded discussions, while Gottipati et al. [GQS$^+$13] developed a topic model for Debatepedia, a semi-structured resource in which arguments are explicitly enumerated. However, all of these models perform their analyses on monolingual datasets. Thus, they are useful for comparing different ideologies expressed in the same language, but not for cross-linguistic comparisons.

4.3 Method

The goal of our model is to analyse large, non-parallel, multilingual corpora and present cross-lingually valid topics and the associated perspectives, automatically inferring the differences in conceptualization of these topics across cultures. Following Boyd-Graber and Blei [BGB09] and Jagarlamudi and Daumé [JD10], our distributions of latent topics range over latent, cross-lingual *topic concepts* that manifest themselves as language-specific *topic words*. We use bilingual dictionaries, containing words in one language and their translations in another language, to represent the topic concepts. These are represented as a bipartite graph, with each translation entry being an edge and each topic word in the two languages being a vertex. While the topic words are tied together by the translation dictionary, the perspective words can vary freely across languages. Following Fang et al. [FSSY12], we treat nouns as topic words and verbs and adjectives as perspective words[1]. The model assumes that adjective and verb tokens in each document are assigned to topics in proportion to the topic assignments of the topic word tokens. Then, the perspective term for this topic is drawn depending on the topic assignment and the language of the speaker.

4.3.1 Basic Generative Model

Given the languages $\ell \in \{a, b\}$, our model infers the distributions of multilingual topics and language-specific perspective-words (Fig. 4.1), as follows:

1. Draw a set C of concepts (u, v) matching topic word u from language a to topic word v from language b, where the probability of concept (u, v) is proportional to a

[1] This approximation was adopted for convenience, computational efficiency and ease of interpretation. However, in principle our method does not depend on it, since it can be applied with all content words as topic or perspective words.

prior $\pi_{u,v}$ (e.g. based on information from a translation dictionary).

2. Draw multinomial distributions:
    - For topic indices $k \in \{1, ..., K\}$, draw language-independent topic-concept distributions $\phi_k^w \sim \text{Dir}(\beta^w)$ over pairs $(w_a, w_b) \in C$.
    - For topic indices $k \in \{1, ..., K\}$ and languages $\ell \in \{a, b\}$, draw language-specific perspective-term distributions $\phi_k^{\ell,o} \sim \text{Dir}(\beta^o)$ over perspective-terms in language ℓ.
3. For each document $d \in \{1, ..., D\}$ with lang. ℓ_d:
    - Draw topic weights $\theta_d \sim \text{Dir}(\alpha)$
    - For each topic-word index $i \in \{1, ..., N_d^w\}$ of document d:
        - Draw topic $z_i \sim \theta_d$
        - Draw topic concept $c_i = (w_a, w_b) \sim \phi_{z_i}^w$, and select w_{ℓ_d} as the member of that pair corresponding to language ℓ_d.
    - For each perspective-word index $j \in \{1, ..., N_d^o\}$ of document d:
        - Draw topic $x_j \sim \text{Uniform}(z_{w_1}, ..., z_{w_{N_d^o}})$
        - Draw perspective-word $o_j \sim \phi_{x_j}^{\ell,o}$

4.3.2 Model Variants

We have experimented with several variants of our model, in order to account for the translation of polysemous words, adapt the translation model to the corpus used, and to handle words for which no translation is found. Two separate approaches to translation exist in the multi-lingual topic modelling literature, focusing on distinct problematic aspects of translation. Boyd-Graber and Blei [BGB09] infer a bilingual matching using both co-occurrence data and a bilingual dictionary as input. On the other hand, by reducing the multiple translations available in multilingual dictionaries to a set of one-to-one correspondences, much information about the meaning of a word in relation to the words in the other language is lost; the MuTo [JD10] model accounts for this by allowing multiple correspondences between terms in the languages. We empirically investigate the effect of different translation models on the performance of the topic model, by integrating both of these approaches into a broader framework according to the presence or absence of three attributes, as follows:

1. Single variants of the model match each topic term in a language with at most one topic term in the other language.

   Multiple variants allow each term to match to multiple other words in the other language.

2. Infer variants allow higher-likelihood matchings to be inferred from the data.

   Static variants treat the matchings as fixed, which is equivalent to assigning a probability of 0 or 1 to every edge in our bipartite graph C.

3. Relegate variants relegate all unmatched words in each language to a single separate background topic distinct from the topics that are learned for the matched topic words. This is akin to forcing the probability for currently unmatched words to 0 in all topics except for one, and forcing the probability of all currently matched words to 0 in this topic.

Include variants do not restrict the assignment unmatched words; they are assigned to the same set of topics as the matched words.

We test the following six variants: SingleStaticRelegate, SingleStaticInclude, SingleInferRelegate, SingleInferInclude, MultipleStaticRelegate, and MultipleStaticInclude. We do not test MultipleInfer variants because of the complexity of inferring a multiple matching in a bipartite graph.

4.3.3 Learning & Inference

For all variants, a collapsed Gibbs sampler can be used to infer topics $\phi^{\ell,o}$ and ϕ^{w}, per-document topic distributions θ, as well as topic assignments $\mathbf{z}$ and $\mathbf{x}$. This corresponds to the S-step below. For Infer variants, we follow Boyd-Graber and Blei in using an M-step involving a bipartite graph matching algorithm to infer the matching $\mathbf{m}$ that maximizes the posterior likelihood of the matching.

S-Step: Sample topics for words in the corpus using a collapsed Gibbs sampler. For topic-word $w_i = u$ belonging to document d, if the word occurs in concept $c_i = (u, v)$, then sample the topic and entry according to:

$$p(z_i = k, c_i = (u,v) \mid w_i = u, \mathbf{z}_{-i}, C) \propto \frac{N_{dk} + \alpha_k}{\sum_j \left(N_{dj} + \alpha_j\right)} \times \frac{N_{k(u,v)} + \beta_k^w}{\sum_{v'} \left(N_{k(u,v')} + \beta_k^w\right)}$$

where the sum in the denominator of the first term is over all topics, and in the second term is over all words matched to u. N_{dk} is the count of topic-words of topic k in document d, $N_{k(u,v)}$ is the count of topic-words either of type u or of type v assigned to topic k in all the corpora.[2] For perspective-word $o_i = n$, sample the topic according to:

$$p(z_i = k | o_i = n, \mathbf{z}_{-i}, C) \propto \frac{N_{dk}}{\sum_j N_{dj}} \times \frac{N^{\ell_d}_{kv} + \beta^o_k}{\sum\limits_m \left(N^{\ell_d}_{km} + \beta^o_k \right)}$$

where the sum in the second term of the denominator is over the perspective-word vocabulary of language ℓ_d; N_{dk} is the count of *topic* words in document d with topic k; and $N^{\ell_d}_{km}$ is the count of perspective-word m being assigned topic k in language ℓ_d. Note that in all the counts above, the current word token i is omitted from the count.

Given our sampling assignments, we can then estimate θ^d, $\phi^{\ell,o}$, and ϕ^w as follows:

$$\begin{aligned} \hat{\theta}_{kd} &= \frac{N_{dk} + \alpha_k}{\sum\limits_k \left(N_{dk} + \alpha_k \right)}, \\ \hat{\phi}^w_{k(u,v)} &= \frac{N_{k(u,v)} + \beta^w_{(u,v)}}{\sum\limits_{v'} \left(N_{k(u,v')} + \beta^w_{(u,v')} \right)}, \\ \hat{\phi}^{\ell,o}_{nk} &= \frac{N_{kn} + \beta^o_n}{\sum\limits_m \left(N^{\ell}_{km} + \beta^o_n \right)}. \end{aligned}$$

M-Step: *(for* Infer *variants only)*: Run the Jonker-Volgenant [JV87] bipartite matching algorithm to find the optimal matching C given some weights. For topic-term u from language a and topic-term v from language b, our weights correspond to the log of the posterior odds that the occurrences of u and v come from a matched

[2] In Relegate variants, for u unmatched z_i is sampled as:

$$p(z_i = k | w_i = u, \mathbf{z}_{-i}, C) \propto \frac{N_{dk} + \alpha_k}{\sum\limits_k \left(N_{dk} + \alpha_k \right)},$$

which can be seen as $\beta^w_{u\cdot} \to \infty$ for unmatched terms.

topic distribution, as opposed to coming from unmatched distributions:

$$\mu_{u,v} = \sum_{k \setminus \{a*,b*\}} \left(N_{k(u,v)} \log \hat{\phi}^{w}_{k(u,v)} \right) \\ - N_u \log \hat{\phi}^{w}_{k(u,\cdot)} - N_v \log \hat{\phi}^{w}_{k(\cdot,v)} + \pi_{u,v},$$

where N_u is the count of topic-term u in the corpus. This expression can also be interpreted as a kind of pointwise mutual information [HLBKK08]. The Jonker-Volgenant algorithm has time complexity of at most $O(V^3)$, where V is the size of the lexicon [JV87].

4.3.4 Inference of Perspective-Word Contrasts

Having learned our model and inferred how likely perspective-terms are for a topic in a given language, we seek to know whether these perspectives differ significantly in the two languages. More precisely, can we infer whether word m in language a and the equivalent word n in language b have significantly different distributions under a topic k? To do this, we make the assumption that the perspective-words in languages a and b are in one-to-one correspondence to each other. Recall that, for a given topic k and language ℓ, N^{ℓ}_{km} is the count for term m and $\phi^{\ell,o}_{k,m}$ is the probability for word m in language ℓ. Just as we collect the probabilities into word-topic distribution vectors $\phi^{\ell,o}_{k}$, we collect the counts into word-topic count vectors $[N^{\ell}_{k1}, N^{\ell}_{k2}, ..]$. Then, since our model assumes a prior over the parameter vectors $\phi^{\ell,o}_{k}$, we can infer the likelihood for that observed word-topic counts N^{a}_{km} and N^{b}_{kn} were drawn from a single word-topic-distribution prior denoted by $\breve{\phi} := \phi^{a,o}_{km} = \phi^{b,o}_{kn}$. Below all our probabilities are conditioned implicitly on this event as well as on N^{a}_{k} and N^{b}_{k} being fixed.

Denote the total count of word tokens in topic k from language ℓ by $N_k^\ell = \sum_m N_{km}^\ell$. Now, we derive the probability that we observe a ratio greater than δ between the proportion of words in topic k that belong to word type m in language a and to corresponding word type n in language b:

$$p\left(\frac{N_{km}^a}{N_k^a}\frac{N_k^b}{N_{kn}^b} \geq \delta\right) \qquad + \qquad p\left(\frac{N_{kn}^b}{N_k^b}\frac{N_k^a}{N_{km}^a} \geq \delta\right) \tag{4.1}$$

By symmetry, it suffices to derive an expression for the first term. We note that the inequality in the probability is equivalent to a sum over a range of values of N_{km}^a and N_{kn}^b. By rearranging terms, applying the law of conditional probability to condition on the term $\breve{\phi}$, and exploiting the conditional independence of N_{km}^a and N_{km}^b given $\breve{\phi}$, N_k^a, and N_k^b, we can rewrite this first term as

$$\sum_{x=0}^{N_k^b} \sum_{y=x\delta N^{a/b}}^{N_k^a} \int p(N_{kn}^b = x|\breve{\phi})p(N_{km}^a = y|\breve{\phi})p(\breve{\phi})d\breve{\phi},$$

where $N^{a/b} = \frac{N_k^a}{N_k^b}$. Recall that $\phi_k^{\ell,o} \sim \mathrm{Dir}(\beta^o)$ under our model. Assume a symmetric Dirichlet distribution for simplicity. It can then be shown that the marginal distribution of $\breve{\phi}$ is $\breve{\phi} \sim \mathrm{Beta}(\beta^o, (V-1)\beta^o)$, where V is the total size of the perspective-word vocabulary. Similarly, it can be shown that the marginal distribution of N_{km}^ℓ given $\phi_k^{\ell,o}$ is $N_{km}^\ell \sim \mathrm{Binom}(N_k^\ell, \phi_i^{\ell,o})$ for $\ell \in \{a, b\}$. Therefore, the integrand above is proportional to the beta-binomial distribution with number of trials $N_k^a + N_k^b$, successes $x+y$, and parameters β^o and $(V-1)\beta^o$, but with partition function $\binom{N_k^a}{y}\binom{N_k^b}{x}$. Denote the PMF of this distribution by $f(N_k^a + N_k^b, x+y, \beta^o)$. Then expression (4.1) above becomes:

$$\sum_{x=0}^{N_k^b} \sum_{y=x\delta N^{a/b}}^{N_k^a} f(N_k^a + N_k^b, x+y, \beta^o)$$

$$+\sum_{x=0}^{N_k^a} \sum_{y=x\delta N^{b/a}}^{N_k^b} f(N_k^a + N_k^b, x+y, \beta^o). \quad (4.2)$$

We cannot observe N_{kb}^a, N_{kn}^b, N_k^a and N_k^b explicitly, but we can estimate them by obtaining posterior samples from our Gibbs sampler. We substitute these estimates into expression (4.2).

4.4 Experiments

4.4.1 Data

Twitter Data. We gathered Twitter data in English, Spanish and Russian during the first two weeks of December 2013 using the Twitter API. Following previous work [PECX10], we treated each Twitter user account as a document. We then tagged each document for part-of-speech, and divided the word tokens in it into topic-words and perspective-words. We constructed a lexicon of 2,000 topic terms and 1,500 perspective-terms for each language by filtering out any terms that occurred in more than 10% of the documents in that language, and then selecting the remaining terms with the highest frequency. Finally, we kept only documents that contained 4 or more topic words from our lexicon. This left us with 847,560 documents in English (4,742,868 topic-word and 1,907,685 perspective-word tokens); 756,036 documents in Spanish (4,409,888 topic-word and 1,668,803 perspective-word tokens); and 260,981 documents in Russian (1,621,571 topic-word and 981,561 perspective-word tokens).

News Data. We gathered all the articles published online during the year 2013 by the state-run media agencies of the United States (Voice of America or "VOA"–English), Russia (RIA Novosti or "RIA"–Russian), and Venezuela (Agencia Venezolana de Noticias or "AVN"–Spanish). These three news agencies were chosen because they not only provide media in three distinct languages, but they are guided by the political world-views of three distinct governments. We treated each news article as a document, and removed duplicates. Once again, we constructed a lexicon of 2,000 topic terms and 1,500 perspective-terms using the same criteria as for Twitter, and kept only documents that contained 4 or more topic words from our lexicon. This left us with 23,159 articles (10,410,949 tokens) from VOA, 41,116 articles (11,726,637 tokens) from RIA, and 8,541 articles (2,606,796 tokens) from AVN.

Dictionaries. To create the translation dictionaries, we extracted translations from the English, Spanish, and Russian editions of Wiktionary, both from the translation sections and the gloss sections if the latter contained single words as glosses. Multi-word expressions were universally removed. We added inverse translations for every original translation. From the resulting collection of translations, we then created separate translation dictionaries for each language and part-of-speech tag combination.

In order to give preference to more important translations, we assigned each translation an initial weight of $1+\frac{1}{r}$, where r was the rank of the translation within the page. Since a translation (or its inverse) can occur on multiple pages, we aggregated these initial weights and then assigned final weights of $1 + \frac{1}{r'}$, where r' was the rank after aggregation and sorting in descending order of weights.

4.4.2 Experimental Conditions

To evaluate the different variants of our model, we held out 30,000 documents (test set) during training. We plugged in the estimates of ϕ^w and C acquired during training using the rest of the corpus to produce a likelihood estimate for these held-out documents. All models were initialized with the prior matching determined by the dictionary data. For each number of topics K, we set α to $50/K$and the β variables to 0.02, as in Fang et al. [FSSY12]. For the MULTIPLE variants, we set $\pi_{i,j} = 1$ if i and j share an entry and 0 otherwise. For INFER variants, only three M-steps were performed to avoid overfitting, at 250, 500, and 750 iterations of Gibbs sampling, following the procedure in Boyd-Graber and Blei [BGB09].

4.4.3 Comparison of model variants

In order to compare the variants of our model, we computed the perplexity and coherence for each variant on TWITTER and NEWS, for English–Spanish and English–Russian language pairs.

Perplexity is a measure of how well a model trained on a training set predicts the co-occurrence of words on an unseen test set $\mathcal{H}$. Lower perplexity indicates better model fit. We evaluate the held-out perplexity for topic words w_i and perspective-words o_i separately. For topic words, the perplexity is defined as $exp(-\sum_{w_i \in \mathcal{H}} logp(w_i)/N^w)$. As for standard LDA, exact inference of $p(w_i)$ is intractable under this model. Therefore we adapted the estimator developed by Murray and Salakhutdinov [MS09] to our models.

Coherence is a measure inspired by pointwise mutual information [NLGB10]. Let $D(v)$ be the the number of documents with at least one token of type v and let $D(v, w)$

be the number of documents containing at least one token of type v and at least one token of type w. Then Mimno et al. [MWT+11] define the coherence of topic k as

$$\frac{1}{\binom{M}{2}} \sum_{m=2}^{M} \sum_{\ell=1}^{m-1} \log \frac{D(v_m^{(k)}, v_\ell^{(k)}) + \epsilon}{D(v_\ell^{(k)})},$$

where $V^{(k)} = (v_1^{(k)}, ..., v_M^{(k)})$ is a list of the M most probable words in topic k and ϵ is a small smoothing constant used to avoid taking the logarithm of zero. Mimno et al. [MWT+11] find that coherence correlates better with human judgments than do likelihood-based measures. Coherence is topic-specific measure, so for each model variant we trained, we computed the median topic coherence across all the topics learned by the model. We set $\epsilon = 0.1$.

Model performance and analysis. Fig. 4.2 shows perplexity for the variants as a function of the number of iterations of Gibbs sampling on the English-Spanish NEWS corpus. The figure confirms that 1000 iterations of Gibbs sampling on the NEWS corpus was sufficient for convergence across model variants. We omit figures for English-Russian and for the TWITTER corpus, since the patterns were nearly identical. Figure 4.3 shows how perplexity varies as a function of the number of topics. We used this information to choose optimal models for the different corpora. The optimal number of topics was $K = 175$ for the English-Spanish NEWS corpus, $K = 200$ for the English-Russian NEWS, $K = 325$ for the English-Spanish TWITTER, and $K = 300$ for the English-Russian TWITTER. Although the optimal number of topics varied across corpora, the relative performance of the different models was the same. In all of our corpora, the MULTIPLE variants provided better fits than their corresponding SINGLE variants. There are several explanations for this. For one, the MULTIPLE variants are able to exploit the information from multiple translations, unlike the Single variants, which discarded all but one translation per word. For another, the matchings produced by the SingleInfer variants can be purely coincidental and the result of overfitting (see some examples below). Include variants performed markedly better than Relegate variants. Infer variants improved model fit compared to Static variants, but required more topics to produce optimal fit.

Recall that we performed an M-step in the Infer variants 3 times, at 250, 500, and 750 iterations. As noted in §4.3.3, the M-step in the Infer variants maximizes the posterior likelihood of the matching. However, Fig. 4.2 shows that this maximization causes held-out perplexity to increase substantially just after the first matching M-step, around 250 iterations, before decreasing again after about 50 more iterations of Gibbs sampling. We believe that this happens because the M-step is maximizing over expectations that are approximate, since they are estimated using Gibbs sampling. If the sampler has not yet converged, then the M-step's maximization will be unstable. We found support for this explanation when we re-ran the Infer variants using 1000 iterations between M-steps, giving the Markov chain enough time to converge. After this change, perplexity went down immediately after the M-step and kept decreasing monotonically, rather than increasing after the M-step before decreasing. However, this did not result in a significantly lower final perplexity or coherence and thus did not change the relative performance of the models. In addition, Fig. 4.2 suggests that the second and third M-steps (at 500 and 750 iterations, respectively) had little effect on perplexity. In light of the high computational expense of each inference step, this suggests in practice a single inference step may be sufficient.

Fig. 4.4 shows that the MultipleStaticInclude variant was also the superior model as measured by median topic coherence. Once again, this general pattern held true for the English-Russian pair and Twitter corpora. Overall, the results show that MultipleStaticInclude provides superior performance across measures, corpora, topic numbers, and languages. We therefore used this variant in further data analysis and evaluation. Incidentally, the observed decrease in topic coherence as K increases is expected, because as K increases, lower-likelihood topics tend to be more incoherent [MWT+11]. Experiments by Stevens et al. [SKAB12] show that this effect is observed for LDA-, NMF-, and SVD-based topic models.

Cross-linguistic matchings. The matchings inferred by the SingleInferInclude variant were of mixed quality. Some of the matchings corrected low-quality translations in the original dictionary. For instance, our prior dictionary matched *passage* in English to *pasaje* in Spanish. Though technically correct, the dominant meaning of *pasaje* is *[travel] ticket*. The Twitter model correctly matched *passage* to *ruta* instead. Many of the matchings learned by the model did not provide technically correct translations, yet were still revelatory and interesting. For instance, the dictionary translated the Spanish word *pito* as *cigarette* in English. However, in informal usage this word refers specifically to cannabis cigarettes, not tobacco cigarettes. The Twitter model matches *pito* to the English slang word *weed* instead. The Spanish word *Siria* (Syria) was unmatched in the prior dictionary; the News model matched it to the word *chemical*, which makes sense in the context of extensive reporting of the usage of chemical weapons in the ongoing Syrian conflict.

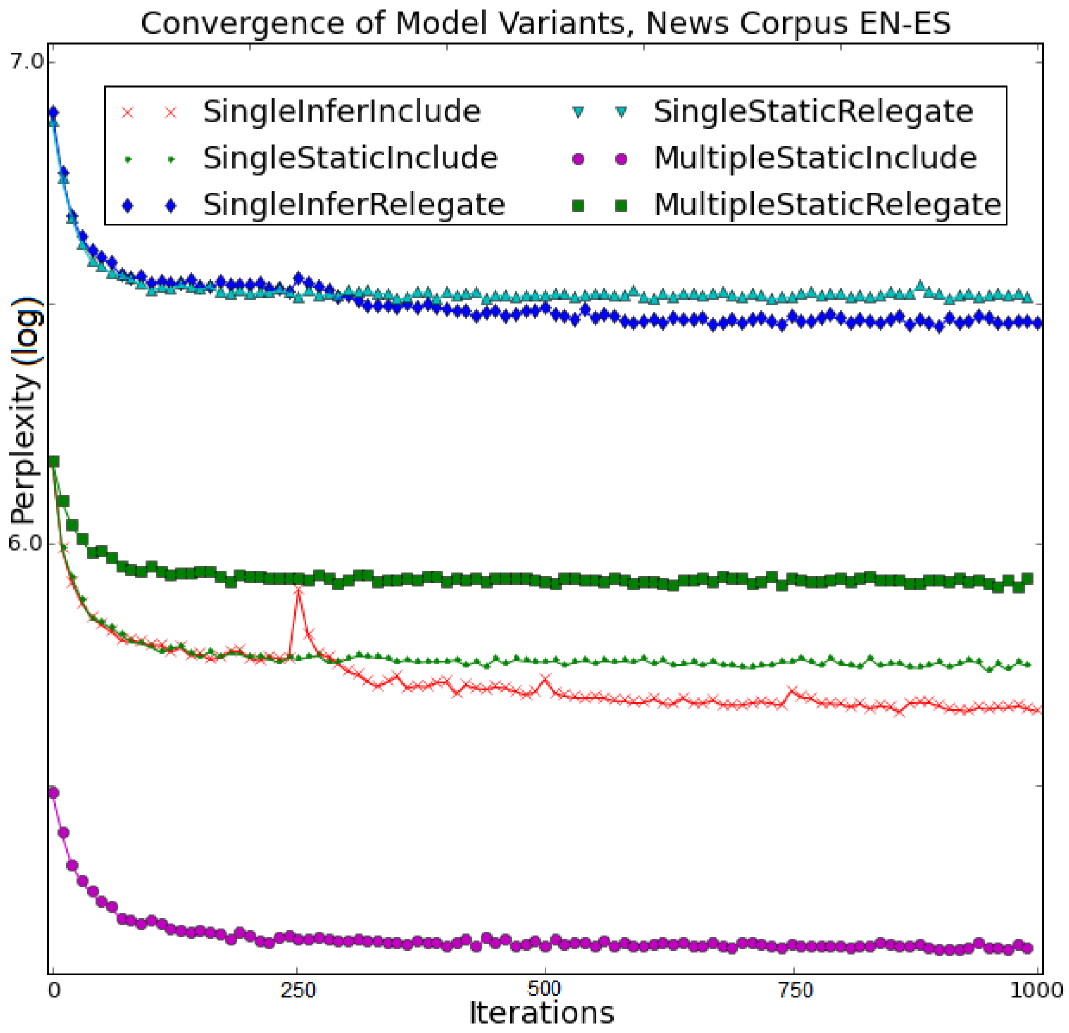

Figure 4.2: Perplexity of different model variants for different numbers of iterations at K=175.

4.4.4 Data analysis and discussion

We have conducted a qualitative analysis of the topics, perspectives and contrasts produced by our models for English–Spanish and English–Russian, Twitter and News datasets. While the topics were coherent and consistent across languages, sets of perspective words manifested systematic differences revealing interesting cross-cultural contrasts. Fig. 4.5 and 4.7 show the top perspective words discovered by the model for the topic of *finance* and *economy* in English and Spanish News and

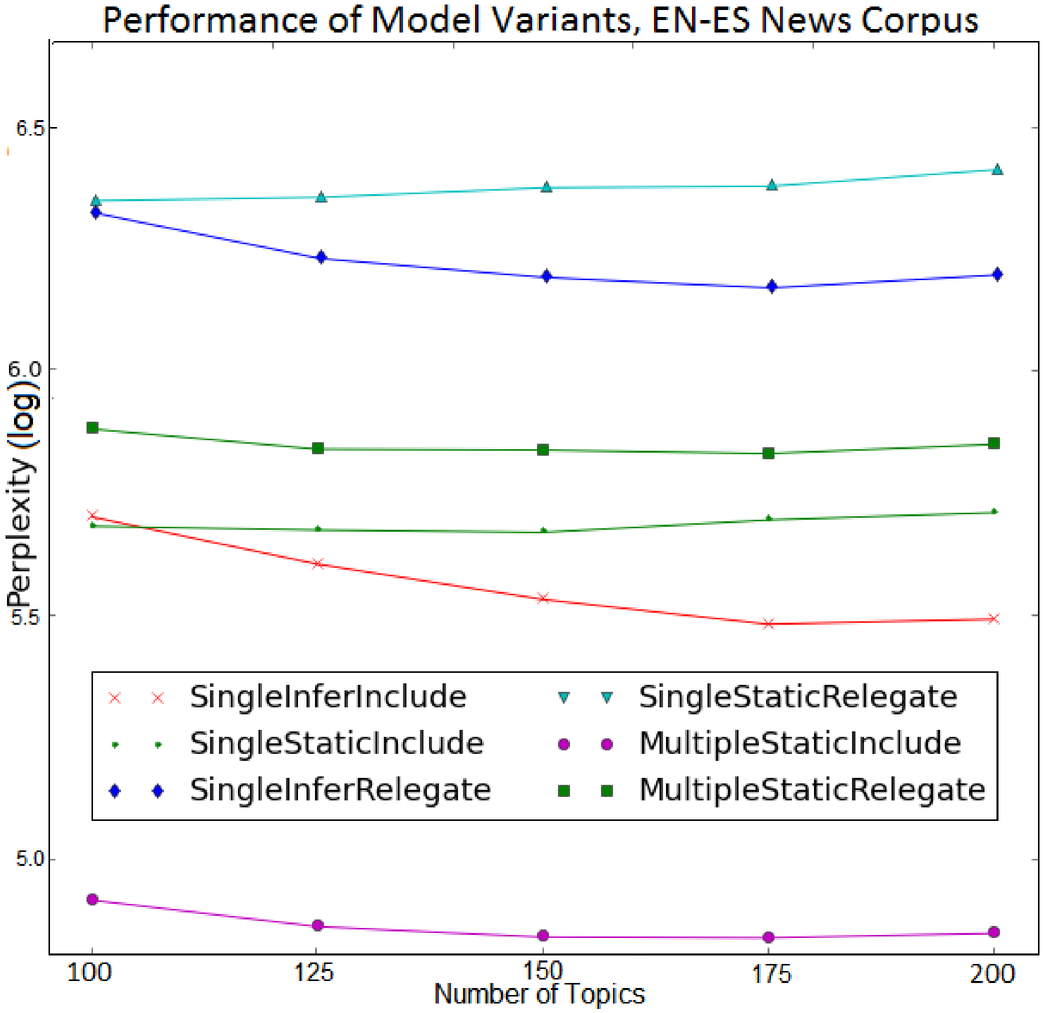

Figure 4.3: Perplexity of different model variants.

Twitter corpora, respectively. While some of the perspective words are neutral, mostly literal and occur in both English and Spanish (e.g. *balance* or *authorize*), many others represent metaphorical vocabulary (e.g. *saddle, gut, evaporate* in English, or *incendiar, sangrar, abatir* in Spanish) pointing at distinct models of conceptualization of the topic. When we applied the contrast detection method (described in §4.3.4) to these perspective words, it highlighted the differences in metaphorical perspectives, rather than the literal ones, as shown in Fig. 4.6 and 4.8. English speakers tend to discuss economic and financial processes using motion terms, such as "*slow, drive,*

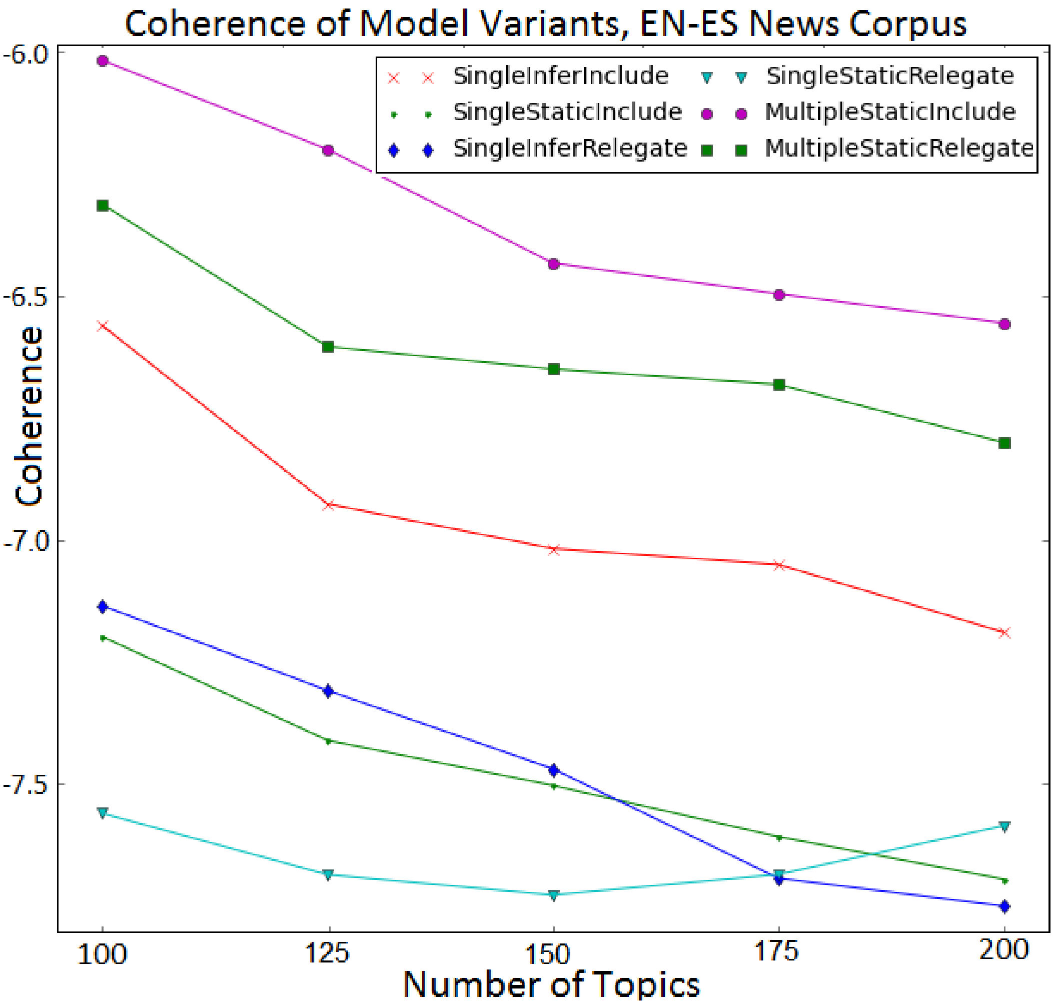

Figure 4.4: Coherence of different model variants.

boost or *sluggish*", or a related metaphor of *horse-riding*, e.g. "*rein* in debt", "*saddle* with debt", or even "*breed* money". In contrast, Spanish speakers tend to talk about the economy in terms of size rather than motion, using verbs such as *ampliar* or *disminuir*, and other metaphors, such as *sangrar* (to bleed) and *incendiar* (to light up). These examples demonstrate coherent conceptualization patterns that differ in the two languages. Interestingly, this difference manifested itself in both NEWS and TWITTER corpora and echoes the findings of a previous corpus-linguistic study of Charteris-Black and Ennis [CBE01], who manually analysed metaphors used in

Topic_EN budget debt deficit reduction spend balance cut increase limit downtown tax stress addition planet

Topic_ES presupuesto deficit deuda reduccion equilibrio disminucion gasto aumentacion tasa sacerdote

Perspective_EN balance default triple *rein* accumulate accrue *trim* incur *saddle slash* prioritize avert *gut* burden *evaporate* borrow pile *cap cut tackle*

Perspective_ES renegociar mejora etiquetado *desplomar recortar* endeudar *incendiar* destinar asignar autorizar aprobado ascender *sangrar* augurar *abatir*

Figure 4.5: Top perspectives in system output for the topic of *finance* in the NEWS corpus (metaphors in red italics).

Contrasts_EN: rein [in debt], saddle [with debt], cap [debt], breed [money], gut [budget], [debt] hit, tackle [debt], boost, slow, drive, sluggish [economy], spur

Contrasts_ES: sangrar [dinero], ampliar, disminuir [la economía], superar [la tasa], emitir [deuda]

Figure 4.6: Contrasts identified by the model in NEWS.

English and Spanish financial discourse and reported that motion and navigation metaphors that abound in English were rarely observed in Spanish.

For the majority of the topics we analysed the model revealed interesting cross-cultural differences. For instance, the Spanish corpora exhibited metaphors of *battle* when talking about *poverty* (with *poverty* seen as an enemy), while in the English corpus *poverty* was discussed more neutrally as a social problem that needs a practical solution. English-Russian NEWS experiments revealed a surprising difference with respect to the topic of *protests*. They suggested that while US media tend to use stronger metaphorical vocabulary, such as *clash, erupt* or *fire*, in Russian protests are discussed more neutrally. Generally, the NEWS corpora contained more abstract topics and richer information about conceptual structure and sentiment in all languages. Many of the topics discovered in TWITTER related to everyday concepts, such as *pets* or *concerts*, with fewer topics covering societal issues. Yet, a few TWITTER-specific contrasts could be observed: e.g., the *sports* topic tends to be discussed using *war* and *battle* vocabulary in Russian to a greater extent than in English.

Our models tend to identify two general kinds of differences: (1) cross-corpus differences representing world views of particular populations whom the corpora characterize (such differences exist both across and within languages, e.g. the metaphors used in the progressive *New York Times* would be different from the ones in the more conservative *Wall Street Journal*); and (2) deeply entrenched cross-linguistic differences, such as the *motion* versus *expansion* metaphors for the economy in English and Spanish. Such systematic cross-linguistic contrasts can be associated with contrastive behavioural patterns across the different linguistic communities [CB08, FMC$^+$11]. In both NEWS and TWITTER data, our model effectively identifies and summarises such contrasts simplifying the manual analysis of the data by highlighting linguistic trends that are indicative of the underlying conceptual differences. However, the conceptual differences are not straightforward to evaluate based on the surface vocabulary alone. In order to investigate this further, we conducted a behavioural experiment testing a subset of the contrasts discovered by our model.

4.5 Behavioural evaluation

We assessed the relevance of the contrasts through an experimental study with native English-speaking and native Spanish-speaking human subjects. We focused

Topic_EN economy growth rate percent bank economist interest reserve market policy

Topic_ES economía crecimiento tasa banco poltica mercado interés inflacin empleo economista

Perspective_EN economic financial grow global expect remain *cut boost* low *slow drive*

Perspective_ES económico mundial agregar financiero informal *pequeño* significar interno *bajar*

Figure 4.7: Top perspectives in system output for the *economy* topic in TWITTER (metaphors in red).

Contrasts_EN: slow [the economy], push [the economy], strong [economy], weak [economy], stable [economy], boost [the economy]

Contrasts_ES: caer [la economía], disminuir, superar [la economía], ampliar [el crecimiento]

Figure 4.8: Contrasts identified by the model in TWITTER.

on a linguistic difference in the metaphors used by English speakers versus Spanish speakers when discussing changes in a nation's economy. While English speakers tend to use metaphors involving both locative motion verbs (e.g. *slow*) as well as expansive/contractive motion verbs (e.g. *shrink*), Spanish speakers preferentially employ expansive/contractive motion verbs (e.g. *disminuir*) to describe changes in the economy. These differences could reflect linguistic artefacts (such as collocation frequencies) or could reflect entrenched conceptual differences. Our experiment addresses the question of whether such patterns of behaviour arise cross-linguistically in response to non-linguistic stimuli. If the linguistic differences are indicative of entrenched conceptual differences, then we expect to see responses to the non-linguistic stimuli that correspond to the usage differences in the two languages.

4.5.1 Experimental setup

We recruited 60 participants from one English-speaking country (the US) and 60 participants from three Spanish-speaking countries (Chile, Mexico, and Spain) using the CrowdFlower crowdsourcing platform. Participants first read a brief description of the experimental task, which introduced them to a fictional country in which economists are devising a simple but effective graphic for "representing change in [the] economy". They then completed a demographic questionnaire including information about their native language. Results from 9 US and 3 non-US participants were discarded for failure to meet the language requirement.

Participants navigated to a new page to complete the experimental task. Stimuli were presented in a 1200×700-pixel frame. The center of the frame contained a sphere with a 64-pixel diameter. For each trial, participants clicked on a button to activate an animation of the sphere which involved (1) a positive displacement (in rightward pixels) of 10% or 20%, or a negative displacement (in leftward pixels) of 10% or 20%;[3] and, (2) an expansion (in increased pixel diameter) of 10% or 20%, or a contraction (in decreased pixel diameter) of 10% or 20%.[4]

Participants saw each of the resulting conditions 3 times. The displacement and size conditions were drawn from a random permutation of 16 conditions using a Fisher-Yates shuffle [FY63]. Crucially, half of the stimuli contained conflicts of information with respect to the size and displacement metaphors for economic change (e.g. the sphere could both grow and move to the left). Overall we expected the

[3] The use of leftward/rightward horizontal displacement to represent decreases/increases in magnitude is supported by research in numerical cognition showing that people associate smaller magnitudes with the left side of space and larger magnitudes with the right side [Deh92, FBGd95].

[4] A demonstration of the English experimental interface can be accessed at http://goo.gl/W3YVfC. The Spanish interface is identical, but for a direct translation of the guidelines provided by a native Spanish/fluent English speaker.

Spanish speakers' responses to be more closely associated with changes in diameter due to the presence and salience of the size metaphor, and the English speakers' responses to be influenced by both conditions. We expected these differences to be most prominent in the conflicting trials, which force English speakers (unlike Spanish speakers) to choose between two available metaphors. We focus on these conflicting trials in our analysis and discussion of the results.

4.5.2 Results

In trials in which stimuli moving rightward were simultaneously contracting, English speakers responded that the economy improved 66% of the time, whereas Spanish speakers judged the economy to have improved 43% of the time. In trials in which stimuli moving leftward were simultaneously expanding, English speakers judged the economy to have improved 34% of the time, and Spanish speakers responded that the economy improved 55% of the time. The results are illustrated in Figure 4.9.

These results indicate three effects: (1) English speakers exhibit a pronounced bias for using horizontal displacement rather than expansion/contraction during the decision-making process; (2) Spanish speakers are more biased toward expansion/contraction in formulating a decision; and, (3) across the two languages the responses show contrasting patterns. The results support our expectation on the relevance of different metaphors when reasoning about the economy by the English and Spanish speakers.

To examine the significance of these effects, we fit a binary logit mixed effects model[5] to the data. The full analysis modeled judgment with native language,

[5] See Fox and Weisberg [FW11] for a discussion of such models including application of the Type II Wald test.

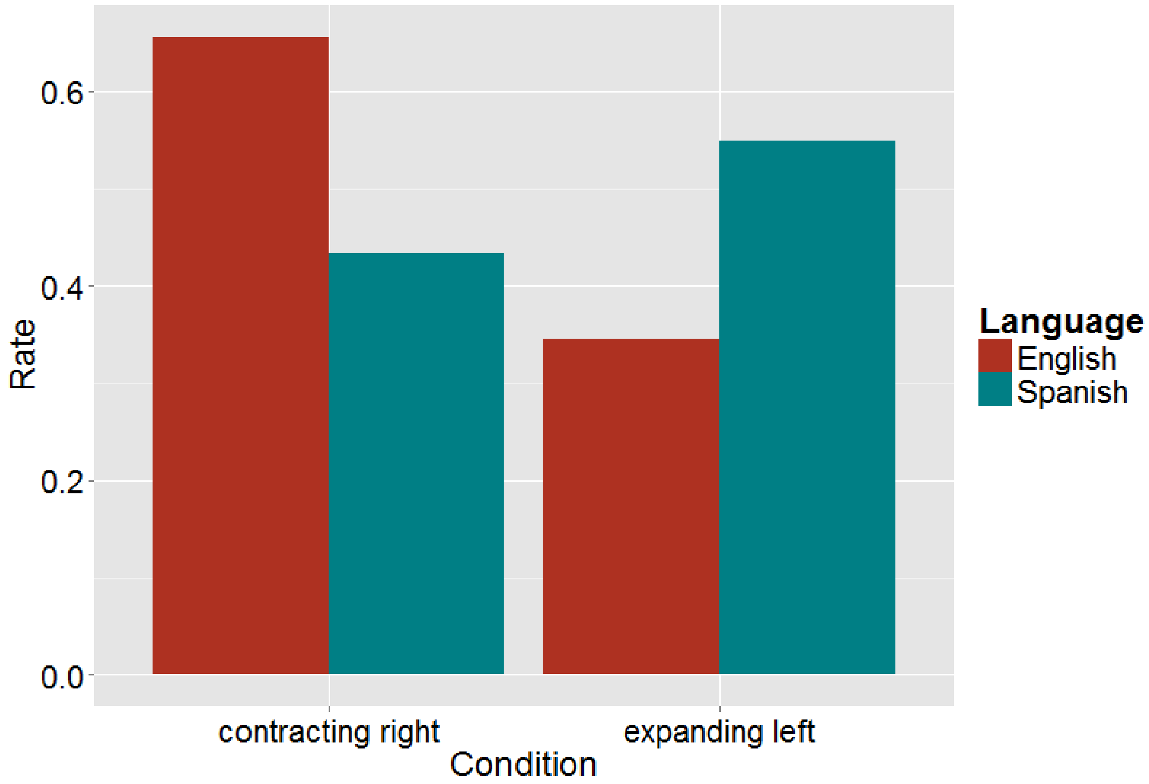

Figure 4.9: "Economy Improved" response rate in conflicting stimulus conditions.

displacement, and size as fully crossed fixed effects and participant as a random effect. This analysis confirmed that native language was associated with judgments about economic change. In particular, it indicated that changes in size affected English speakers' judgments and Spanish speakers' judgments differently ($p < 0.001$), with an increase in size increasing the odds ($e^{\beta} = 2.5$) of a judgment of Improved by Spanish speakers and decreasing the odds ($e^{\beta} = 0.44$) of a judgment of Improved by English speakers. A Type II Wald test revealed the interaction between language and size to be highly statistically significant ($\chi^2(1) < 0.001$).

In summary, the patterns we see in the behavioural data are consistent with the patterns uncovered in the output of our model. While much territory remains to be investigated to delimit the nature of this relationship, our results represent a first step toward establishing an association between information mined from large textual data collections and information observed through behavioural responses on a human scale.

4.6 Conclusion

We presented the first model that detects common topics from multilingual, non-parallel data and automatically uncovers differences in perspectives on these topics across linguistic communities. Our data analysis and behavioural evaluation offer evidence of a symbiotic relationship between ecologically sound corpus experiments and scientifically controlled human subject experiments, paving the way for the use of large-scale text mining to inform cognitive linguistics and psychology research.

We believe that our model represents a good foundation for future projects in this area. A promising area for further work is in developing better methods for identifying contrasts in perspective terms. This could perhaps involve modifying the generative process for perspective terms or incorporating syntactic dependency information. It would also be interesting to investigate the effect of dictionary quality and corpus size on the relative performance of Static and Infer variants. Finally, we note that the model can be applied to identify contrastive perspectives in monolingual as well as multilingual data, providing a general tool for the analysis of subtle, yet important, cross-population differences.

Chapter 4, in part, is a reprint of the material as it appears in Gutierrez, E.D.; Shutova, Ekaterina; Lichtenstein, Patricia; de Melo, Gerard; Gilardi, Luca. "Detecting cross-cultural differences using a probabilistic topic model", Transactions of the Association for Computational Linguistics, vol. 4, 2016. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Conclusion

The work in this thesis highlights some ways that cognitive semantics and computational semantics can benefit from each other. As we have shown However, it skirts around many open questions about how a more complete fusion of these two fields can be achieved. It is unclear, for instance, to what extent computational and statistical models can adjudicate on open scientific questions in cognitive semantics. What does the success of statistical modeling in computational semantics say about the role of statistics versus the role of embodiment in human semantic learning and processing? Conversely, the current trend in computational semantics toward nearly language-agnostic language models (often driven by deep learning—cf. [JVS$^+$16]) raises doubts about what role, if any, cognitive semantic theories, or any theory of language in general, will play in the future of natural language processing. Still, we conjecture that as deep learning matures, researchers will find a productive role for linguistic theory in such language-agnostic language models.

Appendix A

Alternative Approaches to Characterizing Systematicity

This appendix contains some approaches to characterizing the non-arbitrary systematicity in the lexicon that did not make it into chapter 3

A.1 Linear Regression: The Feature-based Model

The orthographic features that are most of interest to us involve the presence or absence of position-dependent orthographic clusters (e.g., word-initial *gl-*; word-final *-ign*), so we attempted to employ several linear regression techniques that detect relationships between binary predictors and continuous responses. Interestingly, these regression methods described below have found broad application in genomics, especially for the task of associating genotypes with phenotypes in large populations [LTB11]. Indeed, this task shares many similarities with the task of finding semantically predictive orthographic features: like our phonetic features, the presence or absence of each polymorphism in the genotype can be modeled as a binary variable (assuming complete dominance of alleles); the most effects of genotype on phenotype are taken to be complex and noisy (just like the effect of orthographic features on our semantic vectors); and the task is to identify a relatively small number of candidate genotypes with an effect from a very large overall number of genotypes.

A.1.1 Regularized Multivariate Regression.

For our regularized linear regression models, our orthographic features consisted of all two- and three-letter combinations that occurred at the beginning and end of words in our lexicon. We regressed the values of our semantic vectors on these responses using multi-task elastic net regression [ZH05]. Elastic net regression is a regularization framework that combines L1 and L2 regularization penalties. Unlike L1 regression, elastic net regularization can deal with collinear predictor variables without shrinking all but one of their coefficients to zero. Unlike L2 regression, elastic net regularization produces a sparse solution that shrinks irrelevant variables to zero. The added cost to elastic net regression comes in having to fine-tune two parameters, the ratio of L1 penalty to L2 penalty, and the overall magnitude of the combined L1/L2 penalty. We used a grid search under 10-fold cross-validation to perform this fine-tuning, as implemented in the Scikit-Learn package for Python.

The greatest drawback of multivariate linear regression is that it encodes strong parametric assumptions. Each dimension of the predictor variable is assumed to have an additive effect on the overall prediction.

Null Results with Multivariate Regularized Regression. Our multivariate regularized regression did not identify any statistically reliable word-initial or word-final feature. This may be due to the strong parametric assumptions of linear regression where each feature has an independent additive effect, and due to the low power of our sample size, which is limited by the number of monosyllabic monomorphemic words.

A.1.2 Mass Univariate Regression.

Another technique common in genetic association studies is mass univariate linear regression. With the mass univariate analysis we hoped to evaluate whether orthographic features exhibit a global, if weak, effect on the values of the semantic vectors. Since our features are binary random variables, this technique involves simply finding the means of the words with each feature and without each feature.

Attempt to Measure Distribution of Systeamaticity across Features. On the assumption that there is no relationship between the orthographic features and the semantic values, the p-values of the mass univariate regressions should roughly adhere to a $\text{Unif}(0, 1)$ distribution. On the other hand, if phonetic features do tend to predict meaning, then the distribution of the p-values should be more skewed toward zero. Define W as the sum of the negative logarithms of the p-values of each of the D features. Then W is distributed chi-squared with N degrees of freedom, so the overall p-value for the hypothesis that features tend to predict meaning is $p = 1 - \chi^2_N(W)$ [War13].

In the case of binary predictors and continuous responses, the problem of finding p-values for the coefficients of univariate regression reduces to a t-test. p-values are readily calculated for the t-test. We first performed a Welch's t-test for independent samples with unequal variance and compiled the p-values. However, the reliability of p-values for Welch's t-test relies somewhat upon the assumption of normally distributed predictor values within each grouping; our data do not hold to this standard, so we applied a transform to our data so that they would closer match a normal distribution. We applied Welch's t-test to each of the 50 principal components of our vector-space model independently. We found that at the significance level of $p < .05$, Bonferroni-corrected for 50 comparisons (i.e., uncorrected $p < .001$), the null hypothesis of no relationship between the orthographic features and the semantic values was rejected for 11 of the 50 semantic dimensions. In other words, for 11 of the 50 dimensions, a significant portion of the p-values were concentrated near or at significance. Looking to eliminate any concerns about whether our transform actually normalized our data, we repeated our analysis using the Mann-Whitney rank-sum test. The Mann-Whitney test is a non-parametric alternative to the t-test that does not rely on any distributional assumptions about the response variables [HM98]; instead the response values are transformed into ordinal ranks. Using the rank-sum test, the null hypothesis of no relationship was rejected for 10 of the 50 dimensions at the Bonferroni-corrected significance level of $p < .05$.

Bibliography

[AFS13] Ekaterina Abramova, Raquel Fernández, and Federico Sangati. Automatic labeling of phonesthemic senses. In *Proeedings of the 35th Annual Conference of the Cognitive Science Society*, volume 35. Cognitive Science Society, 2013.

[AJDDR12] Amjad Abu-Jbara, Mona Diab, Pradeep Dasigi, and Dragomir Radev. Subgroup detection in ideological discussions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 399–409, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[ARW11] Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. OpinioNetIt: Understanding the Opinions-People network for politically controversial topics. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2481–2484, New York, NY, USA, 2011. ACM.

[AX10] Amr Ahmed and Eric P. Xing. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1140–1150, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[BBFZ09] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

[BBZ14] Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9, 2014.

[BDB08] R. Harald Baayen, Douglas J. Davidson, and Douglas M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, 2008.

[Ber04] Benjamin K. Bergen. The psychological reality of phonaesthemes. *Language*, pages 290–311, 2004.

[BGB09] Jordan Boyd-Graber and David M. Blei. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09)*, pages 75–82. Arlington, VA, USA: AUAI Press, 2009.

[BGR10] Jordan Boyd-Graber and Philip Resnik. Holistic sentiment analysis across languages: multilingual supervised latent Dirichlet allocation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 45–55, 2010.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[BH01] Todd M. Bailey and Ulrike Hahn. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44(4):568–591, 2001.

[BHS12] Aurélien Bellet, Amaury Habrard, and Marc Sebban. Good edit similarity learning by loss minimization. *Machine Learning*, pages 5–35, 2012.

[BL04] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 1–4, 2004.

[BLM09] Steven Bethard, Vicky Tzuyin Lai, and James H. Martin. Topic model analysis of metaphor frequency for psycholinguistic stimuli. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 9–16. Association for Computational Linguistics, 2009.

[BLNZ95] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[BLST13] Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278, 2013.

[BNC07] BNC Consortium. British National Corpus, Version 3 BNC XML edition, 2007.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[BPG96] R Harald Baayen, Richard Piepenbrock, and Léon Gulikers. CELEX2 (CD-ROM), 1996.

[BS06] Julia Birke and Anoop Sarkar. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, 2006.

[BVCM12] Gemma Boleda, Eva Maria Vecchi, Miquel Cornudella, and Louise McNally. First-order vs. higher-order modification in distributional semantics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1223–1233. Association for Computational Linguistics, 2012.

[BZ10] Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics, 2010.

[Cam03] Lynne Cameron. *Metaphor in Educational Discourse.* A&C Black, London, 2003.

[Cas14] Daniel Casasanto. Development of metaphorical thinking: The role of language. In Mike Borkent, Barbara Dancygier, and Jennifer Hinnell, editors, *Language and the Creative Mind*, pages 3–18. CSLI Publications, Stanford, 2014.

[CB08] Daniel Casasanto and Lera Boroditsky. Time in the mind: Using space to think about time. *Cognition*, 106(2):579–593, 2008.

[CBE01] Jonathan Charteris-Black and Timothy Ennis. A comparative study of metaphor in Spanish and English financial reporting. *English for Specific Purposes*, 20:249–266, 2001.

[Coh60] Jacob Cohen. A coefficient of agreement for nominal scales. educational and psychosocial measurement, 1960.

[Cro51] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[CSC10] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. In *Linguistic Analysis (Lambek Festschrift)*, pages 345–384, 2010.

[DDL$^+$90] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407, 1990.

[Deh92] Stanislas Dehaene. Varieties of numerical abilities. *Cognition*, 44:1–42, 1992.

[DPB13] Georgiana Dinu, Nghia The Pham, and Marco Baroni. General estimation and evaluation of compositional distributional semantic models. In *Proceedings of the ACL 2013 Workshop on Continuous Vector Space Models and their Compositionality (CVSC 2013)*, pages 50–58, East Stroudsburg, Pennsylvania, 2013. ACL.

[dS16] Ferdinand de Saussure. *Course in General Linguistics*. McGraw-Hill, New York, 1916.

[Dun13a] Jonathan Dunn. Evaluating the premises and results of four metaphor identification systems. In *Computational Linguistics and Intelligent Text Processing*, pages 471–486. Springer, 2013.

[Dun13b] Jonathan Dunn. What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 1–10, 2013.

[EP08] Katrin Erk and Sebastian Padó. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906. Association for Computational Linguistics, 2008.

[EP10] Katrin Erk and Sebastian Padó. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97. Association for Computational Linguistics, 2010.

[FBGd95] Wim Fias, Marc Brysbaert, Frank Geypens, and Géry d'Ydewalle. The importance of magnitude information in numerical processing: evidence from the SNARC effect. *Mathematical Cognition*, 2(1):95–110, 1995.

[FCE69] Joseph L. Fleiss, Jacob Cohen, and B.S. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323, 1969.

[Fir30] John R. Firth. *Speech*. Benn's Sixpenny Library, London, 1930.

[Fis32] R.A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, London, 1932.

[FMC$^+$11] Orly Fuhrman, Kelly McCormick, Eva Chen, Heidi Jiang, Dingfang Shu, Shuaimei Mao, and Lera Boroditsky. How linguistic and cultural forces shape conceptions of time: English and Mandarin time in 3D. *Cognitive Science*, 35:1305–1328, 2011.

[FPC15] Daniel Fried, Tamara Polajnar, and Stephen Clark. Low-rank tensors for verbs in compositional distributional semantics. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*, Beijing, 2015.

[FRMQ07] Anthony Fader, Dragomir Radev, Burt L. Monroe, and Kevin M. Quinn. MavenRank: Identifying influential members of the US senate using lexical centrality. In *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 658–666, 2007.

[FSSY12] Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*, pages 63–72, New York, 2012. New York: ACM.

[FW11] John Fox and Sanford Weisberg. *An R Companion to Applied Regression.* SAGE Publications, CA: Los Angeles, 2011.

[FY63] Ronald A. Fisher and Frank Yates. *Statistical Tables for Biological, Agricultural and Medical Research.* Oliver and Boyd, Edinburgh, 1963.

[FZBB08] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. Introducing and evaluating UKWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54, 2008.

[Gas04] Michael Gasser. The origins of arbitrariness in language. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, volume 26, pages 4–7, 2004.

[GB11] Sean M. Gerrish and David M. Blei. Predicting legislative roll calls from text. In *Proceedings of ICML*, 2011.

[GBNC06] Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. Catching metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, pages 41–48, New York, 2006. Association for Computational Linguistics.

[Geo00] Darren George. *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update (4th ed.).* Allyn & Bacon, London, 2000.

[GH05] Joseph A. Goguen and D. Fox Harrell. 7 information visualisation and semiotic morphisms. *Studies in Multidisciplinarity*, 2:83–97, 2005.

[GH10] Joseph A. Goguen and D. Fox Harrell. Style: A computational and conceptual blending-based approach. In *The Structure of Style*, pages 291–316. Springer, New York, 2010.

[GKCM03] David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English Gigaword. *Linguistic Data Consortium, Philadelphia*, 2003.

[Gog99] Joseph Goguen. An introduction to algebraic semiotics, with application to user interface design. In *Computation for metaphors, analogy, and agents*, pages 242–291. Springer, 1999.

[GQS$^+$13] Swapna Gottipati, Minghui Qiu, Yanchuan Sim, Jing Jiang, and Noah A. Smith. Learning topics and positions from Debatepedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1858–1868, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[GS13] Justin Grimmer and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, January 2013.

[Gue10] Emiliano Guevara. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics, 2010.

[GY09] Nikesh Garera and David Yarowsky. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 710–718, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[Har54] Zellig S. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[HGS$^+$13] Ilana Heintz, Ryan Gabbard, Mahesh Srinivasan, David Barner, Donald S Black, Marjorie Freedman, and Ralph Weischedel. Automatic extraction of linguistic metaphor with lda topic modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66, 2013.

[HJM08] David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language processing*, pages 363–371. Association for Computational Linguistics, 2008.

[HL13] Sterling Hutchinson and Max Louwerse. Language statistics and individual differences in processing primary metaphors. *Cognitive Linguistics*, 24(4):667–687, 2013.

[HL14] Sterling Hutchinson and Max M. Louwerse. Language statistics explain the spatial–numerical association of response codes. *Psychonomic Bulletin & Review*, 21(2):470–478, 2014.

[HLBKK08] Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, ACL-'08:HLT, pages 771–779, Columbus, Ohio, USA, 2008.

[HM98] T. P. Hettmansperger and J. W. McKean. Robust nonparametric statistical methods. In *Kendall's Library of Statistics 5*, pages xiv–467. Edward Arnold, London, 1998.

[Hoc60] Charles F. Hockett. The origin of speech. *Scientific American*, 203:88–96, 1960.

[HSJ$^+$13] Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57, 2013.

[Hut98] Sharon Suzanne Hutchins. *The psychological reality, variability, and compositionality of English phonesthemes.* PhD thesis, Emory University, Atlanta, 1998.

[JD10] Jagadeesh Jagarlamudi and Hal Daumé III. Extracting multilingual topics from unaligned comparable corpora. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, and Suzanne Little, editors, *Proceedings of the 32nd European Conference on Advances in Information Retrieval (ECIR'2010)*, pages 444–456. Springer-Verlag, Berlin, 2010.

[JKPT07] Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. "I know what you did last summer": Query logs and user privacy. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 909–914, New York, NY, USA, 2007. ACM.

[JV87] Roy Jonker and Anton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

[JVS$^+$16] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[KÖ4] Zoltán Kövecses. Introduction: Cultural variation in metaphor. *European Journal of English Studies*, 8:263–274, 2004.

[KF91] Werner Kuhn and Andrew U Frank. A formalization of metaphors and image-schemas in user interfaces. In *Cognitive and linguistic aspects of geographic space*, pages 419–434. Springer, 1991.

[Kli00] Richard R. Klink. Creating brand names with meaning: The use of sound symbolism. *Marketing Letters*, 11(1):5–20, 2000.

[KS$^+$13] Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, et al. Prior disambiguation of word tensors for constructing sentence vectors. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1590–1601, 2013.

[KSP13] Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. Separating disambiguation from composition in distributional semantics. In *Proceedings of the 2013 Conference on Computational Natural Language Learning*, pages 114–123, 2013.

[KZ07] Saisuresh Krishnakumaran and Xiaojin Zhu. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational approaches to Figurative Language*, pages 13–20. Association for Computational Linguistics, 2007.

[Lak89] George Lakoff. Some empirical results about the nature of concepts. *Mind & Language*, 4(1-2):103–129, 1989.

[Lak93] George Lakoff. The contemporary theory of metaphor. In Andrew Ortony, editor, *Metaphor and Thought*. Cambridge University Press, Cambridge, 1993.

[Lam99] Joachim Lambek. Type grammar revisited. In *Logical aspects of computational linguistics*, pages 1–27. Springer, Berlin, 1999.

[LBA95] Kevin Lund, Curt Burgess, and Ruth Ann Atchley. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th annual conference of the Cognitive Science Society*, volume 17, pages 660–665, 1995.

[LBD14] Jiming Li, Marco Baroni, and Georgiana Dinu. Improving the lexical function composition model with pathwise optimized elastic-net regression. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 434–442, 2014.

[LD97] Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211, 1997.

[Lev66] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710, 1966.

[LJ81] George Lakoff and Mark Johnson. *Metaphors we live by.* University of Chicago Press, Chicago, 1981.

[LRM99] Willem J.M. Levelt, Ardi Roelofs, and Antje S. Meyer. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(01):1–38, 1999.

[LRS10] Linlin Li, Benjamin Roth, and Caroline Sporleder. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147. Association for Computational Linguistics, 2010.

[LS09] Linlin Li and Caroline Sporleder. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 315–323. Association for Computational Linguistics, 2009.

[LS10] Linlin Li and Caroline Sporleder. 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. Association for Computational Linguistics, 2010.

[LTB11] Po-Ru Loh, George Tucker, and Bonnie Berger. Phenotype prediction using regularized regression on genetic data in the DREAM5 systems genetics B challenge. *PloS one*, 6(12):e29095–e29095, 2011.

[LW12] George Lakoff and Elisabeth Wehling. *The Little Blue Book: The Essential Guide to Thinking and Talking Democratic.* Free Press, New York, 2012.

[Mag98] Margaret Magnus. *Whats in a Word? Evidence for Phonosemantics.* PhD thesis, University of Trondheim, Trondheim, Norway, 1998.

[Man67] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220, 1967.

[MBHT13] Michael Mohler, David Bracewell, David Hinote, and Marc Tomlinson. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35, 2013.

[MCQ08] Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008.

[ML08] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *ACL-08: HLT*, pages 236–244, 2008.

[MLC14] Padraic Monaghan, Gary Lupyan, and Morten H Christiansen. The systematicity of the sign: Modeling activation of semantic attributes from nonwords. In P. Bello, M. Guarini, M. McShane, and B. Scassellati, editors, *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pages 2741–2746, Austin, TX, 2014. Cognitive Science Society.

[Mon70] Richard Montague. English as a formal language. In B Visentini and et al, editors, *Linguaggi nella Società e nella Tecnica*. Edizioni di Comunitá, Milan, 1970.

[MS09] Iain Murray and Ruslan R. Salakhutdinov. Evaluating probabilities under high-dimensional latent variable models. In *Advances in Neural Information Processing Systems*, pages 1137–1144, 2009.

[MSC$^+$13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[MSCK14] Padraic Monaghan, Richard C. Shillcock, Morten H. Christiansen, and Simon Kirby. How arbitrary is language? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1651), 2014.

[MWN$^+$09] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, pages 880–889. Association for Computational Linguistics, 2009.

[MWT$^+$11] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.

[NAC$^+$13] Yair Neuman, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. Metaphor identification in large texts corpora. *PLoS ONE*, 8:e62343, 2013.

[Nad64] Elizbar A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[NLGB10] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2010.

[Nos86] Robert M. Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39, 1986.

[Oha84] John J. Ohala. An ethological perspective on common cross-language utilization of f0 of voice. *Phonetica*, 41(1):1–16, 1984.

[OS08] Katya Otis and Eyal Sagi. Phonaesthemes: A corpus-based analysis. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 65–70, 2008.

[OS10] Diarmuid Ó Séaghdha. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444, Uppsala, Sweden, 2010. Association for Computational Linguistics.

[Par94] Barbara H. Partee. Lexical semantics and compositionality. In Lila Gleitman and Mark Liberman, editors, *Invitation to Cognitive Science 2nd Edition, Part I: Language*. MIT Press, Cambridge, Mass., USA, 1994.

[PECX10] Kriti Puniyani, Jacob Eisenstein, Shay Cohen, and Eric P. Xing. Social links from latent topics in microblogs. In *Proceedings of the NAACL/HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 19–20. Association for Computational Linguistics, 2010.

[PG09] Michael Paul and Roxana Girju. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1408–1417, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[PGK$^+$11] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English Gigaword Fifth Edition, 2011.

[PL02] Patrick Pantel and Dekang Lin. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619. ACM, 2002.

[PP11] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, Republicans and Starbucks afficionados: user classification in Twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 430–438, 2011.

[QJ13] Minghui Qiu and Jing Jiang. A latent variable model for viewpoint discovery from threaded forum posts. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1031–1040, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[REE10] Alan Ritter, Mausam Etzioni, and Oren Etzioni. A latent Dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434. Association for Computational Linguistics, 2010.

[ŘS10] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. urlhttp://is.muni.cz/publication/884893/en.

[SBT$^+$13] Tomek Strzalkowski, George A. Broadwell, Sarah Taylor, Laurie Feldman, Boris Yamrom, Samira Shaikh, Ting Liu, Kit Cho, Umit Boz, Ignacio Cases, and Kyle Elliot. Robust extraction of metaphors from novel data. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 67–76, Atlanta, Georgia, 2013. Association for Computational Linguistics.

[Sch98] Hinrich Schütze. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123, 1998.

[SDH$^+$10] Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing, Amsterdam/Philadelphia, 2010.

[SHMN12] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics, 2012.

[Shu15] Ekatrina Shutova. Design and evaluation of metaphor processing systems. volume Forthcoming, 2015.

[SKAB12] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961, Jeju Island, Korea, 2012.

[SKMB01] Richard Shillcock, Simon Kirby, Scott McDonald, and Chris Brew. Filled pauses and their status in the mental lexicon. In *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*, 2001.

[SL09] Caroline Sporleder and Linlin Li. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–762. Association for Computational Linguistics, 2009.

[SO08] Eyal Sagi and Katya Otis. Semantic glimmers: Phonaesthemes facilitate access to sentence meaning. In *9th Conference on Conceptual Structure, Discourse, & Language (CSDL9)*, 2008.

[SPH$^+$11] Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics, 2011.

[Spi14] David I. Spivak. *Category Theory for the Sciences*. MIT Press, Cambridge, Mass., USA, 2014.

[SSK10] Ekaterina Shutova, Lin Sun, and Anna Korhonen. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010. Association for Computational Linguistics, 2010.

[Ste02] Anatol Stefanowitsch. Sound symbolism in a usage-driven model. Unpublished manuscript, Rice University, Houston, Texas, USA, 2002.

[Tam06] Monica Tamariz. *Exploring the adaptive structure of the mental lexicon*. PhD thesis, University of Edinburgh, Edinburgh, 2006.

[Tam08] Monica Tamariz. Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, 3(2):259–278, 2008.

[TB11] Paul H. Thibodeau and Lera Boroditsky. Metaphors we think with: The role of metaphor in reasoning. *PLoS ONE*, 6(2):e16782, 2011.

[TBG$^+$14] Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. Metaphor detection with cross-lingual model transfer. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014.

[TDSM13] Masashi Tsubaki, Kevin Duh, Masashi Shimbo, and Yuji Matsumoto. Modeling and learning semantic co-compositionality through prototype projections and neural networks. In *The 2013 Conference on Empirical Methods in Natural Language Processing*, pages 130–140, 2013.

[TMG13] Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, Atlanta, Georgia, 2013. Association for Computational Linguistics.

[TNAC11] Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, EMNLP '11, pages 680–690, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[TP10] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

[Tur13] Peter D. Turney. Distributional semantics beyond words: supervised learning of analogy and paraphrase. *Transactions of the Association for Computational Linguistics (TACL)*, 1:353–366, 2013.

[Uts06] Akira Utsumi. Computational exploration of metaphor comprehension processes. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society (CogSci2006)*, pages 2281–2286, 2006.

[War13] Robert L. Wardrop. Statistics 371, blended: Course notes, 2013.

[WMM15] Bodo Winter, Tyler Marghetis, and Teenie Matlock. Of magnitudes and metaphors: Explaining cognitive interactions between space, time, and number. *Cortex*, 64:209–224, 2015.

[WT07] Killian Q. Weinberger and Gerald Tesauro. Metric learning for kernel regression. In *Eleventh International Conference on Artificial Intelligence and Statistics*, pages 608–615, 2007.

[WWB$^+$04] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Computational Linguistics*, 30(3):277–308, September 2004.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.