A Geometric Method for Context Sensitive Distributional Semantics

by

Stephen McGregor

A thesis submitted to Queen Mary University of London for the degree of Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary, University of London
United Kingdom

September 2017

My university requires me to make the following statement:

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

I add this:

I hereby grant permission to anyone to do anything they so please with the text of this thesis and any information they derive from it or meaning they find in it, with or without acknowledgement of the source.

Abstract

This thesis describes a novel methodology, grounded in the distributional semantic paradigm, for building context sensitive models of word meaning, affording an empirical exploration of the relationship between words and concepts. Anchored in theoretical linguistic insight regarding the contextually specified nature of lexical semantics, the work presented here explores a range of techniques for the selection of subspaces of word co-occurrence dimensions based on a statistical analysis of input terms as observed within large-scale textual corpora. The relationships between word-vectors that emerge in the projected subspaces can be analysed in terms of a mapping between their geometric features and their semantic properties. The power of this modelling technique is its ability to generate ad hoc semantic relationships in response to an extemporaneous linguistic or conceptual situation.

The product of this approach is a generalisable computational linguistic methodology, capable of taking input in various forms, including word groupings and sentential context, and dynamically generating output from a broad base model of word co-occurrence data. To demonstrate the versatility of the method, this thesis will present competitive empirical results on a range of established natural language tasks including word similarity and relatedness, metaphor and metonymy detection, and analogy completion. A range of techniques will be applied in order to explore the ways in which different aspects of projected geometries can be mapped to different semantic relationships, allowing for the discovery of a range of lexical and conceptual properties for any given input and providing a basis for an empirical exploration of distinctions between the semantic phenomena under analysis. The case made here is that the flexibility of these models and their ability to extend output to evaluations of unattested linguistic relationships constitutes the groundwork for a method for the extrapolation of dynamic conceptual relationships from large-scale textual corpora.

This method is presented as a complement and a counterpoint to established distributional methods for generating lexically productive word-vectors. Where contemporary vector space models of distributional semantics have almost universally involved either the factorisation of co-occurrence matrices or the incremental learning of abstract representations using neural networks, the approach described in this thesis preserves the connection between the individual dimensions of word-vectors and statistics pertaining to observations in a textual corpus. The hypothesis tested here is that the maintenance

of actual, interpretable information about underlying linguistic data allows for the contextual selection of non-normalised subspaces with more nuanced geometric features. In addition to presenting competitive results for various computational linguistic targets, the thesis will suggest that the transparency of its representations indicates scope for the application of this model to various real-world problems where an interpretable relationship betweendata and output is highly desirable. This, finally, demonstrates a way towards the productive application of the theory and philosophy of language to computational linguistic practice.

Glossary

base space A high dimensional, sparse vector space of word-vectors, delineated in terms of dimensions of co-occurrence statistics.

context The situation – environmental, cognitive, perceptual, linguistic, and otherwise – in which an agent finds itself and applies language to meaning.

contextual input A set of words characteristic of a conceptual category or semantic relationship used to generate a subspace for the modelling of semantic phenomena.

dimension selection The process of contextually choosing a subset of dimensions in order to project a subspace from a base space.

co-occurrence The observation of one word in proximity to another in a corpus.

co-occurrence statistic A measure of the tendency for one word to be observed in proximity to another across a corpus.

co-occurrence window The boundary defining the proximity within which two words are considered to be co-occurring, typically a distance in terms of words within a sentence.

methodology The process of building base spaces from observations of co-occurrences within a corpus and contextually projecting subspaces through dimension selection.

model An application of methodology to a particular linguistic task or experiment, sometimes including task specific statistical analysis techniques.

subspace A context specific lower-dimensional projection from a base space, effectively mapping semantic relationships to a context by way of the geometric relationships between word-vectors.

word-vector A high-dimensional geometrically situated semantic representation of a word, constructed as an array of co-occurrence statistics.

Table of Contents

List of Figures

List of Tables

Chapter 3

Metaphor and Coercion

In this chapter, I will extend the empirical work on exploring the application of my context sensitive distributional semantic models to two semantic phenomena which involve the application of words in situations where their meanings are in some sense conceptually altered: *metaphor* and *semantic type coercion*. The precise definitions of these terms, which are not without nuance, was explored in Chapter **??** and will be reintroduced in subsequent sections. As an overview, the distinguishing characteristic of these phenomena is that they involve cases where what might be thought of as the stable, encyclopedic understanding of some word sense – a *dictionary definition* of a word, so to speak – is in some way appropriated or subverted in order to, among other things, transfer information via the attributional conduits connecting figurative source to literal target.

My hypothesis is that, because figurative language always involves the contextual specification of word meaning, context sensitive geometries of lexical representations should provide an appropriate framework for identifying when this type of semantic phenomenon is in effect. **?** demonstrates empirically that metaphor interpretation is, when a metaphor is presented to a subject out of context, an ambiguous exercise, and, to the extent that interpretations of de-contextualised metaphors can be predicted, the predicting factors are themselves culturally relative. Along similar lines, **?** propose that metaphor production involves the contextual alignment of overlapping semantic frames, and that this alignment likewise imports structure associated with one frame into the domain of another, evident in, for instance, the additional transposition of syntactic constraints from source to target. From a cognitive perspective, this coordinates a contextual theory of metaphor with the work on conceptual frames from Barsalou (1992,9) discussed at the end of the previous chapter in the context of judgements of semantic similarity. From a modelling perspective, this suggests that a methodology for projecting semantic

spaces where context specific perspectives can reveal *ad hoc* perspectives on semantic relationships should be a productive approach to identifying figurative language.

The idea that metaphor and metonymy are both instances of "a connection between two things where one term is substituted for another," (**?**, p. 260) will quickly call to mind the premise of distributional semantics: if the motivation for building vector space models of word co-occurrence statistics is that related words have similar co-occurrence tendencies, then figurative language might be construed as a special case in which unrelated or at least conceptually divergent words are likewise found in similar sentential situations. The question, then, is whether statistical characteristics of the particular co-occurrences profiles selected by words with different meanings are predictive of figurativeness. A naive hypothesis might be that word combinations that are figurative should simply be further apart in a semantic space than word combination that are literal. If related words have similar co-occurrence profiles, then maybe unrelated words, for instance words with different conceptual entailments, should have less similar co-occurrence profiles. This conjecture, however, is belied first of all by the fact that, in the type of corpus containing a broad range of examples of language use necessary for building distributional semantic models, figurative language will already be built into the data (and at the end of this chapter I will argue, in line with, for instance, **?**, that figurative language is going to built into any sample of language no matter how small or basic). A second problem is that, specifically to overcome the problems with modelling semantic relationships merely in terms of collocations, distributional semantics compares the co-occurrence profiles of words rather than their direct relationships, and it seems likely that word combinations prone to metaphoric interpretation might very well have at least overlapping profiles.

So the objective of the experiments reported in this chapter will be to explore the ways in which and the degrees to which a more fleshed out statistical description of contextually selected distributional semantic subspaces can reveal figurative language. As with the experiments on relatedness and similarity reported in the previous chapter, in addition to the relationship between target word-vectors in the subspaces they select, the statistical properties of the selected dimensions themselves will also be examined. And, again as with previous results, the instrument of analysis will be the geometric features of the subspaces in question, with, again, particular attention paid to the way in which the sets of features can collectively indicate figurative language. The two primary datasets explored represent binary decisions about metaphoricity and coercion respectively, and so my models will be applied to classification tasks here. In the case of metaphor, I test whether a model learned based on classification data is generalisable to graduated human ratings of metaphoricity. With the coercion data, I will examine whether the addition of information about sentential context enhances the classification of word pairs. I will

conclude the chapter with a reflection on some of the theoretical implications of the strongly positive results described here.

3.1 An Experiment on Metaphor

As pointed out by **?**, statistical approaches to metaphor identification and interpretation have generally been formulated in the context of the *conceptual metaphor* theory of **?**. This model is founded on the principle that "we systematically use inference patterns from one conceptual domain to reason about another conceptual domain," (ibid, p. 246). Metaphors are then the mechanism for performing the mapping between these domains, and as such cut right to the core of cognitive processes. Statistical models of metaphor have accordingly treated metaphors as transformations of lexical representations, and vector space models of distributional semantics have naturally leant themselves to this type of approach. The construction of representations with the potential to interact with one another in semantically productive ways has in turn lent itself to the development of models that consider the compositional nature of metaphor, effectively treating the metaphor itself as a transformation of the underlying representations. So **?** constructs candidate metaphor-vectors by calculating the centroid of a number of vectors derived from an analysis of a noun-vector and a predicate-vector learned through latent semantic analysis, and then uses the spatial relationships between these composed vectors to analyse the metaphoricity of certain phrases. **?** similarly consider composition in their approach to metaphor classification, in this case by combining word-vector type representations with a model trained to identify metaphor based on dependency trees of sentences labelled for metaphoricity.

In the tradition of work on compositional distributional semantics explored by the likes of **?**, **?**, and **?**, among others, semantic types such as adjectives and verbs are modelled as tensors which perform transformations on nouns, which are modelled as vectors. In the normal run of things, compositional models therefore represent, for instance, noun phrases modified by adjectives as the product $A\overrightarrow{n}$, where A is a matrix representing an adjective learned from observations of attested instances of the adjective with other noun word-vectors. So the phrase *black dog* becomes a word-vector in the same space as the representation of just *dog*, and can be compared quantitatively and geometrically with other phrases such as *white dog* or *big cat* and so forth. In the case of metaphor, these transformations are expected to map the word-vector representing metaphoric phrases into a region corresponding to the semantic domain of the original noun-vector modified by a metaphoric interpretation of the word associated with the tensor of a modifier or a predicate. So, for instance, in a model that effectively captures metaphoricity, the

composition of the vector space representations corresponding to *brilliant light* would map to a region of space where comparisons between phrases like *dark illumination* and *red glow* are productive, while *brilliant child* might be expected to map into the proximity of *stupid boy* and *boring girl*.[1]

The data that I will use in this section to test my methodology was originally presented by Gutiérrez et al. (2016), along with an accompanying experiment on a novel model. It consists of 8,592 adjective-noun pairs, spanning 23 adjectives chosen for their membership in six different broad semantic categories that are prone to both literal and metaphoric use: so, for instance, *bitter*, *sour*, and *sweet* are considered constituents of the category TASTE. There are 3,473 different noun types used, with only 141 types, represented by 640 tokens, occurring in both literal and metaphoric phrases. Each pair has been rated as either literal or metaphoric by a pair of human annotators, with inter-annotator agreement measuring at Cohen's $\kappa = 0.80$; 4,593 of the pairs have been judged metaphorical. This dataset was conceived as something of an expansion of the similar but smaller corpus of adjective-noun phrases annotated with binary metaphoricity classifications presented by **?** (and those authors tested their own data with an assortment of models, achieving highest f-scores by applying a random forest classifier to the features of an existing library of distributional semantic word-vectors).

In their own experimental treatment of the data, Gutiérrez et al. constructed a pair of compositional models in the mode of **?**, learning adjective matrixes A to map from noun-vectors to noun-adjective phrase-vectors extracted from observations of co-occurrences of both nouns and phrases in a corpus. By creating separate tensor representations for literal and metaphoric instances of a given adjective, the authors can then compare the relationships between the vectors resulting from a noun-vector composed with literal and metaphoric senses of an adjective-vector to try to determine whether a given phrase would generally be classified as a metaphor or a literal expression by comparing the respective compared vectors to the phrase-vector as observed in the corpus. In a further attempt to generalise the method, and, notably, to apply the conceptual metaphor theory of **?** to their computational model, the authors learn matrices performing linear transformations from literal to metaphoric adjective-noun compositions and then compare the similarity between observed phrase-vectors and literal composed vectors versus transformed literal composed vectors to determine whether a given phrase is metaphoric or not.

The data described by Gutiérrez et al. will serve as the basis for testing my own context sensitive distributional semantic methodology's ability to classify phrases as literal or metaphoric, and the results of this experiment will be described in the following

[1] It should be noted that such a methodology at this point begins to assume dim shades of Gärdenfors's (2000) conceptual spaces, with different compositions inherently defining different regions of the space.

section. My hypothesis is that metaphor, and indeed all figurative language, is fundamentally entangled with the context mutually indicated by the representations of the words participating in the composition being analysed. In fact, I think that part of what is captured by the model described by Gutiérrez et al., and indeed a number of other researchers investigating statistical methods for metaphor classification, is precisely that there is a context inherent in the linear algebraic dynamics of composable lexical representations, and this is something which many researchers explicitly recognise. But I also think that the explicit projection of context specific semantic subspaces, the mainstay of my methodology, should provide an ideal testing ground to discover the way in which statistical geometry can directly broadcast the presence or absence and even potentially the degree of metaphor inherent in a given phrase. The following sections will test this hypothesis using a similar methodology to that applied to semantic relatedness and similarity in the previous chapter.

3.1.1 Methodology and Results

My own methodology is clearly less committed to maintaining distinct representations for different semantic types than the compositional models described above, instead modelling all words as untagged word-vectors based on their co-occurrences as observed across a large scale corpus. This feature of my research is in part theoretically motivated: in line with **?**, and *contra* the grammatic nativism or exceptionalism that has been a mainstay in theoretical linguistics (**?**), I would like to investigate the possiblity that "grammar is fully and appropriately describable using only symbolic units, each having both semantic and phonological import," (ibid, p. 290). In other words, the syntactic component of a natural language might be described in terms of the entanglements of the meaning-making structures – the lexical semantic representations – that arise in the course of language use, or maybe even as emergent properties of these entanglements.

With this in mind, I will approach the problem of metaphor classification with a similarly statistical and geometric methodology as was applied to relatedness and similarity in the previous chapter, outside of any prima fascie model of syntax or compositionality. For every pair of words in the data produced by Gutiérrez et al. (2016), I generate subspaces of 20, 50, 200, and 400 dimensions using the JOINT, INDY, and ZIPPED techniques, projected from 2x2 and 5x5 word co-occurrence window base spaces. This data specifies a distince role for each word, one being a metaphoric source (the adjective) and the other being a target (the noun): so, for instance, a *bitter loss* is a loss, but presumably not one with an actual taste, and so the noun *loss* co-opts something of the quality of bitterness into its own conceptual domain. As such, it might be useful to generate subspaces

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.839	0.860	0.878	0.881	0.840	0.862	0.880	0.886
INDY	0.821	0.839	0.855	0.860	0.817	0.840	0.858	0.867
ZIPPED	0.839	0.864	0.876	0.878	0.833	0.854	0.873	0.880
ADJECTIVE	0.771	0.860	0.828	0.845	0.781	0.804	0.828	0.837
NOUN	0.819	0.861	0.843	0.847	0.806	0.821	0.838	0.843
SVD	0.685	0.703	0.703	0.697	0.677	0.694	0.687	0.684
SG	0.679	0.676	0.679	0.673	0.664	0.665	0.672	0.656
CBOW	0.669	0.681	0.677	0.672	0.669	0.673	0.677	0.671

Table 3-A: F-scores for metaphor identification based on a stratified ten-fold cross-validated logistic regression taking geometric features of various subspace types as input.

based simply on an anlysis of the word-vectors corresponding to the adjective and the noun respectively. I do this by simply selecting the top d dimensions, in line with the dimensionality parameter for each model, for the term in question, and these spaces are labelled ADJECTIVE and NOUN in the results that follow.

In each subspace, I extrapolate the same 34 geometric features described in Table 3-J and applied in the previous chapter in the semantic relatedness and similarity experiments. Again because of the semantic asymmetry of the relationship between the input terms, an additional seven features are also available in these spaces: the adjective-vector norm divided by the noun-vector norm (A/B), likewise the lengths of the vectors between the adjective and the generic points divided by the lengths for the noun-generic-point vectors ($\overline{AC}/\overline{BC}$, $\overline{AM}/\overline{BM}$, and $\overline{AX}/\overline{BX}$), and the corresponding fractions of the normalised versions of these points ($\overline{A'C'}/\overline{B'C'}$, $\overline{A'M'}/\overline{B'M'}$, and $\overline{A'X'}/\overline{B'X'}$). These additional measures might offer a sense of wehther there are statistical tendencies that are specific to the semantic role being played by a word moving from literal to metaphorical relationships, and we might expect this to be particularly evident in the spaces selected by either the noun or the adjective on their own. As with the subspaces of relatedness and similarity, I normalise each feature across all word pairs to have means of 0 and standard deviations of 1.

In order to test the capacity of the geometric features of my subspaces to identify metaphor, I perform a stratified ten-fold cross-validated logistic regression taking these features as independent variables and learning to predict the classifications assigned to the word pairs in the dataset. Balanced f-scores based on the precision and recall of my various dimensional selection techniques as well as static SVD factorisations of my base spaces and the `word2vec` models are reported in Table 3-A. The first thing to note is the strong performance across the board of the context sensitive methodology: the

model based on my strongest performing subspace (JOINT, 5x5 window, 400 dimensions) substantially outperform the strongest versions of the static models (the SVD 5x5, 400 dimension model) with $p < .005$ based on a permutation test. The context sensitive models perform better, but only marginally better, in the 5x5 word window subspaces, suggesting that most of the useful information about the semantic properties that indicate a metaphoric projection are captured by the profile of terms co-occurring in close proximity to the target words. That this trend is reversed for the static spaces, with 2x2 word window spaces doing a bit better, further indicates that the peripheral information of wider ranging co-occurrences is specifically useful for a context sensitive analysis.

The JOINT technique gives the strongest results, suggesting that subspaces delineated in terms of co-occurrence dimensions mutually salient to both input terms offer the best platform for analysing metaphoricity. This makes sense: in the case of metaphor versus literalness, it is the co-occurrences that both words have in common that position their respective word-vectors in an indicative relationship relative to one another and the subspace overall. So for instance the co-occurrences salient to both *sweet* and *fruit* will have a particular conceptual profile that will not be evident in the dimensions jointly selected by *sweet* and *revenge*; this effect will be less evident for dimensions independently salient to each word. ZIPPED subspaces, where there will be at least some information about both words along every dimension, accordingly score almost as well as JOINT subspaces, with the INDY subspaces falling further behind.

Interestingly, the ADJECTIVE and NOUN spaces classify metaphor most accurately in 50 dimensional subspaces projected from the 2x2 word window base space. To the extent that part-of-speech can be a component of the analysis of these models, we can expect the smaller co-occurrence window to produce statistics that are more indicative of a particular grammatical class. The degradation of classification at higher dimensionalities for the smaller co-occurrence window setting is a little surprising, and it's worth noting that the INDY subspaces, which are basically blends of the ADJECTIVE and NOUN subspaces, don't exhibit the same tendency. In this case, it would seem the whole really is greater than the sum of the parts, with the dimensional selection of one word providing at least a degree of useful information about the other word not available in spaces salient to a single term. A similar pattern emerges for the static spaces: the SVD, SG, and CBOW models all produce the most accurate classifications in 2x2 word window, 50 dimensional subspaces. One way to explain this is that more ambiguous information about word use begins to leek in at higher dimensionalities, servining to obscure the more standard indications available in either the most salient dimensions or the dimensions containing the most information about variance across the corpus.

There is another possibility to consider regarding the adjectives in this dataset in par-

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.815	0.837	0.854	0.855	0.816	0.837	0.858	0.863
INDY	0.778	0.793	0.828	0.835	0.774	0.805	0.829	0.842
ZIPPED	0.810	0.838	0.847	0.854	0.799	0.828	0.844	0.853
ADJECTIVE	0.606	0.709	0.750	0.777	0.698	0.697	0.757	0.707
NOUN	0.806	0.808	0.828	0.833	0.796	0.812	0.824	0.829
SVD	0.679	0.691	0.695	0.690	0.665	0.674	0.678	0.676
SG	0.668	0.664	0.659	0.657	0.659	0.656	0.644	0.638
CBOW	0.657	0.665	0.665	0.661	0.656	0.660	0.666	0.660

Table 3-B: F-scores for metaphor identification with each of the conceptual categories identified by Gutiérrez et al. (2016) treated as a separate fold for cross-validation.

ticular: as there are only 23 different adjective types, each adjective is observed multiple times in both metaphoric and non-metaphoric contexts. It is therefore possible that, in any given fold of the cross-validation of a classifier, the model might be learning how to guess whether a specific adjective is involved in a metaphor rather than something more general about the statistical geometry of metaphoricity. In order to avoid this trap, I reorganise the data into tranches based on the adjective in each pair, I use the eight conceptual categories outlined by Gutiérrez et al. (2016) in order to structure this new partioning.[2] I use each of these eight new sets of word pairs as a fold in a cross-validated logistic regression, such that the adjective in each phrase in each test set has not been observed in the training data.

Table 3-B presents the results from this reshuffled version of the experiment. The f-scores for metaphor classification returned by the context sensitive models are down slightly, but the difference is only marginally significant

XXX SIGNIFICANCE

The major change here is, as expected, in the ADJECTIVE subspaces: clearly when only information from the adjective in each word-pair is used to train a model, prior observations of a specific word type in the context of some other composition is a benefit. There is also a minor decrease in performance for the static models, which is interesting in that it indicates that, even when a single distance metric is used to classify metaphoricity, observations of a word in training help to subsequently test phrases involving that word. It is worth noting that of the 8,584 noun tokens spread across 3,473 noun types, 1,588 types, represented by 6,724 tokens, occur in more than one of the tranches delineating

[2]Gutiérrez et al. (2016), identifying a similar problem, likewise develop a second model that learns metaphors as mappings between domains rather than just from noun-vectors to phrase, though their methodology requires them to use a reduced version of the data.

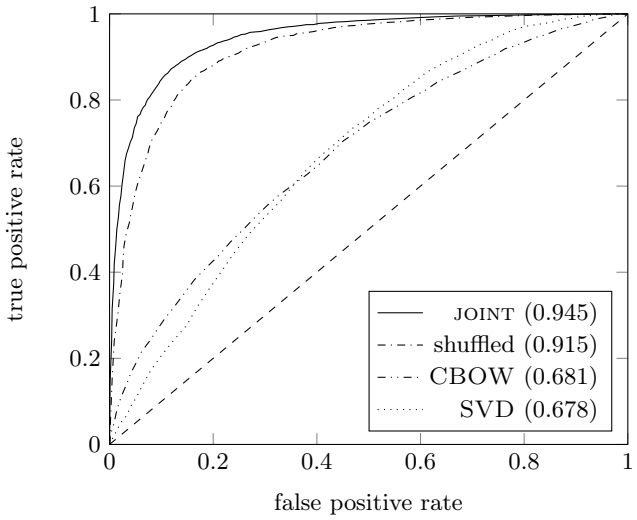


Figure 3.1: Receiver operating characteristic plots for a selection of models, with the area under the curve for each model type indicated in the legend.

the conceptual categorisations of the adjectives, so it is possible that there is a small extent of learning to classify phrases based on previous observations of specific nouns.

In order to take a closer look at the way that different techniques model this data, and in line with the metaphor classification work of **?**, Figure 3.1 illustrates receiver operating characteristic curves for four versions of the approaches that have been described here: the JOINT technique with 400 dimensional, 5x5 word subspaces, the same technique applied to the version of the data shuffled to avoid training and testing on the same adjectives, and the CBOW and SVD models for the optimally performing 50 dimensional, 2x2 word window subspaces. True positive versus false positive rates are correlated at 99 increments in terms of the value of the output of a logistic regression model at which a phrase is determined to be metaphoric. The outcomes visualised here tell a similar story to Tables 3-A and 3-B, with the area under the curve statistics indicating a strong distinction between the context sensitive techniques and the static models. Perhaps the most interesting thing to note is the overall smoothness of the curves, which suggests a steady relationship between precision and recall at various classification thresholds.

With the trade off between true and false positives in mind, Table 3-C presents precision, recall, f-score, accuracy, and Cohen's kappa scores for the same models plotted in Figure 3.1. The trend to notice here is that context sensitive and static models tend to favour recall over precision (and the slight preference for precision in the JOINT 400 dimensional, 5x5 word subspaces for the shuffled version of the data reported here is an anomaly, as other approaches to that data exhibit the tendency towards higher recall). This evident enthusiasm for classifying phrases as metaphoric is a reflection of the data

	precision	*recall*	*f-score*	*accuracy*	*kappa*
JOINT	0.879	0.894	0.886	0.877	0.753
shuffled	0.873	0.865	0.862	0.854	0.678
SVD	0.631	0.794	0.703	0.641	0.265
CBOW	0.638	0.721	0.677	0.632	0.253
Gutiérrez et al. (2016)	0.842	0.793	0.817	0.809	0.618
baseline	0.535	1.000	0.697	0.535	0.000

Table 3-C: Full classification statistics results for the models tested here as well as the results from the original literature and the majority class (metaphor) baseline.

itself, which is slightly skewed towards metaphoric phrases, as described above and indicated in the performance of the majority class baseline, and this is reinforced by the relatively low accuracy scores for both context sensitive and static non-compositional distributional semantic models. It is noteworthy, then, that the model described by Gutiérrez et al. (2016) actually scores better for precision than recall, suggesting it actually tends to under-predict metaphoricity. This could perhaps be expected as a general distinction between statistical models based on unannotated data such as mine, which will arguably tend to favour a majority class, versus likewise statistical models operating on theoretically motivated mappings between representations, which have an apparent propensity for zeroing in with confidence on the properties of a compositional transformation that are indicative of metaphor—but at the expense of sometimes missing what might be considered outliers. In the same spirit, the jumpier nature of the receiver operating characteristic plots presented by **?** is quite possibly an artefact of the decision points inherent in heuristically mapping model features from human made knowledge bases.

As a final point of comparison with other approaches to metaphor classification, I will return briefly to the unannotated character of my lexical representations. One of the most powerful features of the methodology described here is its ability to build a somewhat general model of a semantic phenomenon from a sufficiently comprehensive dataset, and the strong Cohen's kappa score of the best performing subspace selection technique, which begins to approach the aforementioned inter-annotator agreement level of $\kappa = 0.80$, is a testament to this. Following an analysis of the specific geometry of metaphor in the next section, Section 3.1.3 will assess the ability of my methodology to generalise even further from this data to a broader range of metaphors and to moreover move from classification to gradation based on observations of merely binary judgements of metaphoricity. For now, I simply note that it is remarkable that data about nothing more than the way that words tend to be collocated can, with the aid of a mechanism for contextualisation, reveal so much about the nature of the semantic relationship between

JOINT		INDY		ZIPPED	
$\mu(A,B)$	0.787	C	0.767	$\mu(A,B)$	0.788
C	0.771	C/M	0.749	C	0.771
$\mu(A,B)/M$	0.764	$\angle AMB$	0.747	$\mu(A,B)/M$	0.769
$\angle COX$	0.762	C/X	0.746	X	0.767
X	0.762	$\mu(A,B)$	0.734	$\mu(A,B)/X$	0.759

ADJECTIVE		NOUN	
$\mu(A,B)/M$	0.745	$\mu(A,B)$	0.756
$\overline{AC} : \overline{BC}$	0.736	C	0.747
$\overline{AC}/\overline{BC}$	0.734	$\mu(A,B)/X$	0.728
$\mu(A,B)/X$	0.732	$\mu(A,B)/M$	0.721
$\angle ACB$	0.730	C/X	0.721

Table 3-D: Independent f-scores from the metaphor classification data for top five features of each subspace type for 5x5 word co-occurrence window, 400 dimension subspaces.

the lexical components of an previously unseen phrase.

3.1.2 The Geometry of Metaphor

In this section, I will explore the geometric features which prove most productive in the classification of metaphor. As with relatedness and similarity in the previous chapter, I begin by examining the capacity of independent features to predict metaphor. Rather than a proper logistic regression involving multiple independent variables fed into a non-linear function, this analysis amounts to choosing a cut-off point in terms of the value of each feature separating literal and metaphoric phrases in the subspaces which an analysis of their corresponding word-vectors delineate. So the f-scores reported in Table 3-D can be understood as indicating the degree to which the values of a given geometric feature separate the dataset into distinct categories corresponding to human judgements of metaphoricity.

The scores themselves reflect the trend observed in Table 3-A and 3-B: the JOINT and ZIPPED subspaces produce features that are particulary good at classifying metaphor, with a decrease in performance in the INDY subspaces and then another step down in the single-word subspaces. None of the scores themselves come close to the levels of discrimination achieved by the models learned from full feature vectors, though

XXX SIGNIFICANCE

In terms of the actual features indicated by this analysis, two in particular figure prominently in one way or another, namely, the mean of the word-vector norms $\mu(A,B)$

and the norm of the central-vector C. In the first instance, the role of the relationship between word-vectors and the origin of the spaces that their salient co-occurrence dimensions delineate is once again reflective of the preliminary findings on conceptual geometry described in Chapter **??**, where norm was seen to be an effective mechanism for defining a region of conceptual constituency. In the case of the distance of the central vector from the origin, the emergence of this feature, as well of the appearance of the norms M and X as components of various strongly predictive tendencies, indicate that here, as with similarity in the previous chapter, characteristics of dimensions outside of the situation of any particular word-vector along them might be in themselves indicative of metaphor: some words might simply be more likely to co-occur in the context of metaphoric language, and co-occurrence statistics should provide a handle for examining this tendency.

To further delve into the statistical geometry of metaphor, and in line with the results on relatedness and similarity described in the previous chapter, I once again search the state space of possible combinations of features to find the optimal feature vector for classifying metaphor in context sensitive subspaces. This is again treated as a beam search problem, though the search space expanded at each level of the search tree is here limited to the top 500 combinations of features given the larger size of the data being modelled. Table 3-E presents the optimal seven feature combinations discovered for the 5x5 word window, 400 dimensional JOINT subspaces based on both a standard ten-fold cross-validation and the version of the data shuffled in order to test on data not observed in each training phase. The f-scores achieved by these combinations of features, reported next to the respective labels at the top of the table, indicate a marginal decrease in the overall performance as compared to the full featured models of subspaces, but the results are still strong.

Angles between generic vectors, which were already evident as independently predictive features in Table 3-D, have a strong effect here, with the strong negative correlation of $\angle COX$ in the ten-fold cross-validation in particular suggesting that maximal values tend to be relatively similar across dimensions jointly selected by literal adjective-noun combinations, pulling the line of X closer to the centroid described by C. To put this differently, as pairs become more metaphoric, they tend to also become less consistent in the type of dimension that they co-select, as evidenced in the increasing variance in the maximum values of these dimensions. Perhaps the most interesting thing to observe here, though, is the strong correlation between ratios of word-vector to generic vector distances in the case of the version of the data shuffled to test on unseen adjectives, but not in the case of the stratified cross-validation. The positive correlation with the balance of the distances from the word-vectors to the mean vector M means that subspaces where the word-vectors have a relatively even relationship to the weighted centre are, in fact, more

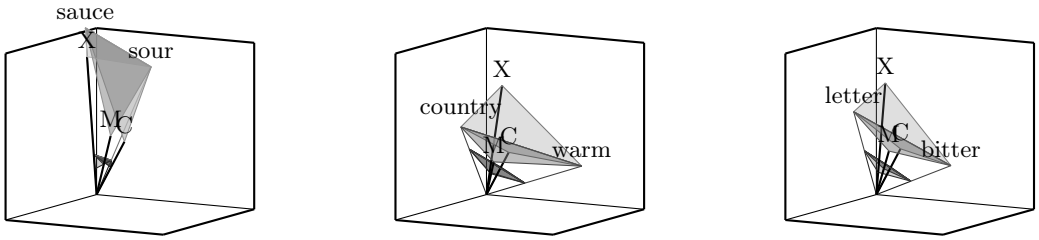
	10-fold ($f = 0.869$)	shuffled ($f = 0.830$)
	DISTANCES	
word-vectors	-	-
generic vectors	$M = -1.448$	-
	ANGLES	
word-vectors	$\angle ACB = -0.775$	-
normalised	-	-
generic vectors	$\angle COX = -1.618$	$-0.271 = \angle COM$
	$\angle COM = 0.974$	$0.045 = \angle MOX$
	MEANS	
word-vectors	$\mu(\overline{AM}, \overline{BM}) = -1.124$	$-1.007 = \mu(\overline{AC}, \overline{BC})$
normalised	-	-
	RATIOS	
word-vectors	-	$0.492 = \overline{AM} : \overline{BM}$
		$-0.620 = \overline{AX} : \overline{BX}$
normalised	-	$-0.168 = \overline{A'C'} : \overline{B'C'}$
	FRACTIONS	
word-vectors	$\overline{AC}/\overline{BC} = 0.325$	-
generic vectors	$M/X = 1.305$	$0.252 = A/B$

Table 3-E: The seven most predictive features for metaphor classification, compared between ten-fold and sight-unseen cross-validation of logistic regression on statistics extrapolated from 5x5 word window, 400 dimensional JOINT subspaces.

metaphoric (and their relationship to the maximum vector is comparatively less balanced, with this vector in turn being less central to the space per the observations regarding $\angle COX$). But more generally, it is noteworthy that the balance between word vectors and generic vectors is informative about metaphoricity specifically in models tested on unseen adjectives: this balance is in effect a projection into space of quotients of joint probabilities of observing words and co-occurrence terms divided by the typical or maximal probabilities of being observed with the co-occurrence terms, and from it we can infer that these quotients are generally predictive of metaphor in context, even without word-specific training data.

Figure 3.2 presents visualisations by way of three dimensional projections of word-vectors and generic vectors from 400 dimensional JOINT subspaces selected from the 5x5 word window base space.[3] In the example of the uncontroversially literal phrase *sweet watermelon*, the word-vectors are characteristically far from the origin and close to one another, corresponding to the predictivity of $\mu(A, B)$ in particular. At the other extent of the spectrum, the highly metaphoric phrase *bitter letter* is characterised by a dropping

[3]These projections have been rendered using the same regression technique as applied to the images for related word pairs in the previous section, but the coordinates of X have been divided by 1.5 instead of 2.



literal: sweet watermelon *neutral: warm country* *metaphoric: bitter letter*

Figure 3.2: Three dimensional projections of word-vectors and generic vectors in subspaces for pairs at the extents and in the middle of the literal-metaphorical spectrum, taken from 5x5 word window, 400 dimensional subspaces selected using the JOINT technique.

of the word-vectors and a widening of the angle between them; the generic vectors, meanwhile, are now further from the origin relative to the word-vectors that select the subspace and, at the same time, draw closer to one another in particular at the normalised layer of the subspace. But most interestingly, at a relatively neutral point, occupied by the intriguingly ambiguous phrase *warm country*, which the logistic regression trained on these subspaces assigned a score of close to 0.5, there is actually evidence of an intermediary widening out of the overall array of points even as the word-vectors remain fairly far from the origin.

The exciting thing about this last observation is that it suggests that, rather than existing on a linear or even monotonic scale, metaphor may itself actually be a multi-dimensional phenomenon, with a characteristic particular to highly ambiguous word combinations that is to some extent separate from the statistical features of straightforward literalness and clear cut metaphoricity. The broad arrangement of word-vectors in space engendered by the contextualisation of the phrase *warm country*, in contrast to the relatively tight relationship of the generic vectors, can be interpreted as revealing an uncertainty regarding the semantic properties being transferred in this small composition, corresponding to a drifting of the word-vectors and a contracting of the generic vectors across the jointly selected co-occurrence profile. Here, once again, the statistical geometry of a subspace can be productively mapped to a theoretical statement about the nature of a semantic phenomenon as characterised by a selectively contextual and quantitative representation of observations about the way that words are used, by and large outside of any strong preconditions symbolically encoded in the computational framework.

3.1.3 Generalising the Model

One of the interesting things about feature-based classification is that there is typically an inherent commitment to degree of class membership, even when the training data used to build a model is simply binary. This is true of any model which uses, for instance, a logistic regression technique for determining class, as there is a cut-off point along the spectrum of model output and a corresponding proximity to that point for any given sample, and it is especially obvious when the features of the model are actually geometrical measures. In this section, I will apply the models learned from the the Gutiérrez et al. (2016) data to another dataset designed to assess metaphor as a matter of degree rather than simply as a binary situation, and a dataset that additionally deals with a different type of metaphor in terms of composition. The question explored here is whether the geometric features of context specific distributional semantic analysis of word-vectors will provide binary classification models with adequate information for projecting metaphoricity along a continuous scale.

The data used for this experiment was originally reported by **?**, and was used to train a model based on an earlier version of methodology as described by **?**. This data consists of 228 predicate-object word pairs selected to cover three degrees of metaphor, consisting of literal pairs such as *announce willingness*, conventionally metaphoric pairs such as *cut pollution*, and novel metaphors such as *smell excuses*. 102 human participants provided metaphoricity scores on a seven point Likert scale, and the average scores were compiled into the dataset that will used to test models learned from the geometric features output by my context sensitive methodology.[4] Specifically, I will experiment with two different classification model techniques. In the first instance, I will take the output of the logistic regression described above, trained on the Gutiérrez et al. (2016) data, as assigning probabilities to the metaphoricity of an input word pair, and I will in turn measure the degree to which these probabilities correlate with the degree of metaphoricity collectively assigned by human raters. In the second instance, I'll use the binary metaphor classification data to train a support vector machine.[5] Applying a radial basis function kernel, I analyse the correlation between distance from the discriminatory hyperplane and the human ratings. In both cases, and in line with results reported in the previous chapter, Spearman's correlations are the unit of analysis.

Table 3-F presents results for both modelling techniques, focussing on features extrapolated from 5x5 word window, 400 dimensional subspaces using the JOINT approach and through an analysis on the adjective in each word-pair from the input data. Feature

[4]Studies were also conducted to gather ratings for *familiarity* and *meaningfulness*, but those ratings will not be modelled in this thesis.

[5]This is implemented using the python scikit-learn `SVC` module.

features	1	3	5	7	9	full
			logistic regression			
JOINT	0.368	0.355	0.033	0.085	0.279	-0.033
ADJECTIVE	-0.377	0.355	0.044	0.513	0.511	0.335
			support vector machine			
JOINT	0.352	0.359	0.042	0.045	0.243	0.158
ADJECTIVE	-0.170	0.247	-0.021	0.407	0.418	0.236

Table 3-F: Spearman's correlation with human verb-noun metaphoricity scales judgements based on logistic regression and support vector machine models trained on adjective-noun classification data, taking feature vectors of various lengths as independent variables.

vectors of various lengths, picking the optimal geometric features for each dimensional selection technique, are used to feed input to each model. In terms of the models trained on features from JOINT subspaces, there is a clear trend towards strong performance with one or three features, weaker performance with five or seven features, stronger performance again with nine features, and then a drop-off again in the full featured space. The relatively low performance with the full set of features is not particularly surprising: there is clearly an encroaching incidence of generalisation error here as the models become flooded with data about various and certainly collinear statistical features of contextual geometry. At the shallow end of the feature selection parameters, on the other hand, the single measure $\mu(A, B)$ (per Table 3-D) once again points to the efficacy of word-vector norm as a predictive characteristic of contextualised co-occurrence subspaces.

The really remarkable outcome here, though, is the very strong performance of the models learned from the top seven and nine features extracted from subspaces selected by PMI values of the adjective word-vectors alone. This is particularly interesting given that the data being tested actually consists of a different type of grammatical relationship, namely, predicate-object pairs. It would seem, then, that the co-occurrence dimensions most salient to either verbs or adjectives generate a geometry in which their relationship to potential arguments can play out in similar ways in terms of the metaphoricity inherent in the semantic context: the interaction between the selecting vector, the noun-vector, and the generic vectors translates from one type of composition to another in an isomorphic way. This explantion, including the claim that the mapping of predictive features from one type of metaphor to the other is to a large extent isomorphic, is supported by the particularly strong performance of the logistic regression at seven and nine dimensions, where the logistic function takes a polynomial with coefficients learned in the training phase as direct input. The more complex non-linearity afforded by the support vector machine appears to actually somewhat confound the mapping from verb-noun to adjective-noun phrases—though the difference between the correlations at nine dimen-

sions is not statistically significant at $p = .104$ based on a Fisher r-to-z transformation.

The one area where a support vector machine provides a clear improvement in performance is in the full dimensional models extrapolated from JOINT subspaces. In this case, it would seem that the radial basis function classification actually does a better job of avoiding the overfitting in a higher dimensional feature space. But, putting questions of model choice aside, there is clear evidence here for the generality of the contextual geometry of metaphor, and also a strong case for the appropriateness of machine learning techniques for providing an appropriate mechanism for the computational manipulation of co-occurrence information to build a more nuanced model of degree of metaphor based on relatively rudimentary classification data. Crucially, it is the context sensitivity of my methodology that facilitates the exploration of a multi-dimensional feature space in which the non-linear nuances of this particular semantic phenomenon can be discovered; a model providing a singular static relationship between lexical representations could not offer the context specific underpinning for generating a geometry replete with interpretable statistical features. Finally, there are signs here to invite further research, and indeed some grounds for hoping that a context sensitive approach might have the scope for handling more sophisticated tasks such as metaphor interpretation and generation.

3.2 An Experiment on Coercion

In this section, I will apply my methodology to the classification of a phenomenon closely related to metaphor, namely, *semantic type coercion*, by which the semantic type of a noun is reassigned in the course of a verb taking that noun as an argument. So, for instance, in phrases like *denied wrongdoing* or *heard footsteps*, the nouns in play are standing in for a conceptually relevant but different type of noun, and the literal versions of these phrases would go something like *denied committing wrongdoing* or *heard the sound of footsteps*, where the verbs select arguments of types along the lines of ACTIVITY and PERCEPTION respectively. This phenomenon is often referred to as *logical metonymy*, identifying it as a subspecies of the more general figurative phenomenon metonymy by which a thing is denoted by a conceptually related lexical representation.

Coercion is one of the semantic phenomena targeted by Pustejovsky's (1995) theory of a *generative lexicon*, by which nouns are semantically modelled as having a *qualia structure* which maps out the way that a thing relates to itself, the world, and the agents interacting with it in that world on four different levels of abstraction, with the general objective of arriving at "a model of meaning in language that captures the means by which words can assume a potentially infinite number of senses in context, while limiting

the number of senses actually stored in the lexicon," (ibid, p. 104). In terms of coercion, qualia provide the basis for a process of *projection* by which a variety of semantic types can be extracted from a complex type (or a *dot object* in Pustejovsky's lingo) in order to fulfil the typing requirements of a predicate in open ended ways. The model that emerges here – one built on dynamically interactive lexical semantic representations contingent on some sort of general conceptual context – begins to look like the general linguistic stance that has motivated my own methodology.

This theoretical commitment suggests a schematic by which a symbol manipulating system might begin to get a handle on productive and context sensitive lexical representations of things in the world. To this end, **?** have described an ontology based on a computational analysis of co-occurrence patterns designed to facilitate the modelling of what is ultimately a sliding scale of statistically enhanced semantic representations, or "shimmering lexical sets," (ibid, p. 19), as the authors put it. Applying a similar notion that coercion is probabilistic rather than discreet, **?** use co-occurrence statistics to try to predict the verbs which, in the role of for instance participles, successfully resolve instances of coercion. And, under the rubric of *logical metonymy*, **?** expand upon the work of **?** by extracting verb senses from WordNet to build a class based model, to some extent recapitulating the categorical distinctions that characterise many theoretical approaches to coercion. The motivation behind this last system is the apt observation that, in the case of coercion, "humans are capable of interpreting these phrases using their world knowledge and contextual information," (**?**, 11:2).

Returning to the theoretical issues regarding grammaticality raised earlier in this chapter, the analysis of coercion within the framework of the generative lexicon points to something more like a graduated typology, sliding from specific instances of processes, things, and the like to more general conceptual categories and finally to entire classes of words. As **?** has pointed out, there is a lurking ambiguity in grammatical class distinctions, with various conceptual schema existing in any natural language for moving between classes: so, to borrow an example from Langacker, phonological and symbolic dynamics facilitate a conceptually coherent progression from *sharp* to *sharpen* to *sharpener*, and the rules that are extrapolated as an explanatory framework for such transitions are just a way of systematising the cognitive networks that underpin these linguistic

6

With this in mind, my hypothesis is that, as with metaphor in the previous section, a syntactically neutral statistical model with a context generating capacity should be able to capture

[6] One is also reminded of **?**'s (**?**) quip regarding "grammatical fictions," (ibid, ¶307).

3.2.1 Methodology and Results

The data which will be used to test my methodology in this section was originally presented by **?** as a task for the ongoing International Workshop on Semantic Evaluation series of computational semantic modelling challenges. The data consists of 2,071 sentences each containing a marked verb and object, with the object classified as either coercive or not. The verbs cover various conjugations of five different verb stems, each identified as selecting for a different semantic type as an argument: the verbs (and the semantic type selected) are *arrive* (LOCATION), *cancel* (EVENT), *deny* (PROPOSITION), *finish* (EVENT), and *hear* (SOUND). The objective, then, is to train a model to indicate that the phrase *finish the party* is not coercive, in as much as we accept that *party* denotes a member of the conceptual category EVENT, whereas *finish the food* is because what is actually being finished is the event of eating food, not the food itself. For the purposes of the original presentation the data is split into a training set and a testing set of roughly equal size, but questions of the most meaningful partitioning of the data will be discussed below.

Two amendments are made to the data as presented. First, of the 2,071 verb-object pairs, 197 contain multi-word objects not compatible with the vocabulary used for my model, reducing the total number of word pairs to 1,874, 543 of which are considered coercive. Second, of these remaining computable word pairs, 903 are duplicates (they are presented in unique sentences, but for the first phase of analysis here only verb-noun pairs will be consider; sentential context will be addressed below). This leaves a total of 971 word pairs, 376 of which are deemed coercive. As with the metaphor data in the previous section, I train a logistic regression model to discriminate between regular argument selection and coercion. I once again take the two words being analysed as input to generate a number of different context specific distributional semantic subspaces, treating the 34 geometric features outlined in Table 3-J plus the seven additional fractional features specific to asymmetric input terms described above in Section 3.1.1 as the independent variables of the regression analysis.

Table 3-G presents the f-scores derived from the precision and recall results of a ten-fold cross-validation of these logistic regression models. Most obviously, these numbers are considerably lower than the comparable results for metaphor outlined in Table 3-A, but this is to some extent mitigated by the relative scarcity of instances of coercion in the data: a minority class baseline always classifying word pairs as coercive would, based on the above data statistics, give $f = 0.558$. The top score of $f = 0.708$ for the context sensitive models, achieved by the 5x5 word window, 400 dimensional verb-only dimensional selection technique, is significantly better than the baseline with

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.604	0.619	0.630	0.657	0.634	0.672	0.673	0.691
INDY	0.666	0.677	0.703	0.693	0.652	0.660	0.707	0.679
ZIPPED	0.568	0.624	0.610	0.647	0.596	0.625	0.658	0.663
VERB	0.664	0.675	0.698	0.704	0.631	0.652	0.699	0.700
NOUN	0.601	0.628	0.643	0.633	0.518	0.565	0.603	0.641
SVD	0.511	0.523	0.539	0.412	0.521	0.409	0.483	0.563
CBoW	0.498	0.508	0.531	0.493	0.496	0.544	0.535	0.496
SG	0.518	0.565	0.575	0.529	0.534	0.523	0.583	0.557

Table 3-G: F-scores for coercion identification based on a ten-fold cross-validated logistic regression taking geometric features of various subspace types as input.

XXX significance

Of the three dimensional selection techniques that use both words as input, the INDY method does best here (as opposed to the JOINT technique for metaphor), indicating that coercion plays out most clearly in the relative co-occurrence profiles of word-vectors across dimensions salient to words individually rather than jointly, and this is corroborated by the strong performance of the VERB and NOUN subspaces generated based on an analysis of either one of the two input words.

In line with the metaphor results is the poor performance of the static models, which generally do somewhat worse than the baseline and substantially worse than the context sensitive models. Of particular note is the decline of the SVD models and the comparative ascent of the `word2vec` skip-gram methodology: the sentential context predicting mechanism of the skip-gram approach seems to better capture the typological relationships between predicates and arguments than a principal component analysis of the dimensional variance in a base space of co-occurrence statistics. But in fact, the results here are across the board less regular in their relationship to parameters of dimensionality and co-occurrence window size, with a more even distribution of relatively high and low scores for both 2x2 and 5x5 word co-occurrence window models, and comparatively strong outcomes occasionally popping up for 20 or 50 dimensional spaces. The seemingly erratic output of the model gives an overall impression of an unanchoring between the statistics of co-occurrence and the semantic phenomenon being explored here. Perhaps in the case of coercion, or at least in terms of the data sampled here, many predicate-object combinations are, regardless of the influence of the verb on the noun's conceptual situation, too conventional for type shifts to be detected in a meaningful way in terms of co-occurrence profiles.

Another telling feature of these results is the quite strong performance of the subspaces

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.338	0.397	0.362	0.381	0.345	0.428	0.404	0.386
INDY	0.454	0.386	0.436	0.459	0.369	0.350	0.411	0.410
ZIPPED	0.256	0.297	0.363	0.358	0.324	0.352	0.377	0.357
VERB	0.233	0.334	0.361	0.448	0.307	0.401	0.352	0.336
NOUN	0.306	0.398	0.406	0.401	0.243	0.293	0.317	0.340
SVD	0.295	0.252	0.276	0.126	0.217	0.173	0.301	0.288
CBOW	0.368	0.329	0.248	0.162	0.302	0.316	0.245	0.177
SG	0.349	0.333	0.281	0.194	0.366	0.351	0.316	0.229

Table 3-H: F-scores for coercion identification taking each verb stem type as a separate fold of a cross-validation.

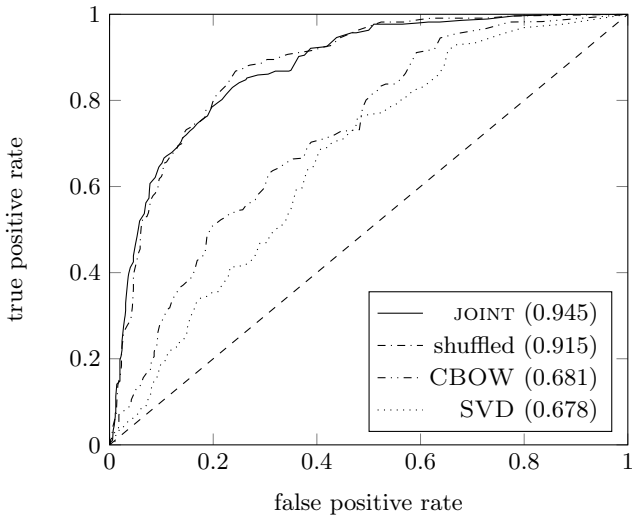


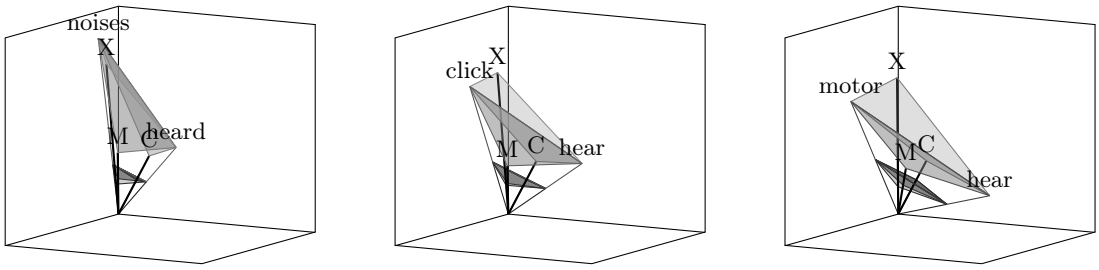
Figure 3.3: Receiver operating characteristic plots for a selection of models for coercion classification, with the area under the curve for each model type indicated in the legend.

selected by an analysis of the verbs alone. In fact, this is likely to be an artefact of the data itself: only five different verb stems are used, and some are arguably marked by their own semantic peculiarities, with, for instance, *finish* coercing 152 out of the 252 arguments it takes in the data, where the rate for *deny* is only 29 out of 183 instances. In order to find out if the models being tested here are actually just learning, in one way or another, specific rules about particular inputs, I rearrange the data into five folds corresponding to the five verb types present, training a model on each combination of four different verbs and then testing the model on the classifications of word-pairs involving the fifth. F-scores are reported in Table 3-H.

JOINT		INDY		ZIPPED	
$\mu(\overline{A'X'}, \overline{B'X'})$	0.526	$\mu(\overline{A'X'}, \overline{B'X'})$	0.547	$\mu(\overline{A'C'}, \overline{B'C'})$	0.392
$\mu(\overline{A'C'}, \overline{B'X'}$	0.496	$\mu(\overline{A'C'}, \overline{B'C'})$	0.544	$\mu(A, B)/C$	0.349
$\mu(\overline{A'M'}, \overline{B'M'}$	0.453	$\mu(A, B)/C$	0.522	$\mu(\overline{A'X'}, \overline{B'X'})$	0.321
$\mu(A, B)/C$	0.442	$\angle AOB$	0.517	$\mu(\overline{A'M'}, \overline{B'M'})$	0.237
$\angle AOB$	0.429	$\mu(\overline{A'M'}, \overline{B'M'})$	0.504	$\angle AOB$	0.209

VERB		NOUN	
$\overline{AC}/\overline{BC}$	0.580	$A : B$	0.528
$A : B$	0.412	A/B	0.486
A/B	0.387	$\mu(A, B)/C$	0.486
$\mu(\overline{A'M'}, \overline{B'M'})$	0.384	$\angle AMB$	0.427
$\mu(\overline{A'X'}, \overline{B'X'})$	0.374	$\angle ACB$	0.423

Table 3-I: Independent f-scores from the coercion classification data for top five features of each subspace type for 2x2 word co-occurrence window, 400 dimension subspaces, validated on unobserved verbs.



literal: heard noises *neutral: hear click* *metaphoric: hear motor*

Figure 3.4: Three dimensional projections of word-vectors and generic vectors in subspaces for pairs at the extents and in the middle of the literal-metaphorical spectrum.

3.2.2 The Geometry of Coercion

('0.683', [('mAMB', '0.848'), ('fABM', '-0.825'), ('aCOM', '0.407'), ('mAXB', '-0.349'), ('aAOB', '-1.756'), ('nACB', '0.537'), ('eACB', '0.207')])

3.2.3 Adding Sentential Context

3.3 Interpretation and Composition in Context

One of the tricky things about figurative language is its ephemerality: if we stare at it for long enough through a theoretical lens, it seems to vanish, as is evident in the deflationary case made by **?**. But on the other hand, if we ask someone in street whether the phase *buy a story* is more metaphoric than *buy a book*, we can reasonably expect the answer

	INDY ($f = 0.676$)	VERB ($f = 0.687$)
	DISTANCES	
word-vectors	-	-
generic vectors	-	-
	ANGLES	
word-vectors	$\angle AMB = -0.831$	-
normalised	-	-
generic	$\angle COM = 0.197$	-
	$\angle COX = -0.159$	-
	MEANS	
word-vectors	$\mu(A, B) = 0.919$	$-0.817 = \mu(A, B)$
normalised	-	-
	RATIOS	
word-vectors	-	-
normalised	$\mu(\overline{A'C'} : \overline{B'C'}) = 0.015$	
	FRACTIONS	
word-vectors	-	$0.490 = \mu(\overline{AB})/M$
		$0.443 = A/B$
		$-0.046 = \mu(\overline{AB}/C)$
normalised	-	$-1.863 = \overline{A'C'} : \overline{B'C'}$
		$-1.602 = \overline{A'M'} : \overline{B'M'}$
generic vectors	$C/M = -0.838$	$0.974 = M/X$
	$M/X = 0.234$	

Table 3-J: Comparison of the seven most effective features for coercion classification in 2x2 word, 400 dimensional subspaces for INDY versus VERB based dimension selection.

window		2x2				5x5			
dimensions		20	50	200	400	20	50	200	400
JOINT	nouns	0.157	0.174	0.244	0.283	0.193	0.244	0.257	0.271
	verbs	0.121	0.155	0.190	0.237	0.117	0.163	0.215	0.229
	adjectives	0.083	0.113	0.179	0.187	0.119	0.131	0.183	0.207
	adverbs	0.042	0.091	0.155	0.154	0.101	0.128	0.171	0.174
INDY	nouns	0.092	0.133	0.147	0.157	0.158	0.170	0.148	0.168
	verbs	0.117	0.126	0.173	0.165	0.147	0.209	0.174	0.201
	adjectives	0.123	0.114	0.162	0.172	0.173	0.161	0.151	0.184
	adverbs	0.115	0.137	0.139	0.120	0.167	0.146	0.121	0.111

will almost always be "yes", and it would be a mistake to dismiss the evidence that in a colloquial sense some compositions are clearly metaphoric, and others are clearly not. This raises a challenging point with regard to the comparison between metaphor and coercion, the two instances of figurative language explored in this chapter: is metaphor perhaps to some extent a more overt case of coercion, or maybe a specific case that is in some way or another a little more subtle? Part of the problem here is that the distinctions between these phenomena begin to exceed the capacity for what can reliably

be quantified about language in a clinical setting, with evaluative criteria that will depend on the opinion of an expert which comes pre-packaged with inevitable biases.

In fact, it is tempting to go so far as to say that figurative language is identified precisely as those instances of language where recourse to a conceptual context is necessary to interpret a lexical composition, and furthermore that the degree of figurativeness correlates with the extent of context construction involved in an interpretation. This proposition is in line with **?**'s (**?**) empirical work treating metaphor interpretation as a mechanism for classification

This, then, raises a valid question: is the role of figurative language exclusively, or even for that matter primarily, to port attributes from one conceptual domain to another? Or is what metaphor does, as **?** has famously suggested, really about something more fundamentally phenomenological than just the efficient transmission of propositions? So, where, for instance, **?** sees polysemy as an intermediate stage bridging the progress from literal to metaphoric usage, my methodology leaves itself open to the possibility that all usage is, in fact, first and foremost pragmatic, and only secondarily lexicalised. By this interpretation, words have semantic affordances in terms of their potential to convey cognitive content intersubjectively, and they are picked up and used in much the same way that a cognitive agent might adapt an object designed or just perceived as being for one purpose as an implement in another activity—using a shoe as a hammer, for example, or a chair to fend off a lion. The cognitive foregrounding of this nascent theory can be found in the ecological psychology of **?** and **?**, and the linguistic correlary seems to be in line with what psycholinguists inspired by biosemiotics such as **?** are saying about the way that language is primarily about affording cognitive value to interlocutors, including but hardly limited to truth values.

This theoretical speculation is a potential extrapolation of my methodology rather than a precondition for it, and is offered primarily as an example of how this statistical approach might become a component of productive line of philosophical enquiry. The point, though, is that with a geometric methodology, relationships between lexical semantic representations can be recast as Gibsonian affordances: there is a mechanism for the direct perception of opportunities for meaning making in the actual layout of the statistical environment

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27.

Agres, K., McGregor, S., Purver, M., and Wiggins, G. (2015). Conceptualising creativity: From distributional semantics to conceptual spaces. In *Proceedings of the 6th International Conference on Computational Creativity*, Park City, UT.

Austin, J. L. (1962). *How to do things with words*. William James Lectures. Oxford University Press.

Banjade, R., Maharjan, N., Niraula, N. B., Rus, V., and Gautam, D. (2015). Lemon and tea are not similar: Measuring word-to-word similarity by combining different methods. In *Computational Linguistics and Intelligent Text Processing - 16th International Conference*, pages 335–346.

Baroni, M., Bernardi, R., Do, N., and Shan, C. (2012). Entailment above the word level in distributional semantics. In *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32.

Baroni, M., Bernardi, R., and Zamparelli, R. (2014a). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9:241–346.

Baroni, M., Dinu, G., and Kruszewski, G. (2014b). Don't count, predict! In *ACL 2014*.

Barsalou, L., Yeh, W., Luka, B., Olseth, K., Mix, K., and Wu, L. (1993). Concepts and meaning. In Beals, K., Cooke, G., Kathman, D., McCullough, K., Kita, S., and Testen, D., editors, *Chicago Linguistics Society 29: Papers from the Parasession on Conceptual Representations*, pages 23–61. Chicago Linguistics Society, Chicago.

Barsalou, L. W. (1992). Frames, concepts, and conceptual fields. In Lehrer, A. and Kittay, E. F., editors, *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, pages 21–74. Lawrence Erlbaum Associates, Hillsdale, N.J.

Barsalou, L. W. (1993). *Theories of Memory*, chapter Flexibility, Structure, and Linguistic Vagary in Concepts: Manifestations of a Compositional System of Perceptual Symbols. Lawrence Erlbaum Associates, Hove.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Bateson, G. (1972). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. Jason Aronson Inc., London.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Birkhoff, G. (1958). Von neumann and lattice theory. *Bulletin of the American Mathematical Society*, 64:50–56.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 136–145.

Bruni, E., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49(1):1–47.

Bullinaria, J. A. and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior Research Methods*, 44(3):890–907.

Burgess, C. and Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive processes*, 12(2/3):177–210.

Carnap, R. (1947). *Meaning and Necessity: A Study in Semantics and Modal Logic*. University of Chicago Press.

Carston, R. (2010). Metaphor: Ad hoc concepts, literal meaning and mental images. In *Proceedings of the Aristotelian Society*, volume 110, pages 297–323.

Casasanto, D. and Lupyan, G. (2015). All concepts are ad hoc concepts. In Margolis, E. and Laurence, S., editors, *The Conceptual Mind: New Directions in the Study of Concepts*. MIT Press, Cambridge, MA.

Chen, D., Peterson, J. C., and Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *CoRR*, abs/1705.04416.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origins, and Use*. Praeger, New York, NY.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. MIT Press, Cambridge, MA.

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, 10(8):370–374.

Clark, S. (2015). Vector space models of lexical meaning. In Lappin, S. and Fox, C., editors, *The Handbook of Contemporary Semantic Theory*, pages 493–522. Wiley-Blackwell.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th*

International Conference on Machine Learning, pages 160–167.

de Saussure, F. (1959). *Course in General Linguistics*. The Philosophical Library, New York. edited by Charles Bally and Albert Sechehaye, trans Wade Baskin.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal for the American Society for Information Science*, 41(6):391–407.

Derrac, J. and Schockaert, S. (2015). Inducing semantic relations from conceptual spaces: A data-driven approach to plausible reasoning. *Artificial Intelligence*, 228:66–94.

Dreyfus, H. L. (2012). A history of first step fallacies. *Minds and Machines*, 22(2):87–99.

Dummett, M. (1981). *Frege: Philosophy of Language*. Duckworth, London, 2nd edition.

Erk, K. and Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 897–906.

Erk, K. and Padó, S. (2010). Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97.

Erk, K. and Smith, N. A., editors (2016). *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany.

Evans, V. (2009). *How Words Mean: Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford University Press.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transanction on Information Systems*, 20(1):116–131.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1606–1611.

Gärdenfors, P. (2000). *Conceptual Space: The Geometry of Thought*. The MIT Press, Cambridge, MA.

Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. The MIT Press.

Geffet, M. and Dagan, I. (2005). The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 107–114.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Miffline, Boston.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors,

Syntax and Semantics Volume 3: Speech Acts, pages 41–58. Academic Press, New York.

Gutiérrez, E. D., Shutova, E., Marghetis, T., and Bergen, B. K. (2016). Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Halawi, G., Dror, G., Gabrilovich, E., and Koren, Y. (2012). Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1406–1414.

Hassan, S. and Mihalcea, R. (2011). Semantic relatedness using salient semantic analysis. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 884–889. AAAI Press.

Herbelot, A. and Ganesalingam, M. (2013). Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 440–445.

Hill, F. and Korhonen, A. (2014). Learning abstract concept embeddings from multimodal data: Since you probably can't see what I mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 255–265.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with genuine similarity estimation. *Computational Linguistics*, 41(4):665–695.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 873–882.

Jäger, G. (2010). Natural color categories are convex sets. In Aloni, M., Bastiaanse, H., de Jager, T., and Schulz, K., editors, *Logic, Language and Meaning: 17th Amsterdam Colloquium, Amsterdam, The Netherlands, December 16-18, 2009, Revised Selected Papers*, pages 11–20.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Kaplan, D. (1979). On the logic of demonstratives. *Journal of Philosophical Logic*, 8(1):81–98.

Kartsaklis, D. and Sadrzadeh, M. (2016). Distributional inclusion hypothesis for tensor-based composition. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16,*

2016, Osaka, Japan, pages 2849–2860.

Kay, P. and Maffi, L. (1999). Color appearances and the emergence and evolution of basic color lexicons. *American Anthropologist*, 101(4):743–760.

Kiela, D. and Clark, S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) @ EACL 2014*, pages 21–30, Gothenburg.

Kiela, D., Hill, F., and Clark, S. (2015). Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2048.

Kornai, A., Ács, J., Makrai, M., Nemeskey, D. M., Pajkossy, K., and Recski, G. (2015). Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, *SEM 2015, June 4-5, 2015, Denver, Colorado, USA.*, pages 165–175.

Kottur, S., Vedantam, R., Moura, J. M. F., and Parikh, D. (2016). Visualword2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4985–4994.

Landauer, T., Laham, D., Rehder, B., and Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pages 412–417.

Lapesa, G. and Evert, S. (2013). Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 66–74, Sofia, Bulgaria. Association for Computational Linguistics.

Lapesa, G. and Evert, S. (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.

Levinson, S. C. (2001). Yélî dnye and the theory of basic color terms. *Journal of Linguistic Anthropology*, 10(1):3–55.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.

Levy, O., Remus, S., Biemann, C., and Dagan, I. (2015). Do supervised distributional methods really learn lexical inference relations? In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976.

Luong, T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Confer-*

ence on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013, pages 104–113.

Ma, Y., Li, Q., Yang, Z., Liu, W., and Chan, A. (2017). Learning word embeddings via context grouping. In *ACM Turing 50th Celebration Conference*.

McGregor, S., Agres, K., Purver, M., and Wiggins, G. (2015). From distributional semantics to conceptual spaces: A novel computational method for concept creation. *Journal of Artificial General Intelligence*.

Melamud, O., Dagan, I., Goldberger, J., Szpektor, I., and Yuret, D. (2014). Probabilistic modeling of joint-context in distributional similarity. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 181–190.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, pages 775–780.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119.

Mikolov, T., tau Yih, W., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 246–251.

Milajevs, D., Sadrzadeh, M., and Purver, M. (2016). Robust co-occurrence quantification for lexical distributional semantics. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 58–64, Berlin, Germany. Association for Computational Linguistics.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language & Cognitive Processes*, 6(1):1–28.

Montague, R. (1974). English as a formal language. In Thompson, R. H., editor, *Formal Philosophy: selected papers of Richard Montague*. Yale University Press, New Haven, CT.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Peirce, C. S. (1932). *Collected Papers of Charles Sanders Peirce*. Harvard University Press. edited by Charles Hartshorne and Paul Weiss.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language*

Processing.

Polajnar, T. and Clark, S. (2014). Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 230–238.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge, MA.

Putnam, H. (1975). The meaning of "meaning". In Gunderson, K., editor, *Language, Mind, and Knowledge*, pages 131–193. University of Minnesota Press.

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, pages 337–346.

Recski, G., Iklódi, E., Pajkossy, K., and Kornai, A. (2016). Measuring semantic similarity of words using concept networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 193–200, Berlin, Germany.

Riedl, M. and Biemann, C. (2013). Scaling to large^3 data: An efficient and effective method to compute distributional thesauri. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 884–890.

Rimell, L. (2014). Distributional lexical entailment by topic coherence. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg.

Rączaszek-Leonardi, J. (2012). Language as a system of replicable constraints. In Pattee, H. H. and Rączaszek-Leonardi, J., editors, *Laws, Lanuage and Life*, pages 295–333. Springer.

Rorty, R. (1979). *Philosophy and the Mirror of Nature*. Princeton University Press.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communincations of the ACM*, 8(10):627–633.

Russell, B. (1905). On denoting. *Mind*, 14(56):479–493.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. In *Proceedings of the 12th ACM SIGIR Conference*, pages 137–150.

Schütze, H. (1992). Context space. In Goldman, R., Norvig, P., Charniak, E., and Gale, B., editors, *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 113–120.

Schwartz, R., Reichart, R., and Rappoport, A. (2015). Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 258–267.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Turney, P. D. and Patel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327–352.

van der Velde, F., Wolf, R. A., Schmettow, M., and Nazareth, D. S. (2015). A semantic map for evaluating creativity. In *Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015)*, pages 94–101.

von Neumann, J. (1945). First draft of a report on the edvac. Technical report, University of Pennsylvania.

von Uexküll, J. (1957). A stroll through the worlds of animals and men: A picture book of invisible worlds. In Schiller, C. H., editor, *Instinctive Behavior: The Development of a Modern Concept*, pages 5–80. International Universities Press, Inc., New York City, NY.

Widdows, D. (2003). Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 136–143.

Widdows, D. (2004). *Geometry and Meaning*. CSLI Publications, Stanford, CA.

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

Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In Rival, I., editor, *Ordered Sets*, pages 445–470, Dordrecht/Boston. Reidel.

Wille, R. (2005). *Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies*, pages 1–33.

Yang, D. and Powers, D. M. W. (2006). Verb similarity on the taxonomy of wordnet. In *3rd International WordNet Conference*, pages 121–128.