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

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 2

Relatedness and Similarity

In Chapter **??**, I laid out the theoretical groundwork for statistical context sensitive models of lexical semantics, and in Chapter **??** I described the actual methodology for building such models, accompanied by a preliminary proof of concept involving conceptual entailment. In this chapter, I will present the first set of experiments designed to evaluate the utility of this methodology. These experiments are intended to probe the productivity of a context sensitive, geometric approach to building a computational model of lexical semantics based on statistics about word co-occurrences. Beyond testing my models' performances on some well-travelled datasets, this will provide an opportunity to explore whether different components of the methodology and, moreover, different aspects of geometric output lend themselves to modelling related but distinct semantic phenomena.

So, moving into familiar computational linguistic territory, I will explore my methodology's performance on two different phenomena: *relatedness* and *similarity*. Each of these objectives have provided reliable but distinct evaluative criteria for computational models of lexical semantics over the years, not to mention grounds for theoretical discourse. One of the hypotheses I will put forward regarding my methodology is that the geometrically replete subspaces generated by my contextualisation techniques should provide features for the simultaneous representation of related, diverse, and sometimes antagonistic aspects of language. Experimenting with these established datasets will provide a platform for exploring the ways in which different features of a semantic structure projected into one of my contextualised subspaces shift as the relationships inherent in the generation of the subspace likewise change, and this will in turn lead to some searching questions about the importance of context in the computational modelling of these particular semantic phenomena in the first place.

A fundamental objective for a general semantic model is a mechanism for measuring

the relatedness inherent in semantic representations. The distributional hypothesis itself is framed in terms of the relatedness between words: if words that tend to have a similar co-occurrence profile should also tend to have similar meaning, then, in some sense of the word, *similarity* is what is being captured by the word-vectors that populate a distributional semantic model. There is, however, an ambiguity at play in terms of what exactly it means for two words to denote things that are semantically *related*, and when this designation should include the more specific quality of *similarity* (or, for that matter, other types of relatedness such as *meronymy*, *analogy*, even *antonymy*, and so forth). So, for instance, the words *tiger*, *claw*, *stripe*, *ferocious*, and *pounce* are all clearly related in the way that they trace out aspects of a very specific conceptual space of TIGERNESS, but none of them are similar in the way that *tiger*, *lion*, and *bear* are all commensurable constituents of a space of WILD ANIMALS.

The compilation of data for the purpose of testing the ability of computational models to identify semantic relationships between words has tended to focus on the general case of relatedness rather than more nuanced similarity, if sometimes simply through a failure to specify between the two. The methodology for generating this data typically goes something like this: human participants are given a set of pairs of words and asked to quantify, for instance, the "similarity of meaning" (Rubenstein and Goodenough, 1965, p. 628) in each pair, or "how strongly these words are related in meaning," (Yang and Powers, 2006, p. 124). Finkelstein et al. (2002) use both the terms *similarity* and *relatedness* in the instructions for generating their WordSim353 data, analysed below, ultimately asking evaluators to rank words from being "totally unrelated" to "very related";[1] Bruni et al. (2012) used only the term *relatedness* in their instructions, with no mention of *similarity*. Faruqui et al. (2016) have discussed the uncertainty inherent in human ratings produced in this manner, pointing out that judgements of similarity and relatedness can be subjective and task specific, an observation which will be revisited at the end of this chapter.

Relatively recently, researchers have made a concerted effort to generate data that focusses on word similarity specifically, rather than a less clearly defined notion of relatedness. Agirre et al. (2009) have taken the widely used WordSim data and split it into two overlapping sets of word pairs, one intended to reflect a range of judgements on word similarity and the other judgements on relatedness, based on human evaluations of the types of relationships inherent in each word pair. Subsequently Hill et al. (2015) have created their SimLex999 dataset by extracting word pairs from an existing set of word associations, sampling from a range of conceptual relationships, and then

[1] Copies of the instructions, along with the data itself, can be found at `www.cs.technion.ac.il/ gabr/resources/data/wordsim353/wordsim353.zip`.

giving human evaluators detailed instructions casting similarity in terms of degree of synonymity.[2] These datasets have proven more resistant to highly accurate modelling through standard distributional semantic approaches—indeed, an interesting corollary to the distinction between relatedness and similarity has been the development of *corpus based* versus *knowledge based* techniques for modelling these semantic phenomena (see Hassan and Mihalcea, 2011; Mihalcea et al., 2006, for a discussion), with corpus based, or statistical, techniques proving more suited to modelling relatedness rather than similarity.

My thoroughly statistical methodologies will be initially tested on the WordSim data in order to explore my subspaces' capacities for capturing semantic relatedness and the SimLex data in order to explore how they handle similarity. Results for each dataset will be examined in turn, first exploring the way that human ratings can be fit to full sets of geometric features using linear models, then examining the correlation between independent features and human ratings, and finally exploring ways to learn combinations of features that should be generally predictive of the phenomena under examination. The most valuable outcome of this set of experiments, however, will be the comparison between the models learned for each of these related but distinct semantic phenomena, and in particular an analysis of the geometric features of subspaces which correlate with different measures of the conceptual interrelations between lexical representations. This meta-analysis will serve to test my hypothesis that different statistical features of an appropriately contextualised semantic space map to different semantic phenomena, and the corresponding claim that context sensitive representations can capture various semantic features as dynamic properties in a single subspace. Finally, the analysis of the different geometric correlates of relatedness and similarity will lend itself to a consideration of the way in which the frames within which humans evaluate semantic relationships may themselves be contextual.

2.1 An Experiment on Relatedness

Standard distributional semantic models have generally tended to capture semantic relatedness over similarity in terms of the proximity between semantic representations. This point, evidenced by the stronger results achieved on relatedness tests by statistical models, is elucidated by imagining the contexts in which words such as *good* and *evil* or *day* and *night* might be expected to regularly occur: there is no serious case to be made that the meaning of a sentence would not be significantly changed by toggling these word pairs in actual sentences (they are closer to being antonyms than to being synonyms), but it

[2]Instructions and data are at `https://www.cl.cam.ac.uk/ fh295/simlex.html`.

is equally reasonable to guess that these words will generally have similar co-occurrence profiles. As such, distributional semantics seems best equipped to capture the sort of broad categorical semantic relationships apparent on a syntagmatic level rather than the more fine-grained conceptual semantic relationships that emerge as we begin to consider specific axes of relatedness.

In this section, I will perform experiments on the WordSim data, which consists of 353 noun pairs rated by humans on a 0 to 10 scale for, as mentioned above, how "related" they are. Many words are involved in more than one comparison, such that the 706 word tokens in the data are spread across 439 word types. The mean word pair ranking is 5.856, with a standard deviation of 2.172. Examples of at least partially corpus derived, distributional semantic type models that have performed well on recapitulating this data include the work of Gabrilovich and Markovitch (2007) and Hassan and Mihalcea (2011), both of whom have applied vector building techniques that exploit Wikipedia page labels to enhance the conceptual knowledge inherent in their lexical representations, achieving Spearman's correlations[3] of $\rho = 0.75$ and $\rho = 0.629$ respectively. Huang et al. (2012) similarly enhance neural word embeddings derived from co-occurrence observations with synonymy information extracted from WordNet, returning a correlation of $\rho = 0.713$. A score of $\rho = 0.646$ is achieved by Luong et al. (2013) using recursive neural networks to actually delve to a level of linguistic abstraction below the word itself, modelling the morphology and the corresponding composition of words based on morphemes as a productive element in predicting relatedness between words. Radinsky et al. (2011) report $\rho = 0.80$ based on a complex model combining distributional semantic representations with detailed information about the way that phrases occur over time across historical collections of documents, and, finally, Halawi et al. (2012) achieve $\rho = 0.850$ by enhancing Radinsky et al.'s method with additional information about the relatedness between words extracted from WordNet. The overall import of this literature is that there is scope for using corpus analytic techniques to build lexical representations that do a good job of capturing semantic relatedness.

Nonetheless, there may be some advantages to identifying context specific subspaces based on an analysis of word pair inputs. For instance in cases where one of the words being compared has multiple senses, the selection of mutually relevant co-occurrence dimensions under the JOINT and ZIPPED techniques might offer a degree of disambiguation. Beyond this, I hypothesise that similar measures to the ones that have proved productive for static vector space models, so, in particular, measures of cosine similarity between word-vectors, anchored at the origin as well as at the generic vectors of the

[3]The standard approach in the empirical literature on word relatedness and similarity has been to report Spearman's correlations rather than Pearson's correlations, and I will follow suit here. The presumption is, perhaps, that word similarity is always relative—more on this in Section 2.4.

space, should be indicative of semantic relatedness. I further predict, following on the results reported at the end of the last chapter on the relationship between the norm of vectors in contextualised subspaces and conceptual entailment, that measures involving the distance of word-vectors from the origin will also correlate positively with relatedness, and here my subspaces, with their sense of interior and exterior, centre and periphery, should have an advantage.

One of the essential features of my methodology is that it is based on a statistical analysis of a corpus with minimal additional annotation. As such, one of the objectives of the experiment described in this section is to see how the performance of context sensitive models generated using the most basic level of large-scale textual data compares with models that have recourse to varying degrees of structured, hand-crafted information about conceptual relationships.

2.1.1 Relatedness: Methodology and Model

In order to test the ability of my statistical methodology to model relatedness, I build JOINT, INDY, and ZIPPED subspaces using each of the 353 word pairs in the WordSim data as input. I project subspaces of 20, 50, 200, and 400 dimensions, extrapolated from base spaces built using 2x2 and 5x5 word co-occurrence windows. For each subspace, I extract the geometric features listed in the previous chapter in Figure **??** and Table 2-H. I normalise each feature across all word pairs to have a standard normal distribution, and then I use these normalised features as the independent variables of a least squares linear regression, taking the WordSim rating of each word pair as the dependent variable. The relatedness ordering of word pairs inherent in the scores assigned by the regression are then compared to human WordSim ratings in terms of Spearman's correlations, as is standard practice in the NLP literature. Results from my model are compared with results from singular value decompositions of my base space using comparable parameters, as well as `word2vec` skip-gram and bag-of-words models, again using commensurable parameters.

Results are reported in Table 2-A. The first thing to note is that the best performance overall is achieved by the 5x5 word window, 400 dimensional version of the SVD factorisation of my base space (though the difference between this correlation and the slightly lower correlation achieved with the same parameters for the INDY dimension selection technique is not significant, with $p = .356$ based on a Fisher r-to-z transformation). More generally, the 5x5 word co-occurrence window versions of all models tend to perform more strongly on this task than the 2x2 versions, suggesting that semantic relatedness is a property of the broader sentential context in which a word occurs rather

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.666	0.681	0.698	0.728	0.704	0.698	0.700	0.709
INDY	0.671	0.676	0.702	0.707	0.703	0.712	0.715	0.729
ZIPPED	0.642	0.674	0.699	0.698	0.652	0.678	0.716	0.717
SVD	0.521	0.618	0.690	0.728	0.527	0.663	0.722	0.742
SG	0.549	0.639	0.696	0.701	0.544	0.635	0.705	0.710
CBOW	0.557	0.648	0.700	0.695	0.584	0.663	0.716	0.716

Table 2-A: Spearman's correlations for word ratings output by a linear regression model of the WordSim data for various subspace types and model parameters, compared to the correlations for cosine similarities output by static models using comparable parameters.

than just the immediate syntagmatic tendencies of a word.[4] It is also notable that my context sensitive methods outperform the static models at lower dimensionality (and here the difference is significant, with $p < .005$ in a comparison between the JOINT 5x5 window, 20 dimensional correlation and the corresponding result for the CBOW model). It seems that the contextually selected dimensions are initially all more informative about relatedness than the degree of general variance captured in lower numbers of dimensions using either factorisation or neural modelling techniques.

In terms of comparing between my dimensional selection techniques, the JOINT and INDY techniques perform somewhat comparably, with the INDY technique doing a bit better in the informationally richer 5x5 spaces in particular, where there is a higher chance of two words both having some non-zero value on a given dimension. While the results for the ZIPPED subspaces begin to tail off as dimensionality approaches 400, presumably reaching a point where the dimensions with non-zero values for both input words become generic and are no longer particularly semantically informative, the JOINT technique seems to still find traction at this dimensionality in the 2x2 word window subspaces in particular, suggesting there is still some difference between dimensions with high PMI values for both words versus one word or the other even at this depth. It's likewise interesting that the ZIPPED technique offers consistently lower correlations, particularly considering that this technique was conceived as something of a hybrid between the comprehensive JOINT approach and the independent INDY approach. It would seem, then, that the dimensions most predictive of semantic relatedness are either those which are substantially informative about both words being compared, or those which are highly informative about one word and only incidentally informative about the other, to the exclusion of the middle ground of dimensions that are highly informative about one

[4]Sahlgren (2008) discusses de Saussure's (1959) semiotic notions of *syntagm* (the way that words are composed into meaningful utterances) and *paradigm* (the way that words are comparable and potentially interchangeable units of meaning) in the context of distributional semantics.

word and at least marginally informative about another. The conclusion to draw here is that the JOINT and INDY spaces are identifying relatedness in two different capacities: in the case of the former, the degree of proximity between two points with fairly high values is being captured, while in the case of the latter the extent to which there is some degree of overlap (or, alternatively, the extent of the orthogonality) between the salient co-occurrence features is being exploited.

Something also must be said about the remarkably strong performance of the SVD models at higher dimensionalities, both in comparison to the context sensitive techniques and to the other static models. It would seem that the step of dimension-wise mean zero, standard deviation one normalisation across the factorised model has served it well in terms of capturing semantic relatedness. Any potentially adverse effects of the translation of the decomposed space, where, at relatively low dimensionality, similar word-vectors could potentially find themselves in proximate positions but on opposite sides of the origin, are ameliorated in the higher dimensional models in particular, and the basic relationships of association inherent in similar co-occurrence profiles are amplified. The overtake of the neural network models, and indeed the contextually selected models, at 400 dimensions calls to mind the comments regarding the commensurability of various distributional semantic techniques, mitigated by the rampant hyperparameterisation of such models, made by Levy and Goldberg (2014): it would seem that the application of this type of normalisation is moving towards a recapitulation of the parameterisation at play in word embedding type spaces.

2.1.2 The Geometry of Relatedness

It must at this point be noted that the context sensitive models described above are instances of fitting the output produced by my methodologies to human generated ratings, and so they should not be construed in some sense as solutions to the problem of computationally modelling the cognitive processes involved in judging semantic relatedness. Given that there are 34 different geometric features associated with any given pair of word-vectors in any subspace, there is a risk of overfitting.[5] In fact, we might speculate that we could begin to arbitrarily extract geometric features for each word-pair and eventually generate enough data to discover a correlation between geometry and human ratings to a likewise arbitrary degree of exactness. Leave-one-out cross-validation will serve to illustrate this point: by producing a relatedness score for each word pair based on coefficients learned from a linear regression of all the other word pairs, peculiarities in

[5]There is also certainly a degree of potential collinearity at play between the features, and this will be addressed below.

JOINT		INDY		ZIPPED	
$\angle AMB$	0.645	$\angle ACB$	0.721	$\angle AMB$	0.636
$\angle ACB$	0.636	$\angle AMB$	0.703	$\angle ACB$	0.607
$\mu(A,B)/M$	0.604	$\angle A'C'B'$	0.663	$\mu(A,B)$	0.603
$\mu(A,B)$	0.604	$\angle A'X'B'$	0.634	$\angle A'M'B'$	0.593
$\mu(A,B)/C$	0.603	$\angle AOB$	0.634	$\angle A'X'B'$	0.587

Table 2-B: Independent Spearman's correlations with WordSim data for top five features of each subspace type for 5x5 word co-occurrence window, 400 dimension subspaces.

the data that give a multi-variable linear model an advantage in data fitting can be eliminated. To this end, a leave-one-out validation of the 2x2 word co-occurrence window, 400 dimensional JOINT space yields a Spearman's correlation of $\rho = 0.663$, as opposed to $\rho = 0.729$ for the full linear model. To delve into this phenomenon a little further, the geometric features for 2x2 word, 400 dimensional subspaces for all three dimensional selection techinques can be concatenated into a single feature vector, resulting in an enhanced full model result of $\rho = 0.795$ but a deflated leave-one-out result of only $\rho = 0.578$. By concatenating all features of all 2x2 word window spaces into a single vector with 408 features for each word pair, a linear model can achieve a perfect Spearman's correlation, but the leave-one-out validation of models based on this amalgamation of the data gives a correlation of merely $\rho = 0.110$.

So it seems that there is a substantial risk of overfitting the data given the quantity of information being extracted from the geometry of my subspaces. In order to get a sense of what's actually happening in these models, I produce Spearman's correlations between the WordSim data and each of the features of different subspaces independently. The top five features for 400 dimensional JOINT, INDY, and ZIPPED spaces generated using 2x2 word co-occurrence windows are reported in Table 2-B. The first thing to note here is that angular measures are significantly predictive for all three dimensional selection techniques—but not the angles that may have been expected based on static distributional semantic models. Where the SVD and `word2vec` results reported in Table 2-A are based on cosine similarity between word-vectors, in my subspaces, the angles at the vertexes of the generic vectors C and M in particular seem to be predictive for all dimension selection techniques, with the measure $\angle AOB$, corresponding to cosine similarity, only figuring as the fifth most predictive feature for INDY type subspaces. All correlations here are positive, which means that words are more likely to be related as their corresponding word-vectors move closer to one another relative to their relationship to the points C and M.

On a dimension-by-dimension level, similar PMI values, or at least similar ratios of values, between word-vectors relative to both the mean values for each dimension and

the average mean across all dimensions tend to indicate semantic relatedness: words that have similar profiles of co-occurrence across the various dimensions selected by these techniques relative to these two typical statistical points are likely to denote conceptually related things. This effect is particularly pronounced in the case of INDY type subspaces, to such an extent that a single feature accounts for most of the correlation captured by the overall model (compare $\rho = 0.721$ for the feature $\angle ACB$ alone versus $\rho = 0.729$ for a model based on all features, a statistically insignificant difference with $p = 0.826$), which is particularly interesting given that each of the dimensions in these subspaces is only guaranteed to be informative about the co-occurrence tendencies of one of the two input words. So it would seem that when a collection independently selected dimensions happen to have a consistent profile of relationships between the two words used to select those dimensions and the mean value of co-occurrence statistics along each dimension, there is a strong chance the words are related.

Beyond the angular relationships between word-vectors and generic vectors, in the case of JOINT subspaces in particular, and also to a lesser extent ZIPPED subspaces, the mean norm of the word-vectors $\mu(A, B)$ correlates positively with relatedness, both alone and as the numerator of fractions where the norms of generic vectors are denominators. This corroborates the findings regarding the relationship between conceptual entailment and word-vector norm presented in Chapter **??**: in an appropriately contextualised subspace, distance from the origin is indicative of conceptual pertinence. This result can be interpreted as meaning that, in subspaces constructed from dimensions containing co-occurrence information about both words being analysed, mutually high PMI scores are indicative of higher degrees of relatedness. In other words, words that tend to have the same terms at the high end of their co-occurrence profiles also tend to be related. It is interesting, then, that this measure isn't more predictive for INDY type subspaces as well, where we might expect that the independent selection of dimensions that are informative about one word and happen to be informative about another word would indicate a strong degree of relatedness and also result in word-vectors with large norms. But these results clearly indicate that, in subspaces delineated by the concatenation of independently derived dimensions, it is the relative situation of word-vectors on these dimensions and correspondingly angular measures that point to relatedness.

It is also worth noting that, while the model learned from the 5x5 word window, 400 dimensional JOINT and ZIPPED subspaces performed well, achieving Spearman's correlations of 0.709 and 0.717 respectively, no individual feature of those subspaces proves nearly as predictive of semantic relatedness, in marked contrast to the $\angle ACB$ measure in the INDY subspaces. There are two possible explanations for this. On the one hand, there may have been a higher degree of overfitting at play in the case of the JOINT and ZIPPED

Hassan and Mihalcea (2011)	0.629
Luong et al. (2013)	0.646
$\angle ACB$	0.721
Gabrilovich and Markovitch (2007)	0.75
Radinsky et al. (2011)	0.80
Halawi et al. (2012)	0.850

Table 2-C: A comparison of Spearman's correlations returned by various models, including my optimal $\angle ACB$ measure.

subspaces. It would actually make more sense to see this effect in the INDY spaces, where the potential for selecting dimensions with unusual profiles based on a single input word, potentially leading to geometric strangeness, is higher. On the other hand, it may be the case that there is a more dynamic interaction between the various features of these spaces. This supposition will be addressed with regards to semantic similarity in particular in the next section, and then will be examined comparatively in terms of similarity and relatedness in Section 2.3.

Finally, in Table 2-C, I compare a sampling of results mentioned at the beginning of this section with the $\angle ACB$ measure in 5x5 word window, 400 dimensional INDY type subspaces. My approach is broadly within the range of results reported in the literature dealing with this dataset, but significantly below the state-of-the-art result reported by Halawi et al. (2012) ($p < .001$). It must be noted, however, that the models achieving higher scores than my own all employ techniques involving the application of structured data, in the form of, for instance, labels from Wikipedia pages (Gabrilovich and Markovitch, 2007), combining this type of labelled data with further historical information about word use (Radinsky et al., 2011), or a further enhancement of these techniques with constraints based on word relationships found in WordNet (Halawi et al., 2012). These approaches clearly return impressive results (approaching inter-annotator agreement in the strongest cases) and tell us something valuable about the ways in which word co-occurrence statistics can be productively interfaced with knowledge bases, but from a theoretical perspective I'm interested in exploring the degree to which semantically productive information can be extrapolated from data in a more raw form. Furthermore, these highly successful techniques are also inherently task specific, in the sense that the heuristic extraction of information from sources such as Wikipedia, WordNet, and so forth is targeted at identified relationships of general relatedness versus more specific aspects of word association. As previously stated, my methodology has been constructed in the hopes that the different aspects of the statistical geometry of context specific subspaces might map to different semantic phenomena. With this in mind, the next section will empirically investigate the more specific case of word similarity.

2.2 An Experiment on Similarity

In this section, I will perform experiments, similar to the ones just described for the WordSim word relatedness data, on the Simlex dataset, which, as mentioned above, has been compiled with instructions for annotators to focus specifically on semantic similarity rather than generally on semantic relatedness. The data consists of 999 word pairs, split up into nouns, verbs, and adjectives, with comparisons only called for between like parts of speech. As with the WordSim data, there are repeated words here, such that the 1,998 word tokens represent 1,028 word types. Also as with the WordSim word pairs, word pairs are rated for similarity on a scale from 0 to 10, but the average rating is 4.562, so approximately a point lower than with WordSim. Hill et al. (2015) have taken care to assemble the word pairs with consideration for the conceptual nuances of semantic similarity, choosing words intended to cover a range of both concrete and abstract concepts. There is a single word token occurring in a single word pair, the verb *disorganize*, which is not included in the vocabulary of my models (which is to say, it is not one of the 200,000 most frequent words in Wikipedia).

Where relatedness has been a fruitful target for statistical semantic modelling, word similarity has typically been the domain of models endowed with a degree of encyclopedic knowledge about the world. A Spearman's correlation of $\rho = 0.76$ with the human evaluations of the SimLex data, a result comparable with inter-annotator agreement, is achieved by Recski et al. (2016) using a statistical model enhanced with a weighted graph of conceptual relationships extracted from the `4lang` conceptual dictionary (Kornai et al., 2015). Banjade et al. (2015) similarly use a combination of statistical and knowledge based models, treating the outputs of individual models developed by various researchers as the independent variables of a range of regression models, achieving correlation of $\rho = 0.658$ in the case of the best performing model. Statistical approaches, on the other hand, have included models such as the one described by Schwartz et al. (2015), which combines `word2vec` word-vectors with vectors of syntagmatic *systematic patterns* of co-occurrence which the authors predict will be particularly indicative of semantic similarity, producing a correlation of $\rho = 0.563$. Most recently, Ma et al. (2017) return a correlation of $\rho = 0.390$ using an updated version of the `word2vec` approach which treats both independent words and groupings of words as co-occurrence terms.

In this section, I apply my own methodology to the SimLex data in order to investigate the extent to which context specific subspaces of word-vectors can accurately represent the similarity between words. As with the previous experiment exploring word relatedness, a primary objective here is to test the extent to which the geometric features of my subspaces both collectively and independently align with human ratings. In addition

to performing a linear regression mapping the full sets of geometric features generated for various combinations of parameters and likewise comparing the correlation between individual features and human similarity ratings, here I will also attempt to extract a set of features which optimally predict similarity while avoiding collinearity and without overfitting the resultant model. This approach will offer a mechanism for interpreting the dynamics at play between different features of contextualised statistical subspaces.

My hypothesis is, first and foremost, that different aspects of statistical geometry will apply to similarity than do to relatedness. In fact, if the methodology is to be even marginally successful, this will necessarily be the case, because in many instances the same word pairs have received significantly different similarity and relatedness ratings. For instance, to take a couple of examples from the small set of word pairs that occur in both the WordSim and SimLex datasets, the pair (*man, woman*) is assigned a relatedness rating of 8.30 out of 10 in the WordSim data, but only 3.33 out of 10 for the SimLex data; (*professor, student*) is likewise rated at 6.81 and 1.95 respectively. This makes sense: professors and students clearly have something to do with one another, but, within the conceptual frame of universities[6], they are different, arguably even diametric, entities. By comparison, the pair (*coast, shore*) is assigned respective scores of 9.10 and 8.83, suggesting that the words denote closely related entities, and the relationship is precisely one of similarity verging on synonymity.

2.2.1 Similarity: Methodology and Model

I initially treat the SimLex data in precisely the same way that I treated the WordSim data: I build 20, 50, 200, and 400 dimensional subspaces from 2x2 and 5x5 word co-occurrence window base spaces using the JOINT, INDY, and ZIPPED dimension selection techniques based on each word pair in the dataset. I then extract the 34 geometric features described in Table 2-H, normalising each feature to a standard normal distribution across the data for each variety of subspace. I use these normalised features as the independent variables for a least squares linear regression trained to model the human similarity ratings provided for the SimLex word pairs. Spearman's correlations between the output of this model and the human ratings on which it was trained are presented in Table 2-D.

As with the relatedness data, the INDY type subspaces once again perform very well here, and in this case notably better than the JOINT and ZIPPED subspaces, where the ZIPPED approach has a slight edge as it moves towards somewhat more independently informative dimensions. So it would seem that subspaces delineated in terms of co-

[6]The role of frames in word association judgements will be discussed in more detail in Section 2.4.

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.414	0.444	0.471	0.459	0.404	0.412	0.425	0.429
INDY	0.411	0.445	0.481	0.503	0.391	0.429	0.462	0.490
ZIPPED	0.425	0.446	0.480	0.471	0.400	0.406	0.430	0.446
SVD	0.235	0.274	0.375	0.423	0.218	0.255	0.353	0.380
SG	0.232	0.273	0.337	0.379	0.215	0.252	0.322	0.355
CBOW	0.245	0.290	0.367	0.404	0.247	0.290	0.372	0.406

Table 2-D: Spearman's correlations for word ratings output by a linear regression model of the SimLex data for various subspace types and model parameters, compared to the correlations for cosine similarities output by static models using comparable parameters.

occurrence dimensions that are definitely informative about either one or the other word being compared but only possibly informative about both collectively offer the most productive grounds for a statistical evaluation of semantic similarity. These subspaces can be seeing as something of a proving ground for similarity: in cases where words do have very similar denotations, it is likely they will independently select subspaces that are more like JOINT subspaces in that the dimensions will tend to have higher PMI values for both words even without the JOINT or ZIPPED constraints for mutual salience in place. It is also interesting to note that here, the JOINT and ZIPPED techniques do begin to trail off as dimensionality increases beyond 200 in the sparser 2x2 word window models. This is possibly an artefact of the broader range of semantic types reflected in this data, with less frequent verbs and adjectives tending to have less fleshed out co-occurrence profiles.

The most striking aspect of these results, though, is the relatively low performance of the non-contextual distributional semantic models. My own SVD model once again performs the best out of the three here, but the result of $\rho = 0.423$ for 400 dimensions generated by a 2x2 word window traversal of the corpus is substantially ($p = .023$) lower than $\rho = 0.503$ for the INDY technique with the same parameters. This corroborates a point made at the beginning of the previous section, raised by Hill et al. (2015) in their original presentation of the SimLex data, and indeed evident throughout subsequent results: where distributional semantic techniques for building lexical semantic representations do broadly capture semantic relatedness, they are less well tuned for modelling the more specific phenomenon of similarity. The two `word2vec` methods fare even worse, with the CBOW approach somewhat outperforming the SKIP-GRAM approach. This difference might again be down to the variety of semantic types at play in this data: recalling that the CBOW technique takes a fuller sample of the co-occurrence windows of vocabulary words than the SKIP-GRAM approach, we could conclude that the representations for these less frequent word types are more filled in for the CBOW models.

Finally, it is worth observing that in the case of similarity, almost across the board, the 2x2 word window models seem to outperform otherwise comparable 5x5 word window models. Hill et al. (2015) have suggested that this correlation between smaller windows and similarity pertains to adjectives and verbs in particular, and less to nouns, but the complementary effect observed in the previous section, where larger context windows tend to capture relatedness in the WordSim data, which contains only nouns, seems to suggest that there is a degree of generality to this observation. So it would seem that shared syntagmatic patterns, more overt in the terms occurring closer to a target word, are indicative of similarity in particular in addition to relatedness in general. This aligns with the findings of Kiela and Clark (2014), who report that distributional models containing information about dependency relationships are especially predictive of similarity, as well as those of Agirre et al. (2009), who achieve stronger results on their similarity focused cut of the WordSim data when they build representations based on co-occurrences with very short sequences of words rather than larger windows of co-occurrence with individual words.

2.2.2 The Geometry of Similarity

Next, as with the relatedness data in the previous experiment, in order to escape overfitting and explore the particular statistical geometry of similarity in context specific co-occurrence subspaces, I consider the predictive capacities of independent geometric features. Table 2-E reports the Spearman's correlations of the five most predicative features for each dimensional selection technique used to pick 400 dimension from a 2x2 word co-occurrence window base space. The features that independently emerge are strikingly similar to those found to be most predictive of relatedness: for the INDY subspaces, a number of different cosine measures, including angles of the vetors converging at the vertexes of generic vectors and the normalised versions of these angles, as well as the cosine similarity between the word-vectors, all correlate positively with similarity, meaning that as these angles grow smaller, the words in question tend to be more similar. Angles are also seen to be predictive of similarity in the JOINT and ZIPPED subspaces, though here the distance from the norm inherent in fractions involving $\mu(A, B)$ as the numerator are even more strongly predictive than before.

That distance from the origin should be particularly predictive of similarity in subspaces delineated by co-occurrence dimensions bearing information about both words being compared makes sense, and lines up with the hypothesis at the beginning of this chapter derived from the observation in the previous chapter that conceptual inclusion, in the appropriate contextualised co-occurrence profile, correlates with overall high PMI

JOINT		INDY		ZIPPED	
$\mu(A,B)/C$	0.377	$\angle ACB$	0.398	$\mu(A,B)/M$	0.361
$\mu(A,B)/M$	0.376	$\angle AMB$	0.375	$\mu(A,B)/C$	0.361
$\mu(A,B)/X$	0.356	$\angle A'X'B'$	0.357	$\mu(A,B)/X$	0.343
$\angle AMB$	0.349	$\angle A'C'B'$	0.351	$\angle AMB$	0.342
$\angle ACB$	0.349	$\angle AOB$	0.333	$\angle ACB$	0.325

Table 2-E: Independent Spearman's correlations with SimLex data for top five features of each subspace type for 2x2 word co-occurrence window, 400 dimension subspaces.

values. Slightly more surprising is that the most predictive measures all involve fractions with generic vectors in the denominator, and not the simple mean norm of word-vectors $\mu(A,B)$. It would seem, then, that distance from the origin is particularly predictive of similarity when it is relative to the mean and maximal values across all dimensions (and we know that there is a degree of correlation between these values, as well, as discussed in Chapter **??**). So it is not merely that these word-vectors are jointly far from the origin of their jointly selected subspaces, but moreover that they are far from the origin in comparison to the characteristic distances of other points from the origin, that indicates that they denote conceptually comparable things, processes, or descriptions.

But the most important thing to note here is that the correlation scores for these independent features are significantly lower than the scores achieved by the multi-variable linear models reported in Table 2-D. This is in contrast to the relatedness results, where the difference in correlation with human ratings achieved by the top feature and the linear model learned from all 34 features were so close that the difference was statistically insignificant. This serves first of all to reiterate a point that has already been made: where judgements of general relatedness can be extrapolated in a fairly straightforward way from a comparison of co-occurrence statistics, the more particular quality of similarity does not yield as readily to the direct quantification of co-occurrence. The critical question, then, is whether there is a combination of geometric features which, in an appropriately contextualised subspace, will reliably indicate semantic similarity between the terms used to generate that subspace—and, if so, whether we can interpret that combination of features in a way which is theoretically productive.

In order to answer this question, I perform a search of possible combinations of up to seven geometric features as the independent variables in a linear model trained to predict the SimLex word similarity ratings. I take as the objective function of the model the Spearman's correlation between the human ratings for each word pair and the corresponding scores returned by a leave-one-out cross-validation of each candidate model, where the score for each word pair is based on the coefficients learned to predict the human scores for all the other word pairs in the dataset. The state space is additionally

JOINT ($\rho = 0.417$)		INDY ($\rho = 0.434$)		ZIPPED ($\rho = 0.418$)	
$\mu(A, B)/M$	3.298	$\angle AOB$	3.467	$\mu(\overline{AC}, \overline{BC})$	-1.617
$\mu(\overline{AX}, \overline{BX})$	2.525	$\overline{AB}$	2.935	$\overline{AB}$	1.572
X	-1.797	$\angle A'M'B'$	-2.156	$\mu(A, B)/M$	1.555
$\mu(\overline{AC}, \overline{BC})$	-1.249	$\angle A'X'B'$	1.811	$\angle A'X'B'$	1.344
C/X	0.817	$\mu(A, B)/C$	-1.378	C/M	0.494
$\angle AMB$	0.397	C	-1.274	$\overline{AX} : \overline{BX}$	0.332
$\overline{A'X'} : \overline{B'X'}$	-0.343	C/M	-0.750	$\angle COX$	-0.270

Table 2-F: The optimal combination of seven non-correlated features for a linear regression modelling SimLex data for 2x2 word co-occurrence window, 400 dimensional subspaces projected using each dimensional selection technique.

constrained through the progressive application of a *variance inflation factor* (O'Brien, 2007) by which, given a set of feature vectors $\{v_1, v_2...v_i\}$, the addition of feature $i + 1$ is only considered if it satisfies the condition $1/(1 - R^2_{i+1}) < 10$ where R^2_{i+1} is the coefficient of determination of $i + 1$ as the dependent variable for a linear model based on the i established features. This constraint eliminates collinearity, which in turn results in features which are optimally informative about the relationships at play within the geometry of a type of subspace and in feature weights which are broadly interpretable in terms of their sign and scale. It also substantially trims the search space of possible combinations.

Rather than exhaustively searching the state space of combinations of features, I treat the discovery of feature combinations as a beam search problem, returning the top 1,000 performing combinations, in terms of Spearman's correlation, for each number of features progressively and then exploring the contribution of adding each of the remaining features to each of these optimal combinations. The top combinations of seven features for each dimensional selection technique, projecting 400 dimensional spaces based on the 2x2 word window base space, are detailed in Table 2-F (leave-one-out Spearman's correlations with human ratings level out with more than seven features). The Spearman's correlations reported here are once again based on a leave-one-out cross-validation, and, unlike with the relatedness data, reveal a marginally significant improvement over the best performing independent features ($p = .166$ in the case of the combined feature score for the INDY type subspaces versus $\angle ACB$ alone, the top feature reported in Table 2-E). These scores are, on the other hand, substantially lower than the scores derived from the coefficients of determination of a linear model trained on all features ($p = 0.049$). So this process of feature combination discovery reveals that, on the one hand, there is something to be gained by considering the overall statistical geometry of a subspace, and, on the other hand, there is a degree of overfitting at play in the full blown linear model.

Another striking thing about these results is the variety of features evidenced both within each subspace type and also between different subspace types. So, for instance, JOINT subspaces optimally predict similarity based on mean word-vector norms divided by average mean values ($\mu(A, B)/M$), mean distance of word-vectors from generic vectors ($\mu(\overline{AX}, \overline{BX}$, $\mu(\overline{AC}, \overline{BC})$), the norm of a generic vector (X), the ratio of the norms of generic vector (C/X), the angle at the vertex of the mean vector ($\angle AMB$), and the ratio of the distances of the word-vectors from the normalised maximum vector ($\overline{A'X'} : \overline{B'X'}$). INDY subspaces, on the other hand, make considerable use of angles, most notably the angle between the word-vectors $\angle AOB$ but also the angles at the vertexes of normalised generic vectors ($\angle A'M'B'$, $\angle A'X'B'$), as well as the actual distance between the word-vectors $\overline{AB}$, the mean norm of the word-vectors divided by the central vector ($\mu(A, B)/C$), the norm of the central vector (C), and the norm of the central vector divided by the norm of the mean vector C/M. ZIPPED subspaces, perhaps predictably, make use of a combination of the features, or at least similar features, that prove useful in analysing JOINT and INDY subspaces, with the interesting addition of the angle between the central point and the maximum point ($\angle COX$), albeit with a very low coefficient in this last case. In line with observations made above regarding the independent predictors of similarity listed in Table **??**, it seems that angles and now additionally distance between word-vectors and some generic features are the most predictive features of subspaces derived from independent analysis of input words, while the norms of word-vectors and related measures are most indicative in subspaces made up of co-occurrence dimensions jointly salient for input words.

In addition to a consideration of the optimal features themselves, there is ground to be gained by analysing the signs of the coefficient associated with these features in each linear model. It is particularly interesting to note the relationship between the angle between the word-vectors $\angle AOB$ and the distance between the word-vectors $\overline{AB}$ for INDY type subspaces. In the case of the angular measure, word-vectors are typically more similar as their cosine similarity increases, which is in line with the general hypothesis applied with standard static distributional semantic models and so is not particularly surprising. In the case of the distance measure, however, there is a likewise positive correlation, which means that words are actually expected to be more similar as the corresponding word-vectors get *further apart* (and it should be noted a similar phenomenon is observed in models learned from INDY subspaces but in the absence of the positive $\angle AOB$ measure, lest it be suggested that collinearity is in effect). This must mean that, in INDY subspaces and, to a lesser extent, ZIPPED subspaces, more similar words actually independently select, by way of high PMI values, co-occurrence dimensions that are less likely to have likewise high values to the words to which they are being compared. One explanation for this is that more similar words are simply more likely to pick less common co-occurrence

Ma et al. (2017)	0.390
INDY *combination*	0.434
Schwartz et al. (2015)	0.563
Banjade et al. (2015)	0.658
Recski et al. (2016)	0.76

Table 2-G: A comparison of Spearman's correlations with SimLex data reported for various models, including my optimal INDY technique.

dimensions, where the PMI value of the selecting word-vector is likely to be magnified by the low frequency of the dimension term in the denominator and at the same time the compared word-vector is liable to have a low or even null PMI value due to the unlikelihood of incidental co-occurrences.

Because words that come up more frequently in a corpus are more likely to acquire a broad profile of co-occurrences including a number of obscure collocations, the geometric affordances of my methodology would seem to suggest that more frequent words can be expected *prima facie* to be considered less similar words. This perhaps initially counter-intuitive claim is marginally supported by an analysis of the data, which indicates a weakly negative correlation of $\rho = -0.097$ between word frequency and similarity rating. Given that different parts of speech are known to occur at different frequencies across corpora, this trend is slightly emphasised by considering adjectives and verbs as separate categories, scoring $\rho = -0.201$ and $\rho = -0.186$ respectively. So analysis indicates that it is not necessarily the case that this frequentist axiom will prove predictive across the board, but the point is that, within some contextual frame of reference, less frequent words will tend to be considered more similar.

A cognitive explanation for the emergence of simple frequency as a predictor of similarity will be discussed in the next section; for now, this analysis is an example of how the statistical geometry of contextual subspaces offers a handle for discovering notable and unexpected tendencies in the way language occurs in a large scale corpus. The fact that more frequent words are more likely to score highly in any given similarity rating is interesting and unexpected, and cognitive explanation for this observation will be offered in the next section. More generally, though, the technique applied here gives rise to another interesting question: along with basic information about word frequency, can data about the statistical profile of a dimension alone indicate the likelihood of that dimension being in a subspace selected by input words which are predictably similar or dissimilar? I propose that the answer to this question is *yes*, and in the following section I will explore how and why this may be by way of a comparison between the statistical geometries of similarity and relatedness.

First, and finally as far as this experiment on word similarity is concerned, Table 2-G offers a comparison between a sampling of results from the literature (and it should be noted that, due to it's relatively newness, the SimLex data has not yet received as much attention as the WordSim data, though there is a growing body of relevant work emerging). Clearly approaches involving the application of heuristics, such as Schwartz et al.'s (2015) trick of mining syntactic patterns specifically indicative of similarity, Banjade et al.'s (2015) construction of a regression based on the output of a variety of models, or Recski et al.'s (2016) recourse to a structured knowledge base do significantly better than my methodology. But again, as with the relatedness experiment described in the previous section, my interest here is not merely in pursuing quantitatively strong results but also in exploring the ways in which models derived from raw word co-occurrence data can be mapped to semantic phenomena and used to explore their cognitive underpinnings (more on that in the next section). If anything, the results here indicate that similarity is clearly a complex phenomenon requiring a great deal of nuance for detection through statistical means, and an expansion of the features used to explore the words that humans deem to denote things that are alike may be in order in future work.

2.3 Comparing the Two Phenomena

The results for correlations between independent geometric features and ratings of relatedness or similarity presented in Tables 2-B and 2-E would at first pass seem to largely refute the hypothesis presented at the beginning of this chapter: the same angular and norm features predict both phenomena in similar ways in similar subspaces. Furthermore, the predictions are substantially more reliable for relatedness than they are for similarity, suggesting that these statistics reflect co-occurrence tendencies that are primarily indicative of a general pattern of semantic association and then only incidentally indicative of similarity to the extent that being similar is a special case of being related, meaning that word pairs that are similar will necessarily tend to receive higher ratings than word pairs that are unrelated. The combinations of non-correlated features obtained in Table 2-F, however, tell a slightly different story. While the best way to bluntly predict similarity based on a single statistical feature might be to guess that words that are related might also be similar, there seems to be a meaningful combination of features that collectively indicates similarity in a way not independently obvious in any of its constituents. The question, then, is whether there is a similarly dynamic and at the same time distinct combination of features indicative of relatedness.

In order to test the hypothesis that relatedness has a different set of statistical correlates than similarity, I use the same ablation technique described in the previous section

	relatedness	*similarity*
	DISTANCES	
word-vectors	-	$2.935 = \overline{AB}$
generic vectors	$X = 0.042$	$-1.274 = C$
	ANGLES	
word-vectors	$\angle ACB = 1.681$	$3.467 = \angle AOB$
normalised	$\angle A'C'B' = -0.707$	$-2.156 = \angle A'M'B'$
		$1.811 = \angle A'X'B'$
generic vectors	-	-
	MEANS	
word-vectors	$\mu(A, B) = 0.135$	-
normalised	-	-
	RATIOS	
word-vectors	$\overline{AM} : \overline{BM} = -0.100$	
normalised	$\overline{A'C'} : \overline{B'C'} = -0.308$	
	$\overline{A'X'} : \overline{B'X'} = 0.183$	
	FRACTIONS	
word-vectors	-	$-1.378 = \mu(A, B)/C$
generic vectors	-	$-0.750 = C/M$

Table 2-H: Comparison of most predictive features for relatedness and similarity in both JOINT and INDY type 2x2 word window, 400 dimensional subspaces, with models optimised for leave-one-out cross-validation.

to discover the combination of seven non-collinear features that achieve the highest Spearman's correlation for the WordSim data. The results are reported in Table 2-H. In the end, angles play an important role in predicting both phenomena, with the angle between vectors $\angle AOB$ being especially indicative of similarity: word-vectors with a similar ratio of PMI values across the set of dimensions they choose are more likely to be considered similar. The offsetting of the positive correlation with the angle $\angle ACB$, formed by the points corresponding to the word-vectors at the vertex of point C, for relatedness by the negative correlation for the angle $\angle A'C'B'$ by normalised versions of the same points suggests that related word-vectors tend to be close to one another relative to their distance from C but at the same time on either side of the central line defined by C. A similar effect can be observed for similarity, where word-vectors tend to pass on either side of the line defined by M, which can be thought of as a kind of weighted centre line, but on the same side of the potentially less central line defined by X.

The really interesting thing to note here, though, is that, outside of angular measures, the two different semantic relationships tend to be associated with different sets of geometric features. Relatedness is strongly associated with ratio type features, with the negative correlation with $\overline{A'C'} : \overline{B'C'}$ indicating that one related word tends to be significantly closer to the centre line than the other in INDY subspaces (this is also supported

by the observation above regarding the negative correlation with $\angle A'C'B'$). Returning to the mathematical analysis of Chapter **??**, the ratios involve a fraction of the norm of a vector of differences between PMI values: so, the likewise negatively correllated ratio $\overline{AM} : \overline{BM}$ involves the difference between scalars of word vectors and mean values of corresponding dimensions, so $\overline{AM} = \sqrt{\sum(A_i - M_i)^2}$ for all dimensions i in a given subspace. The difference $A_i - M_i$ is, in turn, per Equation **??**, can be understood as a logarithm of a ratio of probabilities, in this case the conditional probability of the term associated with i co-occurring with the word associated with A versus the average of all such probabilities across i. Because the values are squared, it doesn't matter which probability is the numerator and which the denominator; the important thing here is that relatedness correlates with a larger differential in the ratio of the conditional probabilities of each selected dimension co-occurring with each word and the average conditional probabilities of co-occurrence across all these dimensions. This is all to say that related words tend to choose subspaces where one of the words is considerably closer to an average co-occurrence profile than the other, which suggests that the relatedness models may be picking up on situations where an exemplar is judged related to a prototype, or a component is considered related to a whole.

Meanwhile, similar words tend to independently choose subspaces where the fraction C/M is relatively small. This observation opens the way for further statistical analysis: because C is the norm of a vector uniformly consisting of the average of the PMI values defining the vector M, C will always be less than or equal to M and will tend to be closer to M as variance in the distribution of M decreases. In other words, similar words tend to independently choose co-occurrence dimensions that together have higher variance across their mean values. Referring back to the discussion of similarity as a product of word frequency, this observation about variance suggests a related postulate that the respective co-occurrence dimensions selected by words that will be considered similar will likewise tend to diverge in terms of frequency, even as the actual words themselves become more frequent. What emerges, then, is a picture of diversity when it comes to similarity. This semantic trait is characterised by scope in terms of words which are similar and variety in terms of the terms with which those prolific words tend to co-occur, where the more general phenomenon of relatedness can be detected in terms of a tight relationship with the central region of a space.

Turning to the cognitive correlates of the frequentist quality of similarity in particular, the observations extrapolated from the geometries of my subspaces call to mind once again the notion of *framing* developed by Barsalou (1992). In maintaining that "human conceptual knowledge appears to be frames all the way down," (ibid, p. 40), Barsalou establishes a model in which framed sets of *attribute values* can be used to generatively

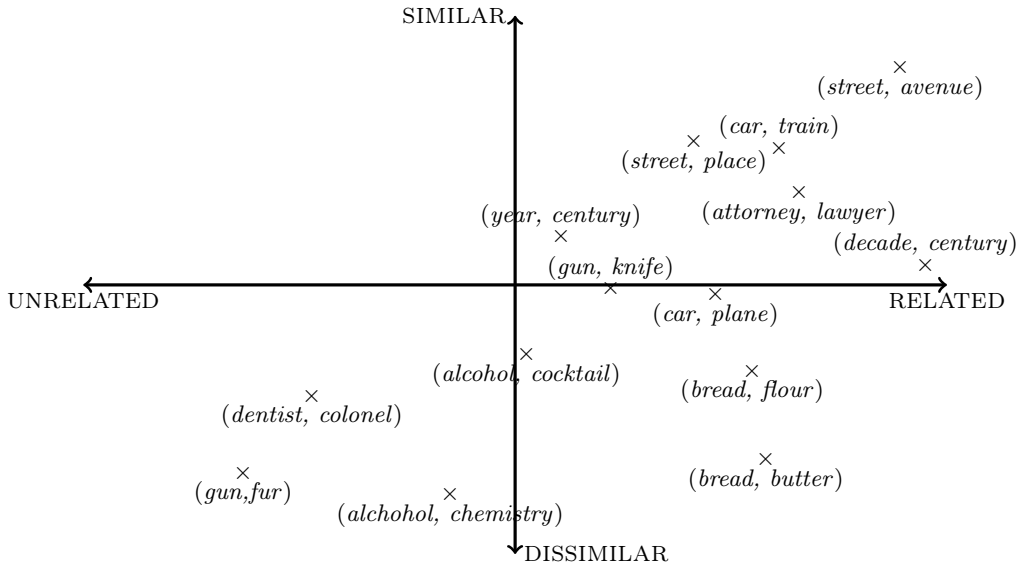


Figure 2.1: Noun pair scores along axes of relatedness and similarity as returned by a model built from features of 2x2 word co-occurrence window, 400 dimensional, INDY type subspaces.

construct conceptual exemplars, and the most typical configurations of these values within a given conceptual frame can be considered as *prototypes*. My proposal is that there is a straightforward correspondence between prototypicality and word frequency: words denoting exemplars characterised by more typical attribute values are the ones that will come up more often, and these words are in fact more likely to be considered dissimilar due to their operation as attractors for competing values along attributional dimensions. So, for instance, it is relatively easy to consider denotations of prototypical exemplars of FRUIT such as *apple* and *orange* as idiomatically opposite, whereas *pear* and *kumquat* would be considered less obviously conceptually diametric despite aligning, in terms of attributes, somewhat with *apple* and *orange* respectively. It is, then, in the dynamics of prototypes as they interact at the extents of compound conceptual fields where we discover the semantic tensions that underlie relationships of antonymy and the like, and this trend plays out in the geometries of my subspaces.[7]

Putting aside for a moment the analysis of individual features, the overall import of this comparison is to a certain extent the vindication of the hypothesis that different

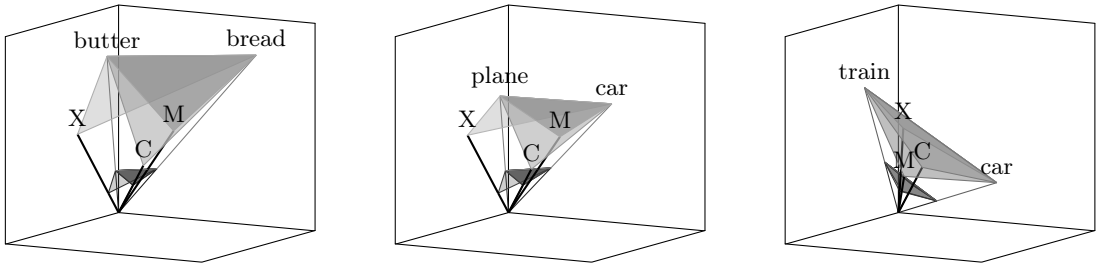
[7]Levy et al. (2015b) have similarly proposed that success in distributional semantic models capturing entailment relationships is in fact down to their ability to identify *prototypical hypernyms* that are simply more likely to be identified as categorically containing some other unseen word—but those authors do not explore whether this may in fact be a cognitively plausible approach to semantic modelling.

features are predictive of relatedness versus similarity.[8] This is illustrated in Figure 2.1, where a selection of word pairs from both the WordSim and SimLex datasets are projected along axes of relatedness and similarity based on the outputs of the respective models learned based on the geometric features of 2x2 word window, 400 dimensional INDY subspaces. So, for instance, *bread* is considered fairly related but not at all similar to *butter*; *flour* is rated as being about equally related to *bread* as *butter*, but somewhat more similar. Similar trends are observed in the progress from (*car, plane*) to (*car, train*) and (*alcohol, chemistry*) to (*alcohol, cocktail*). Meanwhile, and perhaps less explicably, *year* and *decade* are about equally similar to *century*, but *decade* is modelled as being considerably more related. The emptiness of the upper-left region of the field in this selection is characteristic of the models overall: words that are similar are in general *de facto* related to one another, but *relatedness* does not conversely predict similarity.

Figure 3.4 presents an assortment of renderings of three dimensional projections of 400 dimensional subspaces chosen from across the spectrum of both similarity and relatedness ratings as returned by the INDY technique operating on the 2x2 word window base space. The projection to three dimensions preserves the distance of the word-vectors and the generic vectors from the origin, as well as the angles between each vector, keeping the centroid vector C in the centre of the positive region of the space. It should also be noted that the norm of the vector X is scaled by a factor of 0.5 for the sake of visibility. The objective of these renderings is to offer an impression of the shifts in the overall comportment of the statistical geometry of subspaces moving along axes of both relatedness and similarity.

Moving up the scale of similarity from (*butter, bread*) to (*plane, car*), we can observe a tightening of the angle between the word-vectors and a general contractin of the space, followed by an increase in the span between the word-vectors as we ratchet our way up to the highly similar (*train, car*). An almost opposite effect can be observed, on the other hand, as relatedness increases from (*alcohol, cocktail*) to (*bread, flour*), with the word-vectors themselves looming as the angle at C contracts and the ratios of the distances to M even out. Perhaps the most interesting effect of all, though, is the visually evident similarity in the geometries of (*colonel, dentist*), which are equivalently dissimilar and unrelated, and (*train, car*), which are conversely highly similar and highly related: while my projection technique clearly struggles to accommodate the expanse of the angle between the unrelated word-vectors, the congruity of the characteristic spread of the various points in the spaces selected by the word-vectors is striking. This raises an

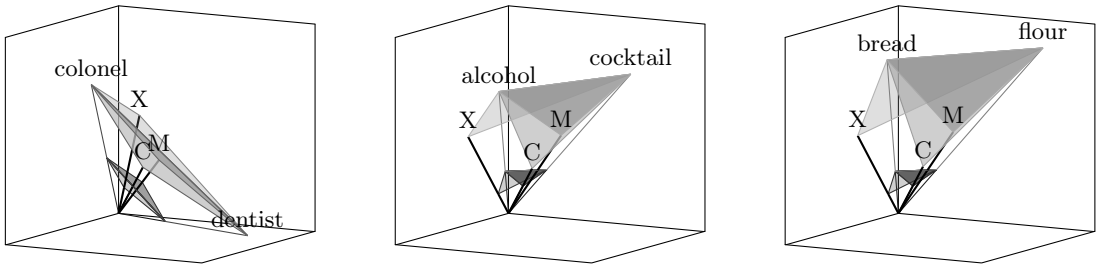
[8] Intriguingly, when identical words are given as input, they are rated as being very related and very dissimilar. The latter outcome is obviously an imperfection, but it also reveals the extent to which the models of each type of semantic phenomenon are making use of different geometric features, or the same features in opposite ways.



relatedness = 7.901
similarity = 1.760

relatedness = 7.320
similarity = 4.828

relatedness = 8.057
similarity = 7.550



relatedness = 2.646
similarity = 2.940

relatedness = 5.129
similarity = 3.714

relatedness = 7.744
similarity = 3.406

Figure 2.2: Subspaces, including word-vectors and generic features, derived from word pairs with an assortment of relatedness and similarity scores.

intriguing possibility that there may be a certain consistency in geometry based on the balance of similarity and relatedness, or, to put it differently, an indication that there is a certain shape to the statistics of a space in which similarity is the primary axis of relatedness, regardless of degree, versus a space in which there is some other specific semantic relationship in play.

2.4 Frames of Similarity

Tversky (1977), in his psychologically motivated reflections on the geometry of similarity, observes that relationships of similarity are fundamentally not symmetric: there tends to be a preference to consider the specific more similar to the general, and the peripheral more similar to the prototypical, than the other way around. So, to use Tversky's own example, *ellipse* is more similar to *circle* than *circle* is to *ellipse*; we might extend this conjecture to predict that *wolf* is more similar to *dog*, *radiologist* more similar to *doctor*, and *limping* more similar to *walking* than the converse propositions. Indeed, the frequentist axiom extrapolated through the geometric analysis of the previous section, stating that more common words denote things that are more likely to generally be a component of a similarity relationship, is broadly in line with this observation. Tversky makes the

point that the conventional conditions of geometric relationships – *minimality*, *symmetry*, and *the triangle inequality* – do not pertain in the case of similarity judgements, a point which if taken seriously serves to foil the project of a vector space model of word similarity.

Chen et al. (2017) carry this point forward experimentally, demonstrating that potential for the arbitrary construction of, for instance, analogies which demand geometrically impossible triangulations: to use one of their examples, *nurse:patient::mother:baby* is a reasonable set of relationships, as is *mother:baby::frog:tadpole*, but the proposition *nurse:patient::frog:tadpole* seems obscure at best. Chen et al. demonstrate that human raters generally identify the failure of the third set of pairings in these types of triads, whereas standard distributional semantics including `word2vec` don't—in fact, they can't, since the semantic relationships in these models are represented as static quantities. The point that emerges here is that semantic relationship emerge within a certain frame of reference, and the reason that the analogy comparing nurses to frogs fails is because both the axis of CARING that sustains the connection between nurses and mothers and the axis of PARENTAGE that connects mothers and frogs have dropped away.

The role of frames in theories of lexical semantics has already been mentioned in Chapter **??** and again earlier in this chapter. To reiterate the point raised there, Barsalou et al. (1993) propose that cognition is organised in terms of frames allowing for a *situated*, *local* representations of concepts: a concept gains its structure through a situationally specific indexing of a variety of established models. One of the consequences of this framework is that a concept emerges as the "collection of all all specialized models for a particular type of individual, together with their associated generic situations," (ibid, p. 48). So, for instance, the concept PROFESSION contains models for constituents such as DENTIST and ATTORNEY and so forth, and the conceptual scheme is structured in such a way as to offer information about the situations which both independently and jointly pertain to the models associated with those constituents. Inherent in this productive nesting of frames within frames and models in terms of their relationships to other models is the idea that concepts are specified in a particular cognitive context and generated on an *ad hoc* basis.

These types of conceptual contexts are evident in the relatedness and similarity datasets which have been explored in this chapter. In the SimLex data, for instance, (*dentist*, *colonel*) is rated as one of the least similar word pairs at 0.40, while (*attorney*, *lawyer*) is, at 9.25, considered one of the most similar pairs. The difference seems reasonable enough in terms of a comparison between the two pairs, but the low rating of (*dentist*, *colonel*) leaves little room for either dentists or colonels to be even less similar to, say, gorillas, or electricity, or democracy, and so forth. What seems to be happening here

is that human evaluators are identifying an implicit conceptual frame in which each word pair is to be evaluated: in the case of attorneys, lawyers, dentists, and colonels, the frame is something like PROFESSION, and so the professional activities of colonels and dentists are judged to be more or less orthogonal, while attorneys and lawyers pursue very similar careers. The inclusion of some additional comparison, for instance (*dentist*, *grandparent*), would suggest a broadening of the conceptual frame to something like HUMAN, and a corresponding drawing together of words denoting professions in particular.

Moreover, it is not particularly clear how a pair such as (*dentist*, *colonel*) should be considered either more or less similar to a pair like (*gun*, *fur*); the comparisons being made here seem just categorically different, and so the project of ranking the similarity of one above the other becomes a bit obscure. Instead, the task at hand really seems to be to determine the conceptual domain in which the comparison is being made, and then to make an inherently relativistic judgement about the proximity of the denotations within the semantic space of that domain. I suggest that my models are beginning to do this. By taking a subset of co-occurrence dimensions expected to exhibit a degree of saliency for either or both of the words being analysed, a subspace with a certain degree of conceptual interpretability is generated. So collectively, the 200 co-occurrence terms that are jointly most predictive of *dentist* and *colonel* also implicate *lawyer* and *attorney*, with those two words ranked 21 and 204 from the mean point of the input vectors respectively (out of a total vocabulary of 200,000), while when *lawyer* and *attorney* are used to generate a 200 dimensional subspace, *dentist* comes in at 1,925 and *colonel* at 1,096.

What begins to emerge is something like a very rough version of the conceptual spaces described by Gärdenfors (2000), in which regions of a space correspond to conceptual constituents and directions within regions can be interpreted as corresponding to values of properties that determine membership. It must be emphasised that this comparison is at a general level of abstraction: my subspaces do not at this stage contain any of the nuanced attributional information of Gärdenfors's conceptual spaces, and my methodology generates unique subspaces for each word pair, so the scores returned by the models learned through linear regression are effectively comparisons between different, albeit potentially overlapping, subspaces. Nonetheless, the reliably distinct respective predictors of relatedness and similarity within any given subspace suggest that there is already an element of conceptual structure at play in my models, even if it lacks much depth in terms of dimensional interpretation.

Faruqui et al. (2016) raise a number of issues with relatedness and similarity datasets, among them the uncertainty surrounding specific semantic phenomena and the lack of applicability of quantified word pair scores to practical NLP tasks. Those authors ultimately propose that quantitative evaluations of vector space models of word meaning

should avoid claims of generality, instead treating particular models as task specific implementations. There is something to be said for this approach, and even more to be said in support of the effort to apply statistical NLP techniques to activities in other fields where heterogeneous data and contextual complexity present potentially confounding factors to the relatively abstract and rigid representational structures of distributional semantic models. All the same, I maintain that word association tasks, particular a battery of tasks spanning a variety of semantic phenomena, can be a productive tool for exploring the capabilities of a methodology, and present the work that has been described in this chapter as a case in point.

A productive next step would be to develop methods targeting the classification of conceptual domains within which word pair comparisons are being performed, so, for instance, to identify that (*dentist*, *colonel*) and (*attorney*, *lawyer*) are both implicitly comparisons between PROFESSIONS, or at least are comparisons within the same unspecified domain. Existing work in the field on conceptual entailment may prove helpful here: Herbelot and Ganesalingam (2013), for instance, use an entropic analysis of co-occurrence statistics to conjecture about hypernymy relationships between sets of words, while Melamud et al. (2014) use a method utilising syntagmatic co-occurrence information to model the probability of words belonging to the same semantic domain. Equipped with an effective method for clustering relationships between words into conceptual domains, or alternatively for rating the degree of relevance inherent in a comparison between two relatedness judgements, my methodology offers, as has been demonstrated in the experiments reported above, a capacity for contextualising the relationships between representation in terms of co-occurrence dimensions and then discovering various geometric axes corresponding to different semantic properties. As the words used as input to define a subspace become more related, the space itself likewise becomes more conceptually coherent, and I predict that these broadly semantic axes will take on a more narrowly Gärdenforsian characteristic, allowing for interpretation as properties specific to the concept implicit in the grouping.

The INDY dimensional selection technique in particular would lend itself to this type of programmatic extension of research into semantic relatedness, as it facilitates the open-ended concatenation of dimensions from an analysis of an arbitrarily large set of constituent word-vectors (the JOINT and INDY techniques, on the other hand, would presumably return increasingly uninteresting dimensions with universally non-zero values as the set of input words expands). A subspace built using the INDY technique based on an analysis of a set of words denoting, for instance, constituents of the concept PROFESSIONALS would acquire co-occurrence dimensions specifically salient to each of the input terms, and the construal of other word-vectors in the space along the collective profile

of dimensions would, I forecast, be indicative of their conceptual situation according to the various properties of being a professional. In such a space, we might predict that we would find, for instance, *surgeon* somewhere in the vicinity of the region between *barber* and *butcher*

This proposition entails a major research project. The data for establishing groups of conceptual relationships needs to be established, and the evaluation of a model's ability to capture the attributes giving these relationships structure presents a daunting task due to the open-endedness of conceptualisation itself. Ultimately, questions of the validity of the assignment of properties to concepts, as they begin to reflect the modelling of situations in the world, are probably better suited for a qualitative analysis, and it is easy to imagine how this work might eventually lend itself to fruitful collaboration with fields such as education and the digital humanities. For now I will leave this line of enquiry where it stands, with some promising results regarding the ability of my methodology to model the overlapping semantic phenomena of relatedness and similarity in a single space. In the next chapter, I will explore my models' capacities for handling a broad and important set of semantic phenomena for which I believe it will be particularly well suited: figurative language.

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 Shutova et al. (2012), 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 Utsumi (2011) 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 Mitchell and Lapata (2010), Baroni and Zamparelli (2010), and Coecke et al. (2011), 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 Tsvetkov et al. (2014) (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 Baroni and Zamparelli (2010), 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 Lakoff and Johnson (1980) 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

[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.

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 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

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.

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 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 2-H 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

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.

the most information about variance across the corpus.

There is another possibility to consider regarding the adjectives in this dataset in particular: 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.

[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.

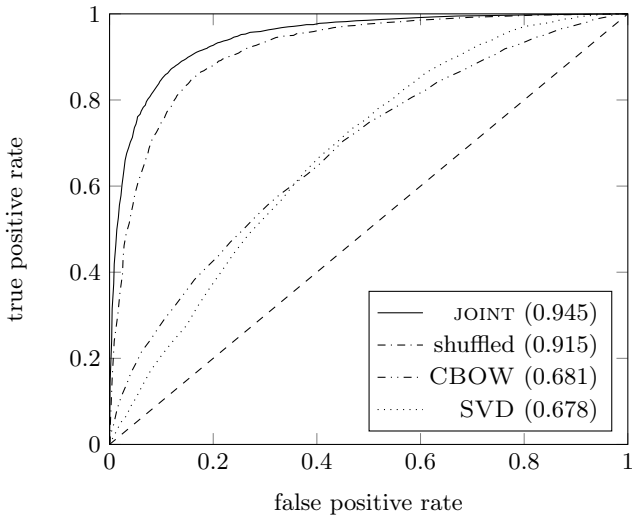


Figure 3.1: Receiver operating characteristic plots for a selection of models, with the area under the curve for each model type indicated in the legend.

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 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 Tsvetkov et al. (2014), 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

	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.

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 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 Tsvetkov et al. (2014) 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

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.

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 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

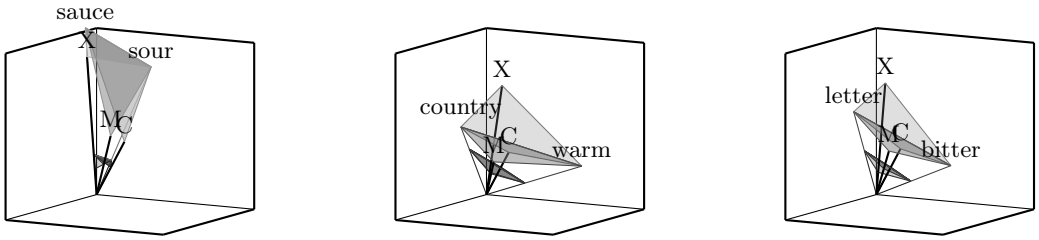
	10-fold ($f = 0.869$)	shuffled ($f = 0.830$)
	DISTANCES	
word-vectors	-	-
generic vectors	$M = -1.448$	-
	ANGLES	
word-vectors	$\angle ACB = -0.775$	-
normalised	-	-
generic vectors	$\angle COX = -1.618$	$-0.271 = \angle COM$
	$\angle COM = 0.974$	$0.045 = \angle MOX$
	MEANS	
word-vectors	$\mu(\overline{AM}, \overline{BM}) = -1.124$	$-1.007 = \mu(\overline{AC}, \overline{BC})$
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.

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 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

[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.

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 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 word is shifted through its interaction with another word: in the cases examined here, verbs that select for a particular semantic type will be seen to coerce nouns from one conceptual category to another by taking those nouns as arguments. 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 this linguistic phenotype.[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 the way in which, in the case of argument type coercion, a predicate

[6]Wittgenstein's (1967) quip regarding "grammatical fictions," (ibid, ¶307) also comes to mind.

specifies some conceptual contingency of the coerced object in order to accommodate its selectional preference. The purpose of this set of experiments is to test this broad hypothesis, and to explore the particular statistical features of co-occurrence which afford appropriate contextualisations.

3.2.1 Methodology and Results

The data which will be used to test my methodology in this section was originally presented by Pustejovsky et al. (2010) as a task for the ongoing International Workshop on Semantic Evaluation series of computational semantic modelling challenges. The data consists of 2,071 (originally split into a test set of

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, 78 contain multi-word objects not compatible with the vocabulary used for my model, reducing the total number of word pairs to 1,992, 591 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 1,029 word pairs, 399 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 2-H 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-

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.

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 $f = 0.707$ for the context sensitive models, achieved by the 5x5 word window, 200 dimensional INDY dimension selection technique, is significantly better than the baseline with

XXX significance

Of the three dimensional selection techniques that use both words as input, the INDY method achieves the overall highest scores (as opposed to the JOINT technique for metaphor), but it must be noted that these top results come at 200 dimensional subspaces selected from both 2x2 and 5x5 word window spaces, suggesting that there is a degradation in the usefulness of information included on dimensions past a certain point of saliency for a given input word. The progression of results as dimensionality increases is evident elsewhere here as well, with the single word input dimensional selection techniques as well as with the static SVD and `word2vec` models. The SVD models in particular perform erratically on this task, hinting that the angular relationships in a centred space of word-vectors which has proved effective on previous tasks provides only marginal information about the selectional relationships between predicates and objects.

In line with the metaphor results is the overall 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

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.

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 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 present, 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.

There is indeed a notable drop-off in scores across the board here, with the difference between the top INDY 400 dimensional, 2x2 word window score here and the corresponding score from the unstratified version of the data significant at

XXX

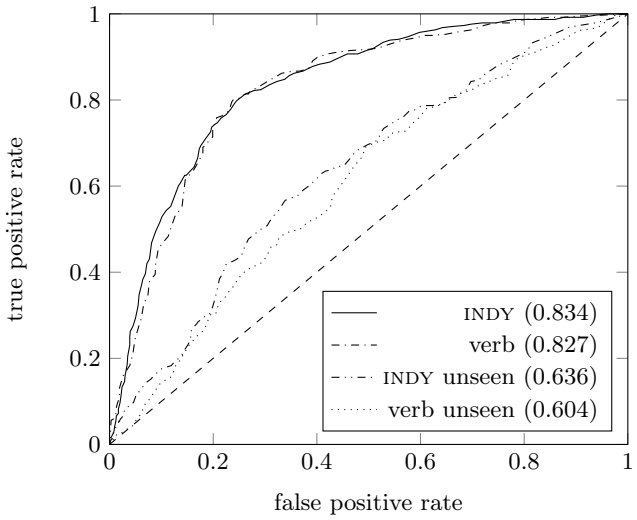


Figure 3.3: Receiver operating characteristic plots for a selection of models for coercion classification, with the area under the curve for each model type indicated in the legend.

On the other hand, the progression of scores as dimensionality increases remains jagged, with the static models particularly notable in their poor performance at higher dimensionalities. So it would seem that a great deal of what is being learned here may be specific to the verbs and the types of the arguments they take, a hypothesis supported by the relatively weak showing for the verb-only dimension selection technique. On the other hand, the verb-only, noun-only, and INDY techniques, unlike the various other methods, do now evince a steady increase in performance as dimensinoality increases, suggesting that with this rearrangement of the data these approaches are now at least discovering much of what can be classified about coercion based on co-occurrence statistics. In fact, it should be remarked that each INDY subspace is composed of the first half of the dimensions selected by the verb-only technique, combined with the first half of the noun-only subspaces, so the correspondence between these approaches isn't surprising. It is noteworthy that here subspaces built from a conjunction of dimensions associated with the two words in play are most indicative of the categorical shifting of a noun's type, rather than the subspaces formed by dimensions which are each in themselves representative of something of a conjunction in the salient co-occurrences of both words, as was the case for metaphor classification.

Figure 3.3 presents a receiver operating characteristic plot comparing between the verb-only and INDY techniques for both the regular and rearranged versions of the data, with areas under the curve indicated in the legend. As expected, the verb-only and INDY techniques are comparable for the unaltered version of the data, but the model learned from verb-only subspaces falls off once each verb stem is treated as its own fold of the

data. Of particular note is the way that the rearranged verb-only curve flattens out in the middle: the relative drop-off in true positives in the mid-range of cut-off points for classifying coercion tells us that there is a lull in the precision of the model here, with mistakes being made on the interpretation of subspaces projected by unfamiliar verbs (keeping in mind that, in the unaltered version of the data, the subspace projected by two different instances of the same verb morpheme would be identical, and so it is only the variance in the relative situation of the noun-vector in these subspaces that needs to be analysed to evaluate coercion). The relative jumpiness of the curves as compared to the smooth trajectories observed for the metaphor data in Figure 3.1 can be attributed to the scale of the data, with the massiveness of the metaphor dataset providing a steadier progression as the criteria for positive classification are relaxed. On the whole, though, the story here is a similar one of a fairly balanced advance of recall and a correspondingly steady decline in precision as the model becomes increasingly permissive in its classification of coercion.

3.2.2 The Geometry of Coercion

Following the procedure which has proved productive for the analysis of semantic phenomena in preceding experiments, I will now study the statistical geometry associated with the contextual classification of coercion, beginning with an analysis of individual features and moving on the a consideration of optimal combinations of features. In Table 3-I, I once again report the top five performing (in terms of f-score) features for each of the context sensitive dimension selection techniques described in the previous section. These features have been tested on the more prohibitive version of the coercion data rearranged to avoid training and testing on the same verb stem types, and with this in mind the improvement in scores here as compared to Table 3-H is remarkable. With the exception of the ZIPPED subspaces, all other techniques exhibit substantial improvements on the models learned from the full set of statistical features, with the difference in the verb-only spaces especially notable and a significant improvement at

XXX

Also of note is the character of the features that are most predictive for each dimensional selection technique. For all three methods involving both words as input for subspace selection, the mean values of distances at the normalised level of subspaces feature prominently (and it should be noted that the angle $\angle AOB$, which also features here, is perfectly correlated with the distance between the normalised word-vectors A' and B'). This indicates that the angles formed between the word-vectors and the generic vectors are especially associated with coercion, and this as opposed to metaphor, where

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.

the distance of various vectors from the origin as well as the ratios of these distances are particular predictive. That the averages of the angles with the generic vectors seem generally more significant than the angles between the word-vectors themselves is evidence that it is the absolute and combined situation of the word-vectors in the context of their subspaces, rather than their relationship to one another, that can be interpreted in terms of a typological semantic relationship such as coercion.

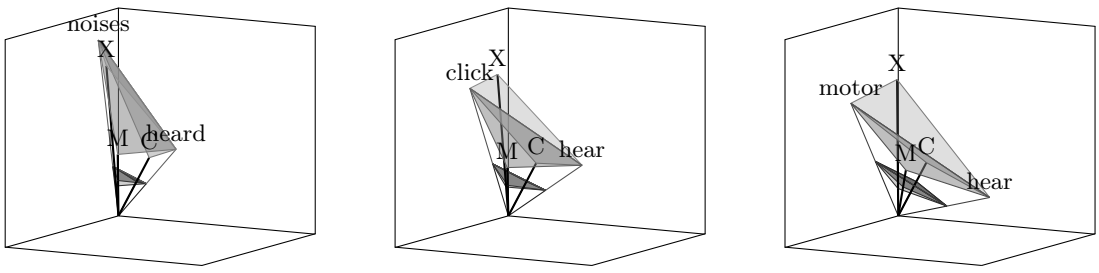
All of this is in opposition to the top features for the subspaces selected based on a single input term, where fractions and ratios are prominent. Of particular note is the significance of the relationship between the word-vectors, and this makes sense: given that the relevance of one word-vector in a subspace selected by the other is only incidental and in no way built into the space itself, differences in the relative lengths of the word-vectors and their relative distances to generic points are particularly indicative of degrees of inclusion in the profile characteristic to the dimension-selecting term. The implication is that, with a co-occurrence profile chosen by one word, it is simply the prominence of the other word with respect to this profile that is indicative of the typological relationship between the verb's selectional constraints and the noun's categorical expectation. Furthermore, the resurgence of verb-only spaces here in terms of a singular feature, namely, the fraction of the verb-centre-vector distance $\overline{AC}$ to the comparable distance $\overline{BC}$ tells us that the comparative situation of the word-vectors to the absolute centre of a subspace varies between coercive and non-coercive cases of argument selection. It's worth noting here that, in verb-selected subspaces projected from the 2x2 word window base space, co-occurrences dimensions will correspond to terms that tend to be observed in close proximity to the verbs themselves, so we can expect these dimensions

	INDY ($f = 0.681$)	VERB ($f = 0.688$)
	DISTANCES	
word-vectors	-	-
generic vectors	-	$-0.833 = X$
	ANGLES	
word-vectors	$\angle AMB = -0.564$	-
	$\angle ACB = -0.103$	
normalised	-	$0.290 = \angle A'M'B'$
generic	-	$1.241 = \angle COM$
		$-0.214 = \angle COX$
	MEANS	
word-vectors	$\mu(A, B) = 1.656$	-
normalised	-	$0.452 = \mu(\overline{A'X'}, \overline{B'X'})$
	RATIOS	
word-vectors	$\overline{AM} : \overline{BM} = 0.450$	-
normalised -	-	
	FRACTIONS	
word-vectors	-	$2.315 = \overline{AM}/\overline{BM}$
normalised	$\overline{A'M'}/\overline{B'M'} = -0.259$	-
	$\overline{A'X'}/\overline{B'X'} = 0.203$	-
generic vectors	$C/M = -1.257$	$-2.398 = C/M$

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.

to be characterised by arguments of the verbs and modifiers of those arguments: it isn't hard to imagine how, in terms of modifiers in particular, the typical characteristics of the arguments normally selected by a verb would serve as a kind of template for testing the typological fit of a new candidate argument, with relative proximity to the centre, along with the extent of the noun vector along these characteristic co-occurrence dimensions, being good metrics for determining the fit.

These independent feature results are suggestive of the types of statistics that are associated with coercion, but not of the direction of these correlations, let alone the dynamics between different statistics. To examine the geometry of coercion more in depth, I once again perform a beam search to discover the top seven features associated with both the INDY and verb-only subspace selection techniques in 400 dimensional subspaces projected from 2x2 word window base spaces, training models on the rearranged version of the data and applying a vector inflation factor in order to avoid collinearity between input features. Results are reported in Table 3-J. Remarkably, a very different picture emerges than what was observed above regarding independent features, with neither the mean distances between the norms of the INDY subspaces nor the angles and ratios individually observed in the verb-only subspaces making an appearance. In fact,



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.

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

Table 3-K: F-scores for coercion detection in full featured subspaces based on JOINT and INDY analyses of parts of speech found in each sentence containing a verb-noun pair.

one of the most notable characteristics of the respective feature vectors is, on the one hand, the spread of the

3.2.3 Adding Sentential Context

We might reasonably speculate that building word-vectors based on dependency relationships – for instance, treating the distance between words in a parse tree rather than absolute distance in a string as the boundary condition for co-occurrence window size, as **?** have proposed – might significantly enhance a model's ability to classify coercion. But this would come at the expense of building a model that doesn't have some degree of syntactic commitment already built into it, and it is likewise easy to imagine how such an approach would open itself up to accusations of tautology: if coercion as a binary case is a grammatical abstraction, then such a model would be to some extent recapitulating the

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.348	0.369	0.352	0.371	0.369	0.392	0.362	0.383
INDY	0.411	0.406	0.467	0.494	0.402	0.358	0.421	0.438

Table 3-L: F-scores for coercion identification using sentential context to generate additional subspaces and corresponding feature vectors.

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 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 Shutova's (2015) 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.

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge 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*.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for distributional semantics. *Computational Linguistics*, 36(4).

Baroni, M. and Zamparelli, R. (2010). 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.

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). Flexibility, structure, and linguistic vagary in concepts: manifestations of a compositional system of perceptual symbols. In Collins, A., Gathercole, S., and Conway, M., editors, *Theories of memory*, pages 29–101. Lawrence Erlbaum Associates, London.

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

Barwise, J. and Perry, J. (1983). *Situations and Attitudes*. MIT Press, Cambridge, MA.

Basharin, G. P., Langville, A. N., and Naumov, V. A. (2004). The life and work of a.a. markov. *Linear Algebra and its Applications*, 386:3–26. Special Issue on the Conference on the Numerical Solution of Markov Chains 2003.

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.

Birke, J. and Sarkar, A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL-06*, pages 329–336.

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

Black, M. (1955). Metaphor. In *Proceedings of the Aristotelian Society*, volume 55, pages 273–294.

Black, M. (1977). More about metaphor. In Ortony, A., editor, *Metaphor and Thought*, pages 19–41. Cambridge University Press, 2nd edition.

Boden, M. A. (1990). *The Creative Mind: Myths and Mechanisms*. Weidenfeld and Nicolson, London.

Boden, M. A. (2006). *Mind as Machine: A History of Cognitive Science*. Clarendon, Oxford.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 1st edition.

Brentano, F. (1974/1995). *Psychology from an Empirical Standpoint*. Routledge, London. Translated by Antos C. Rancurello and D. B. Terrell and Linda L. McAlister.

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.

Carnap, R. and Bar-Hillel, Y. (1952). An outline of a theory of semantic information. Technical Report 247, Research Laboratory of Electronics, MIT.

Carston, R. (2010). Metaphor: Ad hoc concepts, literal meaning and mental images. 110(3):297–323.

Carston, R. (2012). Metaphor and the literal/nonliteral distinction. In Allan, K. and Jaszczolt, K. M., editors, *The Cambridge Handbook of Pragmatics*, pages 469–492. Cambridge University Press.

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.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. The 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. (1959). A review of b. f. skinner's verbal behavior. *Language*, 35(1):26–58.

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

Church, A. (1940). A formulation of the simple theory of types. *Journal of Symbolic Logic*, 5(2):56–68.

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.

Coecke, B., Sadrzadeh, M., and Clark, S. (2011). Mathematical foundations for a compositional distributed model of meaning. *Linguistic Analysis*, 36(1-4):345–384.

Coeckelbergh, M. (2016). Can machines create art? *Philosophy & Technology*.

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.

Colton, S. (2008). Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium on Creative Intelligent Systems*.

Colton, S., Cook, M., Hepworth, R., and Pease, A. (2014). On acid drops and teardrops: Observer issues in computational creativity. In Kibble, R., editor, *Proceedings of the 50th Anniversary Convention of the AISB*.

Croft, W. and Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge University Press.

Davidson, D. (1974). On the very idea of a conceptual scheme. In *Proceedings and Addresses of the American Philosophical Association*, volume 47, pages 5–20.

Davidson, D. (1978). What metaphors mean. In *Inquiries into Truth and Interpretation*. Clarendon Press, Oxford, 2nd edition.

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

Deacon, T. W. (2011). *Incomplete Nature: How Mind Emerged from Matter*. W. W. Norton & Company, New York, NY.

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.

Dennett, D. C. (1991). *Consciousness Explained*. The Penguin Press, London.

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

Descartes, R. (1641/1911). *The Philosophical Works of Descartes*. Cambridge University Press. Translated by Elizabeth S. Haldane.

dos Santos, C. and Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, Dublin, Ireland.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. CSLI Publications.

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.

Dunn, J. (2013). Evaluating the premises and results of four metaphor identification systems. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I*, pages 471–486. Springer Berlin Heidelberg.

Eco, U. (1976). *A Theory of Semiotics*. Indiana University Press, Bloomington.

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.

Everett, D. L. (2005). Cultural constraints on grammar and cognition in pirahã: Another look at the design features of human language. *Current Anthropology*, 46(4):621–646.

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*.

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

Fauconnier, G. and Turner, M. (1998). Conceptual integration networks. *Cognitive Science*, 22(2):133–187.

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.

Firth, J. R. (1959). A synopsis of linguistic theory, 1930-55. In Palmer, F. R., editor, *Selected Papers of J. R. Firth 1952-59*. Indiana University Press.

Floridi, L. (2011). *The Philosophy of Information*. Oxford University Press.

Fodor, J. (2001). *The Mind Doesn't Work that Way: The Scope and Limits of Computational Psychology*. MIT Press.

Fodor, J. A. (2008). *LOT 2: The Language of Thought Revised*. Oxford University Press.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Fox, C. and Lappin, S. (2005). *Foundations of Intensional Semantics*. Blackwell Publishing, Oxford.

Fredkin, E. (2003). The digital perspective. *International Journal of Theoretical Physics*, 42(2):145–145.

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.

Gallie, W. B. (1956). Essentially contested concepts. *Proceedings of the Aristotelian Society*, 56:167–198.

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.

Gargett, A. and Barnden, J. (2013). Gen-meta: Generating metaphors using a combination of ai reasoning and corpus-based modeling of formulaic expressions. In *Proceedings of TAAI 2013*.

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.

Gelder, T. V. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, 92(7):345–381.

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

Grefenstette, E. and Sadrzadeh, M. (2011). Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

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.

Haugeland, J. (1993). Mind embodied and embedded. In Houng, Y. H. and Ho, J., editors, *Mind and Cognition: 1993 International Symposium*, pages 233–267. Academica Sinica.

Hegel, G. W. F. (1816/1989). *Science of Logic*. Humanities Press, Atlantic Highlands, NJ. Translated by A. V. Miller.

Heidegger, M. (1926/1962). *Being and Time*. Basil Blackwell, Oxford. translated by John Macquarrie and Edward Robinson.

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 multi-

modal 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.

Hobbes, T. (1651). *Leviathan*. Andrew Cooke.

Hoffmeyer, J. (1997). *Signs of Meaning in the Universe*. Indiana University Press.

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.

Hume, D. (1738/2000). A treatise of human nature. Oxford University Press.

Husserl, E. (1900/2001). Logical investigations, volume 1. Number v. 1 in (International library of philosophy and scientific method.). Routledge. Translated by J. N. Findlay.

Hutto, D. D. (2001). Consciousness and conceptual schema. In Pylkkänen, P. and Vadén, T., editors, *Dimensions of Conscious Experience*, pages 15–43. John Benjamins.

Indurkhya, B. (1997). Metaphor as change of representation: an artificial intelligence perspective. *Journal of Experimental & Theoretical Artificial Intelligence*, 9(1):1–36.

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.

Johnson, M. (1990). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press.

Jordanous, A. K. (2012). *Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application*. PhD thesis, University of Sussex.

Jr, R. W. G. and Tendahl, M. (2006). Cognitive effort and effects in metaphor comprehension: Relevance theory and psycholinguistics. *Mind and Language*, 21(3):379–403.

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*.

Kant, I. (1787/1996). *Critique of Pure Reason*. Hackett Publishing Company, Indianapolis, IN. Translated by Werner S. Pluhar.

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

Kartsaklis, D. and Sadrzadeh, M. (2013). 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.

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.

Kauffman, S. A. (1995). *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. Oxford University Press.

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.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, 7(2):257–266.

Koch, C. (2004). *The Quest for Consciousness: A Neurobiological Approach*. Roberts and Company.

Koestler, A. (1964). *The Act of Creation*. Hutchinson, London.

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.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. University of Chicago Press.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

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.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent

semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical Prerequisites*. Stanford University Press, Stanford, CA.

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.

Lee, M. G. and Barnden, J. A. (2001). Reasoning about mixed metaphors within an implemented artificial intelligence system. *Metaphor and Symbol*, 16(1-2):29–42.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64:354–61.

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., Goldberg, Y., and Dagan, I. (2015a). Improving distributional similarity with lessons learned from word embeddings. *Transaction of the Association for Computational Lingusitics*, 3:211–225.

Levy, O., Remus, S., Biemann, C., and Dagan, I. (2015b). 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.

Locke, J. (1689/1997). An essay concerning human understanding. Penguin, London.

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

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, 49:199–227.

Malandrakis, N., Potamianos, A., Elias, I., and Narayanan, S. S. (2013). Distributional

semantic models for affective text analysis. *IEEE Transactions on Audio, Speech and Language Processing*, 21(11):2379–2392.

Margolis, E. and Laurence, S. (2007). The ontology of concepts–abstract objects or mental representations? *Noûs*, 41(4):561–593.

Maturana, H. and Varela, F. (1987). *The Tree of Knowledge*. Shambhala, Boston, MA. Translated by Robert Paolucci.

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*.

McGregor, S., Wiggins, G., and Purver, M. (2014). Computational creativity: A philosophical approach, and an approach to philosophy. In *Proceedings of the Fifth International Conference on Computational Creativity*.

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.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439.

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.

Narayanan, S. (1999). Moving right along: A computational model of metaphoric

reasoning about events. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence*, pages 121–127.

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

Ortony, A. (1975). Why metaphors are necessary and not just nice. *Educational Theory*, 25(1):45–53.

Ortony, A., editor (1993). *Metaphor and Thought*. Cambridge University Press, 2nd edition.

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

Pantel, P. (2005). Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 125–132, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pattee, H. H. (2001). The physics of symbols: Bridging the epistemic cut. *Biosystems*, pages 5–21.

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*.

Pierce, J. R. (1980). *An Introduction to Information Theory*. Dover, New York, 2nd edition.

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language*. William Morrow.

Plato (1892). *The Republic*. Oxford University Press.

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.

Pustejovsky, J., Rumshisky, A., Plotnick, A., Jezek, E., Batiukova, O., and Quochi, V. (2010). Semeval-2010 task 7: Argument selection and coercion. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 27–32.

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.

Reimer, M. (2001). Davidson on metaphor. *Midwest Studies in Philosophy*, 25:142–155.

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.

Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17(1):67–99.

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.

Rowlands, M. (2010). *The New Science of the Mind*. The MIT Press, Cambridge, MA.

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. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.

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.

Sapir, E. (1970). *The Status of Linguistics as a Science*, pages 65–77. University of California Press.

Schütze, H. (1992a). 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.

Schütze, H. (1992b). Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE conference on Supercomputing*, pages 787–796.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

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.

Searle, J. R. (1979). Metaphor. In Ortony, A., editor, *Metaphor and Thought*. Cambridge University Press.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Shanahan, M. (2010). *Embodiment and the Inner Life: Cognition and Consciousness*

in the Space of Possible Minds. Oxford University Press.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

Shutova, E. (2010). Models of metaphor in nlp. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 688–697.

Shutova, E. (2013). Metaphor identification as interpretation. In *Proceedings of *SEM 2013*.

Shutova, E. (2015). Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Shutova, E., Teufel, S., and Korhonen, A. (2012). Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Skinner, B. F. (1957). *Verbal Behavior*. Copley Publishing Group, Acton, MA.

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.

Sowa, J. F. (2006). *Semantic Networks*. John Wiley & Sons, Ltd.

Sperber, D. and Wilson, D. (1995). *Relevance: Communication and Cognition*. Blackwell, 2nd edition.

Sperber, D. and Wilson, D. (2012). A deflationary account of metaphors. In Wilson, D. and Sperber, D., editors, *Meaning and Relevance*, pages 97–122. Cambridge University Press.

Thomas, M. S. C. and Mareschal, D. (1999). Metaphor as categorisation: A connectionist implementation. In *Proceedings of the AISB '99 Symposium on Metaphor, Artificial Intelligence, and Cognition*, University of Edinburgh.

Thompson, E. (2007). *Mind in Life*. Harvard University Press, Cambridge, MA.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 248–258. The Association for Computer Linguistics.

Turney, P. D. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, pages 491–502, London, UK, UK. Springer-Verlag.

Turney, P. D., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690.

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.

Utsumi, A. (2011). Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive Science*, 35(2):251–296.

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.

van Genabith, J. (2001). Metaphors, logic and type theory. *Metaphor and Symbol*, 16(1-2):23–57.

Veale, T. (2012). From conceptual mash-ups to bad-ass blends: A robust computational model of conceptual blending. In *Proceedings of the Third International Conference on Computational Creativity*, pages 1–8.

Veale, T. (2016). Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 34–41, San Diego, California. Association for Computational Linguistics.

Veale, T. and Hao, Y. (2007). Comprehending and generating apt metaphors: A web-driven, case-based approach to figurative language. *AAAI*, pages 1471–1476.

Veale, T., Valitutti, A., and Li, G. (2015). Twitter: The best of bot worlds for automated wit. In Streitz, N. and Markopoulos, P., editors, *Distributed, Ambient, and Pervasive Interactions: Third International Conference, DAPI*, pages 689–699. Springer International Publishing.

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.

Whitehead, A. N. and Russell, B. (1927). *Principia Mathematica*. Cambridge University Press.

Whorf, B. L. (2012). *Science and Linguistics (1940)*, pages 265–280. MIT Press.

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.

Wiggins, G. A. (2006). Searching for computational creativity. *New Generation Computing*, 24:209–222.

Wiggins, G. A. (2012). The mind's chorus: Creativity before consciousness. *Cognitive Computing*, (4):306–319.

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.

Wittgenstein, L. (1953/1967). *Philosophical Investigations*. Basil Blackwell, Oxford, 3rd edition. trans. G. E. M. Anscombe.

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

Znidarsic, M., Cardoso, A., Gervás, P., Martins, P., Hervás, R., Alves, A. O., Oliveira, H. G., Xiao, P., Linkola, S., Toivonen, H., Kranjc, J., and Lavrac, N. (2016). Computational creativity infrastructure for online software composition: A conceptual blending use case. In *Proceedings of the Seventh International Conference on Computational Creativity, UPMC, Paris, France, June 27 - July 1, 2016.*, pages 371–379.