# A Geometric Method for Context Sensitive Distributional Semantics

by

Stephen McGregor

A thesis to be submitted to the 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 compu-

tational 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

**Abstract** **i**

**Glossary** **iii**

**Table of Contents** **iv**

**List of Figures** **v**

**List of Tables** **vi**

**3 Relatedness and Similarity** **1**
3.1 An Experiment on Relatedness . . . . . 3
3.1.1 Relatedness: Methodology and Model . . . . . 4
3.1.2 The Geometry of Relatedness . . . . . 6
3.2 An Experiment on Similarity . . . . . 9
3.2.1 Similarity: Methodology and Model . . . . . 10
3.2.2 The Geometry of Similarity . . . . . 12
3.3 Comparing the Two Phenomena . . . . . 16
3.4 Frames of Similarity . . . . . 20

**References** **25**

# List of Figures

3.1 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. . . . . 19
3.2 Subspaces, including word-vectors and generic features, derived from word pairs with an assortment of relatedness and similarity scores. . . . . 20

# List of Tables


3-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. . . . . 5

3-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. . . . . 7

3-C A comparison of Spearman's correlations returned by various models, including my optimal $\angle ACB$ measure. . . . . 8

3-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. . . . . 11

3-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. . . . . 12

3-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. 13

3-G A comparison of Spearman's correlations with SimLex data reported for various models, including my optimal INDY technique. . . . . 16

3-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. . . . . 17

# Chapter 3

# 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" (**?**, p. 628) in each pair, or "how strongly these words are related in meaning," (**?**, p. 124). **?** 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] **?** used only the term *relatedness* in their instructions, with no mention of *similarity*. **?** 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. **?** 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 **?** 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 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 **??**, 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

[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`.

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

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.

## 3.1 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 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 **?** and **?**, 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. **?** 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 **?** 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. **?** report $\rho = 0.80$ based on a complex model combining distributional semantic representations with detailed informa-

[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 3.4.

tion about the way that phrases occur over time across historical collections of documents, and, finally, **?** 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 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.

### 3.1.1 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 3-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 3-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

| *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 3-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.

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 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

[4] **?** discusses **?**'s (**?**) 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.

exclusion of the middle ground of dimensions that are highly informative about one 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 **?**: 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.

### 3.1.2 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 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

[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 3-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.

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 3-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 3-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,

| | |
|---|---|
| **?** | 0.629 |
| **?** | 0.646 |
| $\angle ACB$ | 0.721 |
| **?** | 0.75 |
| **?** | 0.80 |
| **?** | 0.850 |

Table 3-C: A comparison of Spearman's correlations returned by various models, including my optimal $\angle ACB$ measure.

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 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 3.3.

Finally, in Table 3-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 **?** ($p < .001$). 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 (**?**), combining this type of labelled data with further historical information about word use (**?**), or a further enhancement of these techniques with constraints based on word relationships found in WordNet (**?**). 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.

## 3.2 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. **?** 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 **?** using a statistical model enhanced with a weighted graph of conceptual relationships extracted from the `4lang` conceptual dictionary (**?**). **?** 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 **?**, 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, **?** 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.

### 3.2.1 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 3-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 3-D.

[6]The role of frames in word association judgements will be discussed in more detail in Section 3.4.

| *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 3-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.

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-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 **?** 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

| 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 3-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.

models. **?** 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 **?**, who report that distributional models containing information about dependency relationships are especially predictive of similarity, as well as those of **?**, 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.

### 3.2.2 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 3-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

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 3-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 constrained through the progressive application of a *variance inflation factor* (**?**) 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 1000 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 3-F (leave-one-out Spearman's correlations with

| 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 3-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.

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 3-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 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.

| | |
|---|---|
| **?** | 0.390 |
| INDY *combination* | 0.434 |
| **?** | 0.563 |
| **?** | 0.658 |
| **?** | 0.76 |

Table 3-G: A comparison of Spearman's correlations with SimLex data reported for various models, including my optimal INDY technique.

First, and finally as far as this experiment on word similarity is concerned, Table 3-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 **?**'s (**?**) trick of mining syntactic patterns specifically indicative of similarity, **?**'s (**?**) construction of a regression based on the output of a variety of models, or **?**'s (**?**) 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.

## 3.3 Comparing the Two Phenomena

The results for correlations between independent geometric features and ratings of relatedness or similarity presented in Tables 3-B and 3-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 3-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.

| | *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 3-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.

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 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 3-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 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

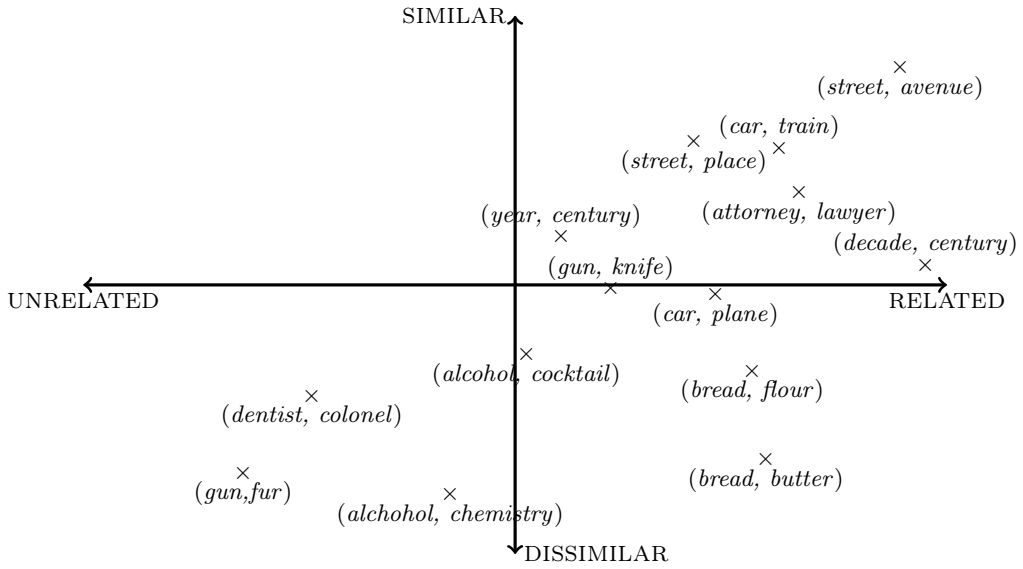


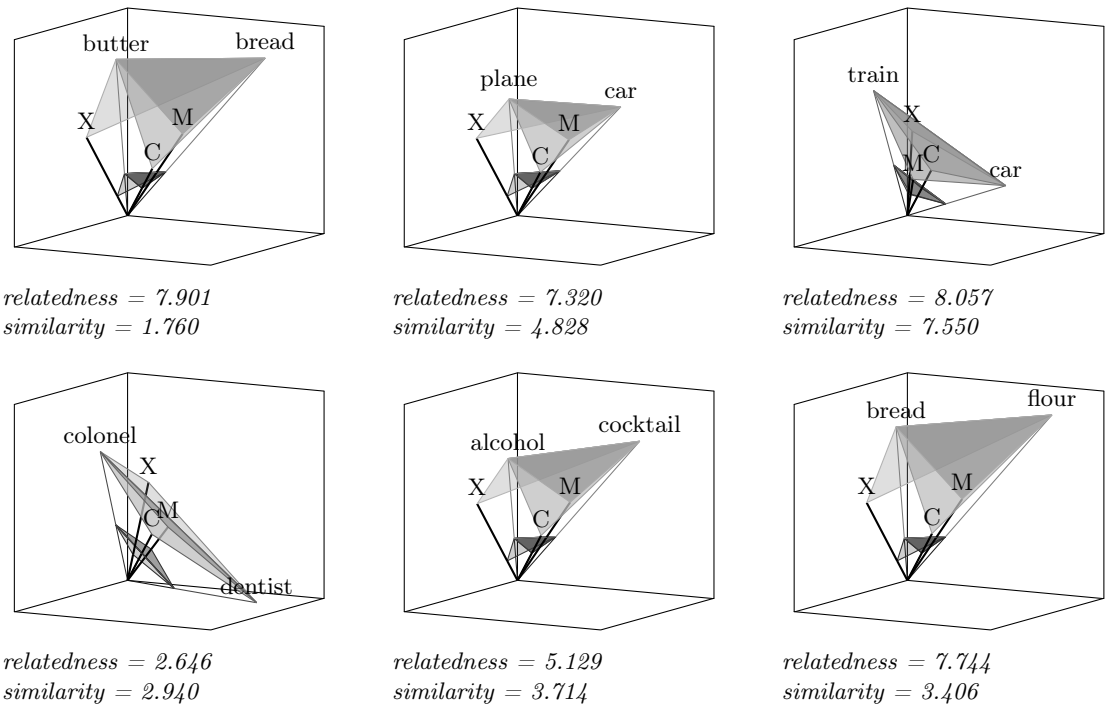

Figure 3.1: Noun pair scores along axes of relatedness and similarity as returned by a model built from features of 2x2 word co-occurrence window, 400 dimensional, INDY type subspaces.

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 features are predictive of relatedness versus similarity.[8] This is illustrated in Figure 3.1, 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

[7] **?** 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.

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

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 3.2 presents an assortment of subspaces

The contrasting

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 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.

## 3.4 Frames of Similarity

**?**, 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.

**?** 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. **?** 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, **?** 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.

EXAMPLES

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.

**?** 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: **?**, for instance, use an entropic analysis of co-occurrence statistics to conjecture about hypernymy relationships between sets of words, while **?** 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.

# References


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

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

Bouveret, M. and Sweetser, E. (2009). Multi-frame semantics, metaphoric extensions and grammar. *Annual Meeting of the Berkeley Linguistics Society*, 35(1):49–59.

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

Fraser, B. (1993). *Interpretation of novel metaphors*, pages 307–341. Cambridge University Press, 2 edition.

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

Gibbs, Jr., R. W. (1994). *The Poetics of Mind*. Cambridge University Press.

Gibbs Jr., R. W. (1993). *Process and products in making sense of tropes*, page 252âĂŞ276. Cambridge University Press, 2 edition.

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*.

Hovy, D., Srivastava, S., Kumar, S., Sachan, J. M., Goyal, K., Li, H., Sanders, W., and Hovy, E. (2013). Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57.

Jezek, E. and Hanks, P. (2010). What lexical sets tell us about conceptual categories. *Lexis*, 4:7–22.

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

Langacker, R. (1991). *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Mouton de Gruyter, Berlin.

Lapata, M. and Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):261–315.

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

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

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

Shutova, E., Kaplan, J., Teufel, S., and Korhonen, A. (2013). A computational model of logical metonymy. *ACM Transactions on Speech and Language Processing*, 10(3):11:1–11:28.

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

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *ACL (1)*, pages 248–258. The Association for Computer Linguistics.

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