[[Something Pithy]]: 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

July 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

Chapter 3

Context Sensitive Distributional Semantics

In the previous chapter, I laid down the theoretical groundwork for a distributional semantic methodology for dynamically establishing perspectives on statistical data about language use. In this chapter, I'll describe the technical details for building a computational implementation of such a methodology. The objective of this implementation is to establish a rigorous procedure for generating subspaces of word-vectors, based on observations of word co-occurrences in an underlying corpus, the geometries of which are semantically productive in particular contexts. This will involve three steps:

1. The selection, processing, and analysis of a large scale textual corpus in order to create a high dimensional base space of co-occurrence statistics;

2. The development of techniques for selecting lower dimensional subspaces based on some sort of contextualising input;

3. The exploration of the geometry of the projected subspaces in search of semantic correlates.

The following three sections will pursue each of these aspects of a technical implementation in turn. The end result is effectively a mapping from text as raw data to geometry as semiotic generator. A fourth section will describe an alternative, general interpretation of the statistical data which underwrites my models and additionally offer a brief overview of another distributional semantic methodology, both to be used as a point of comparison in the empirical results which will be discussed in subsequent chapters.

3.1 Establishing and Analysing a Large Scale Textual Corpus

The first step in a corpus based approach to natural language processing is the selection of the data which will provide the basis for our model. I've picked the English language portion of Wikipedia as my data source, a choice which is in accordance with a good deal of work done in the field. For instance, **?** and **?**, to name just a couple, use Wikipedia as their base data for training distributional semantic models designed to perform tasks similar to the ones explored in subsequent chapters, while **?**, **?**, and **?** use amalgamated corpora that include Wikipedia as a major component. Wikipedia provides a very large sample of highly regular language, meaning that we can expect a certain syntactic and semantic consistency as well as language which, if not always overtly literal, is likewise not typically abstruse or periphrastic. This should supply a source of linguistic data in which, to revisit the central dogma of the distributional hypothesis, words which occur in a particular syntactic and lexical setting can be expected to be semantically similar.

In the case of my implementations, the November 2014 dump of English language Wikipedia has been used.[1] A data cleaning process has been implemented, the first step of which is the chunking of the corpus into individual sentences. Next parenthetical phrases are removed from each sentence, as these can potentially skew co-occurrence data, and all other punctuation other than hyphenation is subsequently removed. All characters are converted into lowercase to avoid words capitalised at the beginning of sentences, quotations, and other places being considered as unique types. Finally, the articles *a*, *an*, and *the* are removed as they can distort co-occurrence distance counts, and then all sentences containing less than five words are discarded. The cleaned corpus contains nearly 1.1 billion word tokens, consisting of almost 7.5 million unique word types spread across about 61 million sentences. The distribution of these types is predictably Zipfian: over 10 million occurrences of each of the top nine word types are observed, while the least frequent 4.27 million words – more than half of all types – only occur once. The top end of this distribution is populated by conjunctions, prepositions, and pronouns, while the bottom end is characterised by obscure place names, one-off abbreviations, unicode representing non-Latin alphabet spellings, and a good many spelling errors.

As is generally the case with data cleaning, these measures are prone to error: for instance, due to the removal of punctuation, the contraction *we're* will be considered identical to the word *were*. One of the strengths of the subspace projection technique that my methodology uses is its resilience to noise. So, for instance, misspellings will be categorised as highly anomalous co-occurrence dimensions and are therefore unlikely to be contextually selected – or, if they are encountered regularly enough to be contextually significant, there may well be useful information in the co-occurrence profile of such mistakes – while, at the other end of the spectrum, essentially ubiquitous words are unlikely to provide context specific information, so the ambiguity between *we're* and *were* is unlikely to be drawn into any of the subspaces actually projected by the model.

From the cleaned corpus, a model's vocabulary is defined as the top 200,000 most

[1] Relatively recent Wikipedia dumps are available at `https://dumps.wikimedia.org/`.

frequently occurring word types. This cut-off point is very close to the point where the total number of word tokens included by selecting all instances of all vocabulary words equals the total number of word types excluded. Given the Zipfian distribution of word frequencies as observed throughout the corpus, this means that more than 95% of the co-occurrence data available from the corpus will be taken into account by the model, while the number of word-vectors used to express this data represents less than 5% of the potential vocabulary—a fairly efficient way of extrapolating statistics from the corpus. The selection of this as a cut-off point means that the least frequent words in the vocabulary occur 83 times throughout the corpus.

Having processed the corpus and established the target vocabulary, the next step of this methodology is to build up a base space of co-occurrence statistics. Here, following the example of the majority distributional semantic work, co-occurrence between a word w and another word c will be considered in terms of the number of other words between w and c. In the case of my methodology, and again in accord with the a great deal of work within the field, a statistic for word w in terms of its co-occurrence with c will be derived from the consideration of all the times that c is observed within k words to either side of w within the boundary of a sentence, where k is one of the primary model parameters that will be considered in the experiments reported in later chapters of this thesis. Based on these co-occurrence events, a matrix M is defined, where rows consist of word-vectors, one for each of the 200,000 words in the vocabulary, and columns correspond to terms with which these vocabulary words co-occur. These column-wise co-occurrence dimensions include the words in the vocabulary as well as many, many words that are not in the vocabulary, to the extent that every word type in the corpus is considered as a candidate for co-occurrence. A *pointwise mutual information* metric gauging the unexpectedness associated with the co-occurrence of two words is calculated in terms of this equation:

$$M_{w,c} := \log_2 \left(\frac{f_{w,c} \times W}{f_w \times (f_c + a)} + 1 \right) \tag{3.1}$$

Here $f_{w,c}$ represents the total number of times that c is observed as co-occurring in a sentence within k words on either side of w, f_w is the independent frequency of occurrences of w, and f_c is likewise the overall frequency of c being observed as a co-occurrence term throughout the corpus. W is the overall occurrence of all words throughout the corpus– and it should be noted that, excluding the term a, the ratio in Equation **??** is equivalent to the joint probability of w and c co-occurring. The term a is a skewing constant used to prevent highly specific co-occurrences from dominating the analysis of a word's profile, set for the purposes of the work reported here at 10,000.[2] Finally, the entire ratio is skewed by 1 so that all values returned by the logarithm will be greater than 0, with a

[2] Anecdotally, the first combination of input words analysed during an early stage of the development of this model that didn't use a smoothing constant was the phrase *musical creativity*, and the very first dimension indicated by the analysis was labelled *gwiggins*—the email handle of one of my supervisors. Prof. Wiggins's deep connection with music and creativity meant that every instance of *gwiggins* occurring throughout Wikipedia was in the vicinity of both *musical* and *creativity*, and so the dimension was indicated by its very high PMI value for each of these terms, which makes sense, but it was still a bit eerie to have such a personally relevant result generated by a model based on such general data.

value of zero therefore indicating that two words have never been observed to co-occur with one another.

This last step of incrementing the ratio of frequencies in order to avoid values tending towards negative infinity in the case of very unlikely co-occurrences is again a departure from standard practice, where, in word counting models, a *positive pointwise mutual information* mechanism involving not skewing the ratio and instead treating any ratio of frequencies less than 1 – that is, any co-occurrence that occurs with a lower probability than the combined joint probability of independently observing w and c – as being equivalent to zero (**?**, have considered a more general variable ratio shifting parameter). The motivation for this more typical technique is again to avoid incorporating unnecessary and potentially confounding information into a model, but, again, in the case of my model, the dimensional selection process will tend to ignore such information, and at the same time, as will be seen, data regarding relatively unlikely co-occurrences can sometimes also be quite informative. Other variations on the distributional semantic approach include alternative treatments of the co-occurrence window, where some researchers have taken weighted samples or considered word order (**?**), and also the processing of corpora, where part-of-speech and dependency tagging have been applied to positive effect (**?**). **?** and **?** offer comparative overviews of the effects of parameter variations on the performance of distributional semantic techniques.

The net result of my methodology is a matrix of weighted co-occurrence statistics, where higher values indicate a high number of observations of word w co-occurring with word c relative to the overall independent frequencies of w and c. Values of zero indicate words which have never been observed to co-occur in the corpus, and, as most words never co-occur with one another, the matrix is highly sparse. The weighting scheme results in a kind of semi-normalisation of the matrix: infrequent words will tend to correspond to more sparse dimensions, but the non-zero values along these dimensions will for the same reason tend to be higher due to the lower value of the word's frequency in the denominator. So far this technique sits comfortably within the scope of existing work in the field. It is what I propose to do with this base matrix that will begin to distinguish my methodology, and this next step in the process of projecting context sensitive spaces of word-vectors will be discussed in the following section.

3.2 Selecting Dimensions from a Sparse Co-Occurrence Matrix

Context has thus far remained a somewhat abstract concept in this thesis. In principle, the context in which conceptualisation occurs for a cognitive agent is its environment with all its affordances, linguistic and semantic but also more generally perceptual: in a word, the agent's *umwelt* (**?**). In the world of physical entanglements, language presents itself with precisely the same open-ended opportunities for action as other modes of cognition (**??**)—and, in the case of language, the action afforded is meaning making. In practice, however, context will be specified lexically, in terms of a word or set of words which are fed to a model, analysed in terms of their co-occurrence profiles, and then used to generate a subspace of conceptually relevant co-occurrence dimensions.

The intuition behind this approach is that there should be a set of dimensions which collectively represent a semantic tendency which can be mapped to a context, and this tendency should be discoverable in an analysis of the co-occurrence statistics of words which are exemplary of this way of talking about things.

So, notwithstanding interesting work on multi-modal approaches to distributional semantics from, for instance, **?** and **?**, with regard to the present technical description, I will treat *contextual input* as meaning some set of words T which have been selected for the purpose of performing some type of semantic evaluation and act as input to a context sensitive distributional semantic model. The exact mechanisms for specifying T will be discussed in subsequent chapters with regard to each of the individual experiments to be performed using my methodology; for now, I offer a general outline. Each component of T points to a word-vector in the matrix M described in the previous section, and the collection of word-vectors corresponding to T serve as the basis for an analysis leading to the projection of a context specific subspace S. I propose three basic techniques for generating these projections, with the model parameter d indicating the specified dimensionality of the subspace to be selected:

Joint A subspace of d dimensions with non-zero values for all elements of T and the highest mean PMI values across all elements of T is selected;

Indy The top $d/|T|$, where $|T|$ is the cardinality of T, dimensions are selected for each element of T regardless of their values for other elements of T, and then these dimensions are combined to form a subspace with dimensionality d;

Zipped The top dimensions for each element of T are selected as in the INDY technique, with the caveat that all selected dimensions must have non-zero values for all elements of T and no dimension is selected more than once.

These techniques are used for the purpose of analysis, and, once this analysis has been performed, the subset of dimensions returned is used to project the entire model vocabulary onto a d dimensional subspace. The JOINT technique requires the greatest finesse, as there is an element of cross-dimensional comparison at play. As such, for the purposes of this technique, the word-vectors selected by T are merged, dimensions with non-zero values for any of the word-vectors are discarded, and the resulting truncated word-vectors, each consisting of an equal number of non-zero dimensions, are normalised. This ensures that certain elements of T won't dominate the analysis: because the frequency of each word in T applies a deflationary pressure on the PMI values associated with the corresponding word-vectors, very infrequent words would be liable to dominate the analysis with the associated high PMI values in their profile. This effect is illustrated in Table **??**, where PMI values for the top dimensions selected using the JOINT type subspace by the words *guitar*, which at 88,285 occurrences is ranked 1541 in frequency, are compared with those for the word *dulcimer*, which occurs 516 times and is ranked 62,313 (the base model here was constructed using a 5x5 word co-occurrence window). Among the dimensions with non-zero values for both words, normalisation brings the high end of the respective co-occurrence profiles more in line with one another, facilitating the selection of a subspace which is jointly characteristic of the input terms.

	guitar			*dulcimer*		
	dimension	PMI	normalised	dimension	PMI	normalised
HIGH	*mandolin*	8.30964	0.10719	*hammered*	13.97749	0.09354
	bass	8.08501	0.10429	*dulcimer*	12.73992	0.08526
	12-string	8.07679	0.10418	*autoharp*	11.50399	0.07699
	acoustic	7.99076	0.10308	*appalachian*	11.23224	0.07517
	banjo	7.96400	0.10057	*zither*	10.98302	0.07350
LOW	attacked	0.05222	0.00067	*him*	0.25698	0.00172
	report	0.04768	0.00062	*school*	0.25340	0.00170
	country	0.04418	0.00057	*would*	0.23825	0.00159
	champions	0.02644	0.00034	*into*	0.21336	0.00143
	regions	0.02538	0.00033	*there*	0.21320	0.00143

Table 3-A: The top five and bottom five dimensions by PMI value for the words *guitar* and *dulcimer*, out of all the dimensions with non-zero values for both words, with scores tabulated independently for each word.

The intuition behind the construction of JOINT subspaces is that their dimensions should represent a profile of co-occurrences capturing the collective semantic characteristics of the contextual input. By focussing on the terms that have strong co-occurrence tendencies for all of the word-vectors indicated by the input, the expectation is that these words will occupy a central region near the perimeter of the projected subspace, and other words in this region should be likewise conceptually associated with the input. Expressed formulaically, a JOINT subspace is delineated by a set J of d dimensions generated by the contextual input T consisting of k input terms mapping to word-vectors $\{t_1, t_2...t_k\}$. These word-vectors are analysed to establish a $k \times j$ matrix N consisting of $\{n_1, n_2...n_k\}$, the vectors of T truncated such that they contain only the j dimensions with non-zero values across T:

$$n_h := \left\{ t \in t_h : \prod_{g=1}^{k} t_{g,i} > 0 \right\} \tag{3.2}$$

J is then composed by taking the d dimensions with the highest mean values across a row-wise normalisation of M:

$$J := \left\{ f_{1...d} \in \underset{f}{\operatorname{argmax}} \left(\sum_{g=1}^{k} \frac{M_{g,f}}{||m_g||} \right) \right\} \tag{3.3}$$

In the cases of the INDY and ZIPPED techniques, the selectional process is more straightforward, since mean values between features of word-vectors are not being considered. Where the JOINT technique is intended to discover subspaces that represent an amalgamation of the input terms, the INDY technique is expected to produce a subspace where individual conceptual characteristics of the input terms, captured as collections

of co-occurrence dimensions, are distilled into distinct geometric regions. So the set of d dimensions I returned by the INDY technique will delineate a subspace in which the relative geometry of contextual input word-vectors will reflect the degree to which the independent co-occurrence profiles of those word-vectors overlap. So, given the set B of all dimensions and the input word-vectors $\{t_1, t_2...t_k\}$, I can be selected from this base set of dimensions:

$$I := \left\{ binB : t_{h,b} \geq \max_{d/k} t_h \right\} \tag{3.4}$$

The ZIPPED technique might be seen as something of a hybrid of the JOINT and INDY techniques, since it used the INDY approach to make selections from the intermediary space of non-zero dimensions available to the JOINT technique. Here we know there will be some information about every co-occurrence dimensions for each word-vector associated with the contextual input, and so we might expect a subspace that offers a more nuanced interpretation of semantic relationships between the contextual input in particular. The set of dimensions Z delineating this space is selected from the same set N described in Equation **??**, in this case simply selecting the dimensions with the highest values for each input word-vector, as they have non-zero values for all the input word-vectors:

$$Z := \left\{ n \in N : t_{h,n} \geq \max_{d/k} t_h \right\} \tag{3.5}$$

An import feature of the INDY and ZIPPED techniques is that in these subspaces, rare co-occurrence dimensions of the input terms are liable to have an impact on their geometric situation when these dimensions are selected by another input word-vector, so the preservation of all co-occurrence information in my methodology might be expected to prove valuable in these cases. In each instance, these techniques are formulated to return a set of dimensions which, with varying degrees of cohesion, delineate a space that is in some sense salient to the contextual terms T serving as the basis for the analysis. In all cases, these techniques are used for the purpose of analysis, and, once this analysis has been performed, the subset of dimensions returned is used to project the entire model vocabulary onto a d dimensional subspace.

In order to offer a sense of what's happening with these dimension selection techniques, a preliminary and intuitively motivated case study of dimension selection is outlined in Table **??**, again derived from a base space generated through observations made within a 5x5 word co-occurrence window over the course of the corpus. The top dimensions selected by each technique are presented for two different three term sets of input words: *lion*, *tiger*, and *bear*, on the one hand, which are taken to represent in their union exemplars of wild animals, and on the other hand *dog*, *hamster*, and *goldfish*, which are prototypical pets. The dimensions selected by the JOINT technique in response to the WILD ANIMAL type input include the names of other wild animals, as well as *paw*, a component of many wild animals, *mauled*, an activity performed by wild animals, and, interestingly, *mascot*, presumably because many sports teams take these types of ani-

lion, tiger, bear			*dog, hamster, goldfish*		
JOINT	INDY	ZIPPED	JOINT	INDY	ZIPPED
leopard	cowardly	cowardly	pet	sled	dog
cub	crouching	sumatran	hamster	hamster	hamster
hyena	localities	grizzly	goldfish	goldfish	goldfish
sloth	rampant	tamer	hamsters	hound	pet
lion	sumatran	leopard	domesticated	djungarian	hamsters
mascot	grizzly	teddy	breed	koi	fancy
paw	wardrobe	tamarin	fancy	nassariidae	breed
tiger	leopard	tiger	pets	ovary	siberian
rhinoceros	stearns	polar	bred	carp	domesticated
mauled	teddy	passant	robotic	ednas	cat

Table 3-B: The top 10 dimensions returned using three different dimensional selection techniques, featuring one set of input terms collectively referring to wild animals and another set collectively referring to pets.

mals as their mascot: while this connection may not be immediately intuitive, it seems likely that this word would probably select for other wild animals in terms of salient features of its co-occurrence profile. The dimensions returned by the INDY technique, on the other hand, are, as expected, more independently characteristic of each of the input terms, with culturally referential words like *cowardly* (presumably from many mentions of the Cowardly Lion character from *The Wizard of Oz*) and *crouching* (indicating the context of the popular Chinese movie *Crouching Tiger, Hidden Dragon*), as well as other species-specific terms such as *sumatran* and *grizzly*. Notably, the term *stearns* pops up here, certainly because of prolific references on Wikipedia to the defunct investment bank Bear Stearns, illustrating ways in which the INDY technique might allow for dimensions indicative of underlying polysemy in some of their input terms.

Similar effects are observed in response to the PET type input. The word *pet*, two of the three input terms themselves, and the names of other types of pets appear in the output from the JOINT technique, as well as descriptive terms such as *domesticated*, *breed*, and, amusingly but not irrelevantly, *robotic*, presumably because of the phenomenon of robotic pets, which has its own page on Wikipedia. The INDY technique, on the other hand, returns some very term specific dimensions, again indicating a degree of ambiguity, such as *djungarian* (a breed of hamster popular as a house pet), *nassariidae* (in fact a species of snail, known colloquial as the *dog whelk*), and *ednas* (Edna's Goldfish was a short-lived but often cited American punk rock band). In the cases of both PETS and WILD ANIMALS, the dimensions returned by the ZIPPED technique represent something of an intermediary between the two other techniques, tending to include some of the terms generated using the JOINT technique but also some more word-specific terms. The actual geometry of these spaces will be discussed generally in the next section, and will be explored in detail in relation to specific semantic applications in subsequent chapters.

A very broadly similar approach to distributional semantics has been proposed by **?**, who describe a *context selection* methodology for generating word-vectors, involving building a base space of co-occurrence statistics and then transforming this space by

preserving only the highest values for each word-vector up to some parametrically determined cut-off point, setting all other values to zero. Setting the cut-off point relatively stringently – generating a base space of more sparse word-vectors, followed by various dimension reduction techniques – led to improvements in results on both word similarity and compositionality tests. This suggests that allowing word-vectors to shed some of their more obscure co-occurrence statistics leads to a more sharply defined semantic space, and indeed there may be an element of disambiguation at play here, as well, with vectors dropping some of the features associated with less frequent alternate word senses.

In the end, though, the method described by **?** results in a space which, while the information contained in the representation of a particular word is to a certain extent focused on the most typical co-occurrence features of that word, is still fundamentally general and static. To the extent that any contextualisation takes place here, it happens *a priori* and is cemented into a fixed spatial relationship between word-vectors. This is anathema to the theoretical grounding of my methodology, which holds that conceptual relationships arise situationally, and that semantic representations should therefore likewise come about in an *ad hoc* way. The novelty, and, I will argue, the power of my approach lies in its capacity to generate bespoke subspaces in reaction to semantic input as it emerges, and the expectation is that these subspaces will have a likewise context specific geometry which can be explored in order to discover situationally significant relationships between the projected semantic representations. The next section will begin to examine how these geometries might look.

3.3 Exploring the Geometry of a Context Specific Subspace

Before delving into the question of the types of geometries my method might be expected to generate, I would like to raise a point regarding the typical application of the term *geometry* to vector space models of distributional semantics in the first place. **?** makes an enthusiastic and compelling case for the representational power of geometry, while **?** has pointed out that treating words as geometric features endows lexical representations with "significant internal structure" (**?**, p. 509) which can be applied towards modelling the meaning making compositionality of language. **?** go so far as to suggest that their distributional semantic model effectively instantiates the abstract principles of Frege's work on the logic of natural languages (**?**) in a geometric mode. These are powerful points touching on the essence of semiotics, and the idea that representations that map from data to interpretable features in a space are core to my own methodology, as discussed in Chapter **??**.

The point I would like to make now, though, is that there are different degrees of geometry that can be in principle accessed in a vector space of real valued dimensions. The great majority of approaches surveyed here, taken to be representative of the historical and ongoing trend in the field, present models consisting of spaces of normalised word-vectors, in which there is a monotonic correlation between the distance and the angle between two word-vectors. In the case of models built using a principal component analysis, this is because when the eigenvectors of a matrix factorisation are used as

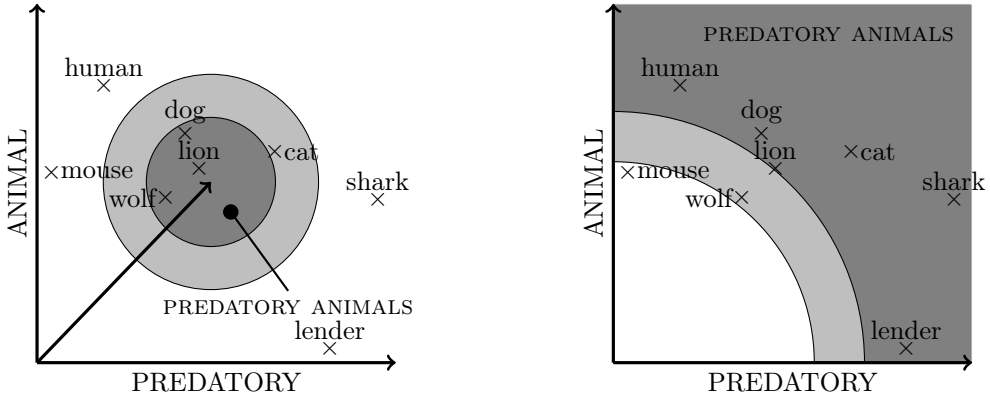
dimensions of maximal variance, there is no meaningful interpretation of the actual values along these dimensions; in fact, mean values along a dimension will tend towards zero and the signs of values along any dimension discovered through a singular value decomposition can be reversed without any degradation of the information available from analysis (**?**). So, while Euclidean distance is strictly meaningful in such a dimensional reduction, there is no sense of a centre of the space other than the centre of gravity of the data as projected onto the selected number of eigenvectors, and cosine similarity is in practice the measure used to determine the similarity between two word-vectors. And in the case of models built using neural networks, there is no meaningful interpretation of dimensions to begin with, so the resulting space is a *de facto* hypersphere of word vectors that are only relative in terms of their relationship to one another, not their relationship to any objective features of the space.

In the case of my methodology, however, precise values along dimensions, and, correspondingly, overall Euclidean distances are significant: because base dimensions are preserved in the spaces projected through any of the dimension selection techniques described above, the actual position of word-vectors in space, not just their relative situations on the surface of a normal hypersphere, are significant, with a number of potentially desirable effects. The first effect to note is that in my subspaces distance from the origin is expected to be a meaningful feature. In a subspace of contextually selected dimensions, word-vectors with strong co-occurrence tendencies for that set of dimensions should have high PMI values across all dimensions, and so a relatively high norm of a word-vector is anticipated to correspond to semantic saliency within that context.

The second effect is that there is a notion of centre and periphery in my subspaces. Since all values are positive, a word-vector with high scores across all or most dimensions in a subspace will be far from the origin and in the central region of the space. A further consequence of the positivity of these subspaces is that word-vectors with mainly low or null PMI values will be far from the centre, so in the end two word-vectors may be both close to the centre of a subspace, or at the periphery of a subspace but close to one another, or at the periphery and far from each other, at two different edges of the positively valued space, and each of these situations can be predicted to have a particular semantic interpretation. The third effect, which follows from the first two points, is that a subspace can be characterised in terms of a set of key points based on an analysis of the collective profiles of the dimensions delineating the subspace, by which I mean some straightforward assessments of the statistical distribution of each dimension involved. This aspect of my subspaces will be examined in more detail in Chapter **??**; first, though, I'll consider a couple basic measures for analysing word-vectors in context.

3.3.1 Two Measures for Probing a Subspace

In order to take a first pass at examining these robustly Euclidean features of my contextualised subspaces, I propose two geometric measures for exploring the conceptual geometry of a subspace, illustrated in Figure **??**. The first is a distance metric, which defines a central point in a subspace and then considers the relationship of words to the semantic context of the subspace in terms of the distance of the corresponding word-



(a) Word-vectors measured by proximity to a central point.

(b) Word-vectors meansured in terms of distance from origin.

Figure 3.1: Co-occurrence statistics for a small vocabulary construed along two hand-picked dimensions. Darker regions are expected to be more conceptually prototypical for the context captured by these dimensions.

vectors from this central point. The central point is defined as the mean point between the input word-vectors used to generate the subspace, or, for the purposes of Figure **??**, the central point of the eight word-vectors being analysed in this context. In this subspace featuring two hand picked co-occurrence dimensions selected from a base model built from a 5x5 word co-occurrence window traversal of Wikipedia, word-vectors relatively closely associated with the concept PREDATORY ANIMAL turn up near this central point.[3] So, for instance, cats (certainly in their taxonomical sense), more specifically lions, dogs, and, again more specifically, wolves all fall close to the central point, while sharks (certainly predators, and also animals, but perhaps less prototypically so), mice, humans, and lenders are more distant.

The second measure deployed here will be to analyse the norms of the word-vectors projected into the contextualised subspace, with my hypothesis being that word-vectors that are relatively far from the origin will be correspondingly relevant to the conceptual context from which the subspace has been generated. This prediction does not entirely play out in the subspace depicted in Figure **??**, where words like *human* and *lender* are about as far from the origin as *cat* and *shark*, and have higher norms than more prototypical denotations such as *lion* and *wolf*. As will be seen in subsequent results, beginning here and extending into the experiments described in the next chapter, in higher dimensional subspaces selected using the techniques outlined above, norm does prove to be a predictive measure of semantic relevance. Here again, the preponderance of co-occurrence statistics associated with a word over the course of a set of dimensions gives a higher dimensional subspace an advantage: if the selected dimensions are appropriately aligned, there will be a tendency for those word-vectors with some consistency of co-

[3]Here it happens to be the case that choosing dimensions which actually nominate a concept likewise delineate a space where, at least in terms of the restricted vocabulary evoked in Figure **??**, conceptual membership plays out in a geometrically predictable way, but I will not generally presume this to be the case.

lion, tiger, bear					
JOINT			INDY		
norm	distance	angle	norm	distance	angel
leopard	cat	and	leopard	wild	and
langur	wild	like	dhole	cat	as
hyena	wolf	also	hyena	giant	which
dhole	elephant	as	rhinoceros	elephant	like
boar	animals	such	leopards	lions	also
tapir	giant	well	tapir	wolf	be
macaque	animal	including	passant	animals	more
chital	bears	include	langur	tigers	including
civet	dog	from	sumatran	cats	been
sloth	panther	which	gules	golden	one

Table 3-C: The top word-vectors in subspaces selected by input terms characteristic of WILD ANIMALS, for the JOINT and INDY dimension selection techniques, measured in terms of top norms within each subspace (*norm*), word-vectors closest to the mean point between the input word-vectors (*distance*), and also the smallest angle with this mean vector regardless of actual position in the subspace (*angle*).

occurrence across all dimensions to extend towards the central fringe of the space, while those with inconsistent co-occurrence profiles will move towards the edges while remaining closer to the origin.

In the cases of both the distance from mean and norm measures, a threshold could, in principle, be established in order to determine a cut-off point for conceptual membership, either in terms of an absolute geometric measure – a radius from either the central point or the origin – or in terms of a set of nearest neighbours. This move would begin to move these subspaces towards **?**'s (**?**) notion of a region within a conceptual space, particularly in the case of the distance based metric illustrated in Figure **??**: here a clear sense of convexity as a criterion for a conceptual region exists, and likewise of betweeness as an indicator of conceptual inclusion. Importantly, though, these spaces as they stand lack the dimensional interpretability that characterises Gärdenfors's spaces, in that it is not possible to say that there is a dimension of size, or strength, or ferocity, or so forth along which a boundary for inclusion in the concept of PREDATORY ANIMAL can be identified.

Examples of the tendencies of both norms and relative distances are explored in Table **??** and Table **??**, where, as with the examples offered earlier in this chapter, input terms denoting things exemplary of the respective concepts WILD ANIMALS and PETS are used to generate subspaces, in this case using both the JOINT and INDY dimension selection techniques, once again using a base space built using a 5x5 word co-occurrence window. In these cases, the top 200 dimensions derived using each technique have been used to project subspaces, and then within those subspaces, the top ten word-vectors based on their norm and their distance from the mean point between the input word-vectors are reported. In addition to the two geometric measures described above, as a point of comparison, I also present results using an angular measure, where the word-vectors with the highest cosine similarity with the vector of the mean point between the

dog, hamster, goldfish					
JOINT			INDY		
norm	distance	angle	norm	distance	angle
hamsters	cat	and	dogs	cat	also
gerbils	pet	also	hamsters	giant	as
rabbits	monkey	as	sheepdog	animal	in
chinchillas	pig	of	terrier	wild	which
pet	rabbit	in	canine	animals	and
ferrets	rat	such	kennel	like	like
pigs	animal	well	akc	rabbit	is
rats	dogs	-	spaniel	include	called
pets	giant	called	poodle	pig	of
chickens	cats	which	jerboa	cats	has

Table 3-D: The top word-vectors in subspace, as in Table **??** but selected by input terms characteristic of PETS.

input word-vectors are returned. This is offered as an approximation of what would be a typical approach in a standard static distributional model, to demonstrate why this measure doesn't work for the context sensitive spaces built using my methodology and also as a mechanism for further exploration of what's happening in these subspaces.

Notably, in the case of the norm measure, word-vectors that are exemplary of the conceptual category suggested by the intersection of the input terms seem to rise to the top of the subspace, so to speak: for both dimension selection techniques for the WILD ANIMAL type inputs, a list of wild animals, some rather exotic, are returned. A similar outcome is observed for the norm measure in the case of the pet inputs, with some admittedly disputable admissions such as *rats* coming up in the JOINT output; jerboas, which are indicated in the INDY output, are apparently a somewhat popular pet, and *akc* presumably refers to the American Kennel Club, so, not a pet, but an institution related to pet keeping. An interesting side effect of the INDY technique in particular is that it returns a list including names of various dog breeds. It would seem that the co-occurrence dimensions of the word-vectors for *hamster* and *goldfish* are characteristic enough of these more specialised words relating to particular types of pets that the corresponding word-vectors are pushed towards the outer fringe of the subspace. It's also interesting that *passant* and *gules*, terms associated with the depiction of animals in heraldry, have high norms in the INDY subspace for WILD ANIMAL input in particular—of course all three of the input terms here are denotations of animals typical of heraldic devices, so it is not particularly surprising that some of their independently strong co-occurrence features combine to select for these word-vectors.

The distance measure returns roughly similar results, including a number of denotations of appropriate animals. Here it is interesting to observe that other semantic types – in particular, adjectives in addition to nouns – begin to creep into the output: *wild*, *giant*, and *golden* are returned in the JOINT and INDY subspaces for the WILD ANIMAL input, and *giat* again comes up in response to the PETS input, along with, perplexingly, the verb *include*. It makes sense that the region near the mean point between the input

vectors, where consistently high but perhaps not absolutely maximal PMI scores across these contextually characteristic dimensions are to be found, feature some of the descriptors and predicates associated with the concept being modelled, while the region at the outer fringe of the space, where the words with the highest overall PMI values across the dimensions of the subspace, would be pointed denotations of instances of the concepts in question. The word-vectors corresponding to some of the more esoteric animals in particular are likely to have high co-occurrence frequencies with the same dimensions selected by the combination of the input terms relative to low independent frequencies precisely because of their rareness.

Turning to the angular results, where words that are closest to the line extending through the mean point are returned, a sharp contrast to the other two geometric measures is observed. Here, very generic words which serve as the structural components of language, contributing little in terms of specific meaning but crucial to the functional cohesion of an utterance, are found in abundance. This is completely logical: these types of words are liable to have a very consistent, albeit relatively low, profile of PMI scores across all dimensions in a subspace, since they are likely to have a high frequency of co-occurrences with any given word mitigated by a correspondingly high independent frequency across the corpus influencing the denominator of the PMI calculation. The result is a word-vector populated by relatively low but also relatively consistent PMI values, situated not far from the origin and also very close to the centre line of the subspace. This phenomenon highlights the discrepancy between the Euclidean, positively valued subspaces generated by my context sensitive methodology and the normalised, hyperspherical spaces built by conventional static distributional semantic models. Because my subspaces have a sense of centre and periphery, as well as a sense of distance from the origin, it is possible to make both semantic and functional predictions about the types of words that will be found in different regions of a subspace, and accordingly to predict where to look – and where not to look – to discover geometries mapping to desired conceptual properties.

3.3.2 Replete Geometric Analysis

I will now propose a general method for a replete geometric analysis of a contextually projected subspace, based on the position of word-vectors in a space as well as the relationship between those word-vectors and points based on a more general analysis of the dimensions delineating the subspace which I will characterise as *generic points*. For the purposes of explicating this method, I will presume a subspace projected from an analysis of two input word-vectors A and B using one of the dimension selection techniques described earlier in this chapter, a presumption in line with the experiments to be described in Chapters **??** and **??**. The premise is that these word vectors are to be analysed in terms of their semantic relationship; the precise nature of the relationship being analysed could be more or less anything, and in the next two chapters this method will be applied to the assessment of lexical similarity, relatedness, metaphor, and metonymy. The objective of this analytic method will be first to test the hypothesis that the geometry of contextually projected subspaces should be semantically informative, and second to compare the aspects of the geometry that are most informative for different semantic

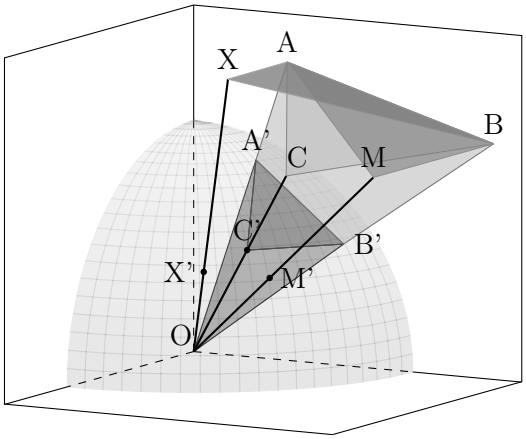


Figure 3.2: The geometric features of a subspace contextually projected based on an analysis of two input word-vectors.

phenomena.

Figure **??** illustrates a generic three dimensional subspace, with point O as the origin. Points A and B are the two word-vectors that have been used to select the dimensions which define this subspace, and are likewise the word-vectors which will be analysed through the geometry of the subspace. In addition to these two points explicitly defined in terms of the values of projected word-vectors, two points are established based on an overall analysis of the dimensionality of the subspace: the *mean point* M and the *maximal point* X. M is defined as the vector of all the mean values for all the dimensions J delineating the subspace, so, if the dimensionality of J is d, M can be defined formally as follows:

$$M = \{\mu(J_1),\ \mu(J_2)...\ \mu(J_d)\} \tag{3.6}$$

And likewise, X can be expressed in terms of an equation:

$$X = \{max(J_1),\ max(J_2)...\ max(J_d)\} \tag{3.7}$$

Finally, a generic central point C, a vector with all dimensions set to the same value, is defined. The universal value chosen to define the dimensions of this vector is the mean value of the mean point M, so, formally, this point is the vector of that mean value repeated d times:

$$C = \{\mu(M),\ \mu(M)...\mu(M)\} \tag{3.8}$$

In the analysis of the semantic relationship between A and B in a given projection, these three vectors will be used as anchor points to establish the situation of A and B relative to the subspace overall: where C is an objectively central point in the subspace, M is

	MEAN	MAX
TOP	*sofla:* 6.984	*nico:* 15.690
	olya: 6.326	*yeah:* 15.610
	non-families: 6.035	*superfamily:* 15.598
	gmina: 5.364	*eel:* 15.483
	crambidae: 5.485	*kermanshah:* 15.455
BOTTOM	*it:* 0.748	*he:* 3.903
	they: 0.812	*in:* 3.449
	you: 0.804	*of:* 3.379
	this: 0.789	*to:* 3.120
	he: 0.719	*and:* 2.993
mean	2.312	11.066
std	0.396	1.607

Table 3-E: Dimensional profiles in terms of mean and maximum PMI values along dimensions, including mean values and standard deviation as well as the top five and bottom five dimensions for each statistic.

in a sense central to a subspace relative to its particular dimensional constitution, and X is similarly indicative of the outermost possible extent of a particular subspace. The underlying intuition here is that, due to the frequentist components of the information theoretic co-occurrence statistics used to build the base space, different dimensions have different distributional profiles. To demonstrate this point, Table **??** presents the mean values and standard deviations for the distribution of mean and maximum points from the top 20,000[4] most frequent co-occurrence dimensions, as well as the top five and bottom five values for each of these statistics for illustrative purposes.

The co-occurrence dimensions that tend to have lower mean and maximum values are clearly quite frequent words, and this is to be expected, given that the high frequency of independent observations of the word will drive PMI scores down for that word across the board. The emergence of relatively infrequent words at the top end of the spectrum is then also to be expected.[5] The main point to note here, though, is that there is a broad range of possible mean and maximum values for a given dimension, and so the points M and X might be expected to vary considerably from subspace to subspace. Moreover, this variance may in turn correspond to semantic features of a given subspace: it may be the case that a given type of relationship between input terms – terms which are similar or dissimilar, literal or figurative in relationship to one another – select for a subspace which has a particular orientation in terms of its dimensional profile. A final observation here regards the way that the distribution of mean and maximum dimensional values skew, with means tending to clump towards the low end of the spectrum while maximums are

[4]less frequent dimensions tend to have higher PMI values overall, and also tend to be products of co-occurrences observed in quite obscure passages of the base corpus—it's worth recalling that a little more than half of the co-occurrence dimensions are observed only once.

[5]The appearance of *yeah* as one of the dimensions with a particular high maximum value is interesting, and perhaps surprising, though it should be noted that this is a particularly un-Wikipedian word, and is likely to occur in the context of things like quotations and band names, where co-occurrence with likewise obscure terms is more likely.

more dense at the high end of the spectrum. More specific conjectures and results will be presented throughout the next two chapters.

In addition to the situation of the points A, B, C, M, and X in a subspace, a normalised version of the subspace is considered, in which each vector is effectively measured at its intersection with a hypersphere of radius 1 emanating from the origin. These points are represented as A', B', C', M', and X' respectively in Figure **??**. The purpose of considering these points is to take measure of the way in which the various vectors in a given subspace relate to the subspace as a whole, regardless of the extent of these vectors. So, for instance, the vectors A and B might have very different norms, but the distances A', B', C', M', and X' might still be very small—and, even then, the angle $\angle A'M'B'$ might be very large, suggesting that A and B both pass through the central region of the subspace but on different sides of the generic central point of the subspace. One of the objectives of this analytical method is to test whether this kind of information, which can be captured through a robust geometric description of a subspace, is semantically indicative.

So finally the various geometric features available for the analysis of a subspace are systematically outlined in Table **??**. The points to be found in the space are broken down into three types, namely, the word-vectors themselves (points A and B), the generic points that emerge from an analysis of a subspace (points C, M, and X), and the normalised versions of all these points (A', B', C' M', and X'). The relationships between these points are construed across five categories as follows:

Distances Euclidean distances, such as the distance between the two word-vectors A and B as well as the norms of the generic points, and, additionally, the mean distance of A and B from the origin;

Angles The angles at the vertexes of the generic points of a subspace, so for instance $\angle ACB$ formed by lines $\overline{AC}$ and $\overline{BC}$, as well as the normalised versions of these angles, and also the angles formed between the vectors of the generic points such as $\angle COM$;

Means The average norms of the word-vectors and the average distances from the word-vectors to generic points as well as the average distances of the normalised versions of these points;

Ratios The ratio of the norms of the word-vectors and of the distances from the word-vectors to generic points, taking the lower of the two distances as the denominator, as well as the normalised version of the same measures;

Fractions The ratio of the mean distance from the origin of A and B to each of the three generic points, as well as the ratios of the generic points to one another.

These features have been selected as indicative of the overall comportment of the subspaces from which they are extracted, and, both independently and in conjunction, are expected to serve as indicators of the semantic phenomena characteristic of the word-vectors used to generate the subspace into which they are projected. So, for instance, I will predict (incorrectly, it turns out) that the distance $\overline{AB}$ will be one of the strongest

	DISTANCES
word-vectors	$\overline{AB}$
generic points	$C, \quad M, \quad X$
	ANGLES
word-vectors	$\angle AOB, \quad \angle ACB, \quad \angle AMB, \quad \angle AXB$
normalised	$\angle A'C'B', \quad \angle A'M'B', \quad \angle A'X'B'$
generic points	$\angle COM, \quad \angle COX, \quad \angle MOX$
	MEANS
word-vectors	$\mu(A,B), \quad \mu(\overline{AC},\overline{BC}), \quad \mu(\overline{AM},\overline{BM}), \quad \mu(\overline{AX},\overline{BX})$
normalised	$\mu(\overline{A'C'},\overline{B'C'}), \quad \mu(\overline{A'M'},\overline{B'M'}), \quad \mu(\overline{A'X'},\overline{B'X'})$
	RATIOS
word-vectors	$A:B, \quad \overline{AC}:\overline{BC}, \quad \overline{AM}:\overline{BM}, \quad \overline{AX}:\overline{BX}$
normalised	$\overline{A'C'}:\overline{B'C'}, \quad \overline{A'M'}:\overline{B'M'}, \quad \overline{A'X'}:\overline{B'X'}$
	FRACTIONS
word-vectors	$\mu(A,B)/C, \quad \mu(A,B)/M, \quad \mu(A,B)/X$
generic points	$C/M, \quad C/X, \quad M/X$

Table 3-F: Geometric features extrapolated from a subspace projected based on an analysis of two two input terms A and B.

indicators of semantic relatedness. Furthermore, the extrapolation of the generic features of a subspace is expected to indicate more general patterns of co-occurrence that are associated with semantic phenomena such as similarity and metaphor. When dimensions with similar mean value are jointly selected by a pair of words, a (more correct) expectation will be that this indicates a high degree of conceptual overlap between the words' referents, and therefore a high degree of similarity.

As a more general hypothesis, I surmise that different sets of geometric features will collectively be predictive of different semantic phenomena. One of the primary objectives of the empirical work described in the next two chapters will be to establish a methodology for mapping features to phenomena and then using these correspondences as a mechanism for understanding the statistical characteristics that allow for the computational extraction of semantically and contextually useful information from large scale corpora. It will therefore ultimately be the comparison of the groupings of features corresponding to specific semantic phenomena that will provide the most significant outputs of the research reported here, and so the arrangement of features in terms of types and categories as outlined in Table **??** is in this regard a schematic for the computational experimentation and corresponding evaluation and analysis at the core of this thesis.

3.4 A Mathematical Justification for Geometric Analysis

The application of geometry as a productive analytical tool for extrapolating semantic information from contextualised co-occurrence statistics has been, thus far, presented as a somewhat intuitive decision. There is certainly a certain elegance to using quantifiable distances and angles as the analytical representation of choice, and this approach will, it

will be seen, assist in the visualisation of what's happening statistically in the subspaces produced by my model. Notwithstanding these benefits, this section will offer a more mathematically thorough explanation of why a geometric approach is the right one for the types of statistics that are being used here, and in probabilistic models in general.

In order to understand the usefulness of geometry, it is worthwhile to consider again the information theoretical nature of the statistics being used here, and more generally in a plethora of distributional semantic models. Specifically, revisiting and restating Equation **??**, the scalars of the base model are defined by considering a ratio of frequencies approximately equivalent to a ratio of probabilities:

$$PMI(w, c) \approx \log\left(\frac{p(w, c)}{p(w) \times p(c)}\right) \tag{3.9}$$

In other words, PMI values are logarithms of probabilities, and logarithms have the natural property of translating products and ratios into sums and differences. So, for instance, if we have an operation such as $PMI(w_1, c) - PMI(w_2, c)$, we can express this as a log of a ratio of products of probabilities, with the term $p(c)$ dropping out of the equation:

$$PMI(w_1, c) - PMI(w_2, c) \approx \log\left(\frac{p(w_1, c) \times p(w_2)}{p(w_2, c) \times p(w_1)}\right) \tag{3.10}$$

This, in turn, actually just reduces to a product of conditional probabilities:

$$||PMI(w_1, c) - PMI(w_2, c)|| \approx \log\left(p(c|w_1) \times p(c|w_2)\right) \tag{3.11}$$

Next it must be noted that the construction of many of the geometric features described in Table **??** involves precisely the kind of operation described in Equation **??**. In particular the distance between two word-vectors, in linear algebraic terms, correlates with the sum of the squares of the point-by-point differences between the scalars of the two vectors, or, in other words, the sum of the squares of a sequence of logarithms of products of conditional probabilities:

$$\overrightarrow{w_1} - \overrightarrow{w_2} \approx \sqrt{\sum\left(log\left(p(c|w_1) \times p(c|w_2)\right)\right)^2} \tag{3.12}$$

Because a product of probabilities is itself a conditional probability ($p(a) \times p(b) = p(a, b)$), and because the square of a logarithm is the logarithm of the argument to it's own power ($(\log a)^2 = \log(a^{\log a})$), distances under the geometric regime I've outlined can be understood as essentially highly compounded probabilities, or, in other words, Bayesian inferences about complex concatenations of linguistic co-occurrence events.

Now, regarding the expression $\log p(c)$ on the right hand side of Equations **??** and **??**, this value will be stable for all word-vectors across a given dimension, and so effectively

becomes a skewing constant, universally shifting the relationship of word-vectors to the origin, but not to one another. The same holds for the various measures involving the distance between word-vectors and the generic points of a subspace, and, as will be seen in subsequent chapters, linear combinations of features not involving the origin turn out to be most predictive of a variety of semantic phenomena.

Furthermore, a similar skewing effect can be observed in the approximate nature of the equations offered above. In Equation **??**, the addition of the constant a is a necessary feature of the calculation, as, without it, our subspace selection technique drifts into obscurity, picking dimensions containing very little useful information. The addition of 1 to the ratio of frequencies, which is a convenient move given the geometric desirability of having an entirely non-negative space, turns out to also be a mathematically effective move, as it somewhat offsets the skewing effects of a. In fact, given the constant value W and the mean values for f_w and f_c, it turn out that the addition of 1 to the overall ratio pretty nearly offsets the addition of a to f_c in the denominator, so, in the average case, these two constants restore the natural PMI value of w co-occurring with c as though it were simply a logarithm of a joint probability. As we move away from the average frequencies, there is a bit of a twisting of the space, but on the whole these constants preserve the relationships of word-vectors fairly well well contributing some very desirable qualities to the space overall.

The implication of this insight into the mathematics of information theory is that many of the geometric relationships described in Table **??** are in fact encodings of complex Bayesian inferences about the way that words are expected to co-occur with one another.

Moreover, linear combinations of these features will likewise be

With this understanding of the mathematics of distributional semantics in place, we can now effectively use the techniques of machine learning – linear regression, logistic regression, feature selection methods, and so forth – to learn geometric relationships which can subsequently be converted into Bayesian

and just such an approach will be applied to considerable effect in the following chapters.

So, finally, we see how the *geometrisation*, so to speak, of the co-occurrence statistics inherent in a textual corpus are an effective way not just to help humans visualise these relationships between frequencies, but also to provide a computationally tractable mechanism for handling the state-space of belief-quantities.

This more or less sets the stage for the empirical section of the thesis. The only outstanding issue is the establishment of the models which will serve as consistent points of comparison for my methodology.

3.5 Comparing to Alternative Approaches

In order to evaluate the effectiveness of my methodology, it will naturally be necessary to compare the performance of the models I develop against other models. One way of doing this will, of course, be to compare to results other researchers have obtained experimenting with the data which will serve as the foundation for the results reported in the next two chapters. In the cases of results reported by other researchers, though, similar but variously different corpora have been used to train other models described in the literature. This is to be expected, and the results for large scale corpora should be fairly generalisable assuming a sensible choice of data (and the use of Wikipedia as all or a large portion of base data is quite common in the field), but nonetheless it will be useful to establish a baseline of results generated using models trained on the exact corpus to which I apply my methodology. And in the cases of metaphor and semantic type coercion in particular, which will be examined in Chapter **??**, the datasets explored are relatively new and have not been approached by many researchers in the field, so any additional point of comparison will be valuable in evaluating my methodology.

Moreover, in most cases, other models have been designed in a task specific way: so, for instance, **?** have developed a syntactic heuristic for identifying semantic similarity as compared to relatedness in particular, and **?** describe a model that generates compositional adjective-noun representations geared towards metaphor detection. One of the key features of my models is that they are intended to be *general*: the geometries generated by my methodology are expected to be replete with semantic interpretability, allowing for the same potential for diverse and often surprising conceptualisation corresponding to the infinitely combinatory characteristic of natural language in use. For this reason, it is desirable to have a base case of a generic model that can be compared across the board to all the different tasks handled by my methodology.

With all this in mind, I propose two different points of comparison that, in addition to results extracted from existing literature, will be applicable to all subsequent experiments described here. The first is a pair of technique interpreting my base space in a non-contextual way, and so will serve as a way of measuring the degree to which building context specific subspaces enhances the ability to model semantic phenomena. The second is an application of a well known and highly productive neural network model to the same underlying data that I've used. This will serve as a mechanism for comparing my results to what has proved to be another very effective methodology for the statistical modelling of semantics in general.

3.5.1 Static Interpretations of the Base Space

In cases where the geometry being explored involves just target word-vectors and generic points of a space – so, for all the features described in Table **??** – it is computationally tractable to treat the sparse base matrix from which subspaces are projected as a semantically interpretable space in its own right. This is because universal generic points (the mean point for all dimensions, the maximum point for all dimensions, and the cen-

tral point based on the average value of the mean point) can be discovered through a one-off caluculation, and the word-vectors themselves will be relatively sparse. In the most honerous case of comparing two orthogonal word-vectors using the INDY technique, the total number of scalars involved in the computations of geometric features would be the sum of the number of non-zero dimensions for each word-vector, so something on the order of thousands to tens of thousands of values—not that bad, computationally speaking.

Of course, the norms of the generic points in such a general space will be extremely high compared to any given actual word-vector, since these generic points will have non-zero values in several million dimensions. With this in mind, a second and more typical approach to building a general and computable distributional semantic space out of my base space of co-occurrence statistics is to through matrix factorisation: using singular value decomposition, I project the base space onto the top most informational eigenvectors up to a dimensionality to match the parameters tested using my context specific dimensional selection techniques.[6] Because of computability constraints, I take the top 100 to 50,000 most frequent word types as the vocabulary for this model, and consider the top 10,000 most frequent co-occurrence terms as the dimensions of the matrix to be factorised. This means that 1,979 of the 1,998 word tokens in the SimLex999 dataset (**?**, discussed in detail in Chapter **??**) are included in the vocabulary, and almsot 90% of the co-occurrence observations tabulated in the base space are represented in the decomposed model.

Because SVD produces a space in which dimensions are characterised by variance rather than extent (meaning that signs can be reversed along a given dimension, and the barycentre is typically at the origin), the factorised model is not suitable for generating the generic points which are key features of my contextual models. In order to make the most fair comparisons possible, this factorised model will be shifted in the case of each analysis of a set of word-vectors W such that those word-vectors have the highest possible value along a dimension D in the set of top eigenvectors, with the value furtherest from the mean of the word-vectors being reset to zero:

$$d'_i = |d_i - \underset{d \in D}{\operatorname{argmax}}(\mu\{d_w : w \in W\} - d)| \tag{3.13}$$

This shifting procedure in practice introduces a degree of context to the generic dimension reduction technique, allowing for new geometric relationships between word-vectors and emergent generic points of the model to be established for each set of inputs; the distances between the word-vectors themselves, meanwhile, are unaffected. In the end this will simply re-enforce the point that the difficulty of systematically applying context using more typical dimension reduction techniques is one of the strengths of my methodology.

[6] In practice, the `sklearn` python module's PCA method is used to do this.

3.5.2 A Model Trained Using a Neural Network

In addition to the interpretations of the statistical base space described above, the neural network based models outlined by **?** under the rubric `word2vec` will be used as a point of comparison. These models have received a remarkable degree of attention in the NLP literature since their introduction a few years ago, so much so that the software was mentioned by name in 116 out of the 230 long papers published in the 2016 Proceedings of the Meeting for the Association for Computational Linguistics (**?**). The models have been taken, sometimes in modified form, as a source for representations of words *embedded* in vector spaces trained on large scale textual data, applied to tasks ranging from word relatedness and similarity ratings (**?**) to analogy completion (**?**), and have also been applied to multimodal tasks such as image labelling (**?**).

The `word2vec` framework includes two different neural network architectures for generating word-vector representations based on traversals of large scale corpora. The *contextual bag of words* (CBOW) technique treats the terms in a co-occurrence window surrounding a target word w as input and attempts to learn a representative word-vector $\overrightarrow{w}$ that is predicted by processing the input word-vectors through a recursive neural network. The *skip-gram* technique, on the other hand, treats the representation $\overrightarrow{w}$ itself as input to a network which learns to predict word-vectors representing words on either side of the target word. In both cases, the model updates the scalars of the target word vectors in order to move them closer to the vectors representing each co-occurrence in which they're observed through backpropagation. In the case of the CBOW model, the terms co-occurring within a given window of the target word are combined into an average vector for the purpose of each training observation; with the skip-gram model, the selection of target output word-vectors is weighted based on their distance from the input word-vectors, and the model optimises the probability of two word vectors interpreted via the softmax function (see **?**, for more details).

In addition to the size of the co-occurrence window, model parameters include the number of iterations of the corpus, the architecture of the single-layer network connecting input to output vectors, and, in the case of the skip-gram model, a rate of negative sampling by which random sets of words are taken as instances of non-co-occurrences and used to push the corresponding word-vectors away from the input word-vector. The skip-gram model, with its sensitivity to word order, has been reported to perform particularly well on analogy completion task involving semantic similarity, so for instance in discovering the relationship *king:queen::man:woman*. The CBOW model, on the other hand, has performed better on what the authors have described as *syntactic* analogies such as *good:better::bad:worse*.

Here, the skip-gram and CBOW techniques of `word2vec` will be taken as exemplars of general-purpose distributional semantic modelling. For the purposes of a fair comparison, I've trained instances of both models using the same cleaned corpus described in the previous chapter and used to train my own model. The presumption, corroborated by the wide applications found for the models and described by various authors over the past three years, is that this approach provides a general framework for generating a space in which word-vectors relate to one another in conceptually productive ways. A primary

difference between the vectors learned by `word2vec` and the vectors representing word co-occurrence statistics derived by my model is that `word2vec` produces dense vectors whose dimensions cannot be individually interpreted as corresponding to any specific set of observations across a corpus, whereas my model generates a base space of sparse vectors for which each dimension maintains its status as an indication about a tendency of co-occurrences with a specific term. This dimensional interpretability gives my model its power of contextualisation.

Following from this, it should also be noted that in the `word2vec` models, as is likewise typically the case with models generated using principle component analysis, semantic relationships are measured in terms of cosine similarity between word-vectors, which means that the models are treated as effectively normalised vector spaces centered at the origin. A consequence of this normalisation and centering is that these spaces lack a sense of perimeter and extent, which means that they can't be interpreted in terms of the relationship between word-vectors and generic points characteristic of a contextual subspace, as described above. These two features of my methodology, its ability to generate subspaces contextually and its capacity for nuanced geometric interpreation, are the two essential points that will be examined in the experiments described in the next two chapters.

Chapter 4

Conceptual Clusterings, Similarity, and Relatedness

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 much. In this chapter, I will now 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 semantics based on statistics about word co-occurrences. They encompass two different experimental set-ups and corresponding varieties of data, one of which has been designed specifically for the purpose of testing my ideas and one of which involves an assortment of data used pervasively by computational linguistics interested in semantic models.

The first experiment, presented as a proof of concept, involves using multi-word phrases as input and evaluating the methodology's capacity for building subspaces where words associated with the conceptual category denoted by the input term can be reliably discovered. This experiment expands upon the notion of proto-conceptual spaces outlined in the previous chapter, considering whether the word vectors that populate regions of subspaces are characterised by a certain categorical coherence. In the case of the data explored here, the experiment is specifically set up to feel out the contextual capacity of my methodology and compare it to a standard generic semantic space. The question asked is whether the shifts from subspace to subspace based on particular input yield productive alterations in the way that words both cluster and emerge from the melange of word-vectors that circulate around my base model.

The second experiment moves into more familiar computational linguistic territory, using some well-travelled datasets to examine the methodology's capacity for identifying two related but distinct semantic phenomena: relatedness and similarity. Each of these objectives have provided reliable but distinct evaluative criteria for computational models of lexical semantics. 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.

4.1 A Proof of Concept

In this section, I present the first experiment performed using my contextually dynamic distributional semantic model. The gist of this experiment is to take a word pair representing a compound noun – for instance, *body part* – and see if my methodology can use the word pair to contextually generate a space where other words conceptually related to that compound noun can be found in a systematic way. This is conceived of as an entailment task, in that I will attempt to find phrases considered to be categorical constituents of the concept represented by the word pair, taking the WordNet lexical taxonomy as a ground truth. There is a scholastic back story here.

An early version of this experiment was reported in **?**. That first effort arose out of a question posed by a colleague regarding the feasibility of using a statical NLP technique for generating categorical labels that could be used to evaluate computational creativity in a domain specific way (for a psychological perspective on the difficulty of generating such terms in an objective way using human subjects, see **?**). So, for instance, given a creative domain such as MUSICAL CREATIVITY, could a distributional semantic model generate terms that are reliably relevant to the concept denoted by that phrase, rather than the potentially disparate properties independently associated with MUSIC and CREATIVITY? Intuitively there seems to be little reason to hope that the space halfway between these points in a general semantic space would somehow adequately represent the properties of the overall concept. The early work explored the dimensions contextually selected by analysing the co-occurrence features of word-vectors corresponding to inputs along the lines of the expository results presented anecdotally in Chapter **??**, but without any rigorous evaluation.

Reviewer responses to a subsequent journal article (**?**), designed as a more thorough introduction of the methodology, inspired a computationally oriented mode of evaluation. The experiment that has emerged involves attempting to recapitulate taxonomical conceptual relationships from the WordNet database (**?**). Wordnet is a lexical taxonomy of *synsets*, basically semantic word senses, arranged into a hierarchy of entailment relationships, with each synset associate with a number of *lemmas*, word types indexed by that synset according to human annotators. This experiment takes as input instances of synsets labelled by compound noun phrases and seeks to output as many of the lemmas listed associated with synsets that are hyponyms of the input synset. So, for instance, the synset body part has a hyponym EXTERNAL BODY PART, which has a hyponym EXTREMITY, which has a synset LIMB, which has a synset LEG associated with the lemma *leg*,

and so *leg* would be considered a positive output for the input *body part.*[1]

4.1.1 Experimental Set-Up

12 of the top synset labels consisting of compound noun phrases are extracted from WordNet. These labels are extracted through a breadth first traversal of the tree of noun synsets, selecting the highest 12 synsets with multi-word labels with the constraint that none of the 12 can be parent nodes of any of the others: in this way, 12 distinct, non-overlapping conceptual categories are choosen. The experimental vocabulary is considered to be the intersection of the list of all WordNet noun lemmas associated with the vocabulary of my model (the 200,000 most frequent word types in Wikipedia), resulting in a total vocabulary of 32,155 words. The lemmas associated with all the hyponyms of each synset are extracted and grouped, and these words become the target words for my models' output.

With the target output established, the terms labelling a given synset are passed to my model as contextual input, with the corresponding word-vectors serving as the basis for dimensional selection using the JOINT, INDY, and ZIPPED techniques as outlined in Chapter **??**. Here, the base space generated using a 5x5 word co-occurrence window is used, and 200 dimensional subspaces are returned; variations of these parameters will be tested in subsequent experiments. The subspaces returned by each of these techniques are explored to return the top terms using both of the procedures outlined in Chapter **??**: the terms closes to the mean point between the input word-vectors in a subspace are returned, and the terms furthest from the origin – the terms with the largest norm – in a given subspace are returned. The top 50 terms found in a subspace each according to each measure are returned, as well as the top terms up to a limit n where n is the total number of lemmas associated with the target multi-word label. Accuracy scores for each of these sets of output are computed, so the total number of positive matches for hyponyms of the input synset out of the top 50 and top n terms returned.

As a point of comparison, results are likewise returned from two different `word2vec` models, one using the skip-gram methodology and one using the bag-of-words methodology, as described in Chapter **??**. In line with the subspaces generated using my methodology, 200 dimensional models are used, and these models are built across 10 iterations of the corpus, using a 5x5 word co-occurrence window, applying a negative sampling rate of 10 and an initial learning rate of 0.025, as discussed in Chapter **??**. Here the top terms in terms of proximity by cosine similarity to the mean point between the word-vectors associated with the input terms are returned, again taking the top 50 and top n for each input.

[1] In keeping with the convention used elsewhere in this thesis, synset labels will be presented in small caps and lemmas will be presented in italics.

		JOINT		INDY		ZIPPED			
		norm	dist	norm	dist	norm	dist	SG	BoW
top-50	accuracy	0.292	0.208	0.240	0.189	0.273	0.199	0.247	0.270
	ratio	10.304	6.129	7.731	5.270	8.625	5.719	6.733	7.168
full	accuracy	0.235	0.160	0.198	0.149	0.210	0.153	0.081	0.079
	ratio	4.967	3.525	3.967	2.997	4.290	3.221	2.397	2.551

Table 4-A: Average accuracy scores and average ratio of accuracy to baseline for reconstructing the lemmas entailed by 12 different multi-word WordNet synsets, for both the top 50 terms returned by models and the full set of terms returned up to the number of lemmas associated with each input.

4.1.2 Results and Analysis

Results for the set-up described in the previous section can be found in Table **??**, with both the average accuracy scores and the average ratio of model accuracy to baseline reported. Results for both the norm and distance from mean point methods are reported for subspaces derived using the JOINT, INDY, and ZIPPED dimension selection techniques, followed by results for the skip-gram and bag-of-words `word2vec` techniques. The first thing to note about these results is that all of the results are substantially above the baseline: the average ratios of model accuracy to the baseline (the likely accuracy achieved by randomly choosing words from the vocabulary for each input) are all above 2.5, and are above 3.2 for all of my methodologies. So it is clear that all these techniques are generating semantically significant relationships between word-vectors.

Results across the board are strongest for the JOINT dimension selection technique applying the norm measure for returning output: in these subspaces selected by choosing dimensions with high PMI values across all contextual inputs, word-vectors that are far from the orgins – and that therefore likewise tend to have high values across all these dimensions – are most characteristic of the conceptual category indicated by the input. This is not surprising. Results for the norm measure applied to ZIPPED and INDY type subspaces follow in kind, with intermediary performance from the in-between ZIPPED technique, where all dimensions bear at least some tendency for co-occurrence with the input terms, and then another step down for the INDY subspaces. In all cases the norm measure outperforms the two `word2vec` results.

More surprising is the distinction between the strong performance of the norm measures and the less impressive performance of the mean point measure. In the case of accuracy among the top 50 terms returned by each model, my methodologies results using this Euclidean measure consistently fall short of the `word2vec` techniques. It would seem, then, that in the subspaces returned by my models, proximity to the input word-vectors is not in itself an indicator of categorical inclusion in the conceptual space traced by the intersection of the correspond contextual input terms. Upon further consideration, there is a plausible explanation for this: revisiting the outputs for subspaces projected using denotations of animals as input, reported last chapter in Tables **??** and **??**, the norm measure produced specialised terms such as *chital* and *poodle*, while the distance

		top-50				full			
	baseline	norm	dist	SG	BoW	norm	dist	SG	BoW
psychological feature	2.39	0.240	0.660	0.400	0.401	0.417	0.130	0.102	
causal agency	0.177	0.000	0.140	0.100	0.180	0.125	0.170	0.043	0.102
human action	0.156	0.180	0.460	0.500	0.480	0.300	0.346	0.127	0.116
animate being	0.044	0.020	0.060	0.020	0.020	0.030	0.031	0.007	0.006
cognitive content	0.043	0.360	0.260	0.320	0.300	0.168	0.188	0.065	0.050
mental object	0.043	0.120	0.240	0.140	0.180	0.130	0.188	0.068	0.053
physical process	0.035	0.520	0.260	0.160	0.200	0.205	0.138	0.056	0.065
social group	0.031	0.080	0.220	0.320	0.380	0.075	0.114	0.059	0.064
body part	0.025	0.760	0.120	0.100	0.220	0.407	0.080	0.047	0.087
taxonomic category	0.024	0.460	0.180	0.540	0.540	0.147	0.026	0.163	0.164
physiological condition	0.020	0.640	0.160	0.320	0.280	0.365	0.099	0.155	0.139
woody plant	0.012	0.120	0.060	0.080	0.060	0.143	0.127	0.046	0.062

measure generated relevant but not always categorical terms such as *wild*, *giant*, and *golden*.

We might characterise this tend in terms of a distinction between words which denote semantic *relatedness* versus *similarity*, a topic which will be addressed in depth in the next section.

SOME EXAMPLES

Focusing on the accuracy of the results returned by the models up to the full length of each target set of lemmas, here results are weaker all around, which is not particularly surprising: as we move away from the regions where we expected to see the highest degree of conceptual consistency, mismatched terms begin to creep into the results. It is notable, though, that my methodologies outperform the neural network based models across the board, especially for the norm based measures but also in the case of this larger sample of the respective semantic spaces for the distance based measures. In fact, the stronger relative performance for the distance measure in these expanded regions of each type of subspace makes sense, since, as the norms measure moves closer to the origin in search of output and the distance measure likewise expands from the locus of its mean point, the results output by each measure will increasingly overlap (an overlaying of Figures **??** and **??** will illustrate this phenomenon). But the main point to take here is that, in the case of my methodologies, there is clearly a more persistent conceptual organisation to the space. As we expand from any point in the static type of semantic model generated by `word2vec`, we will undoubtedly begin to encounter the vagary and the messiness inherent in language and problematic for fixed lexical relationships. My methodologies, on the other hand, afford the *ad hoc* construction of semantic spaces which afford the situational corralling of the looseness and ambiguity inherent in a dynamic lexicon.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Press, Cambridge, MA.
Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, 10(8):370–374.
Clark, S. (2015). Vector space models of lexical meaning. In Lappin, S. and Fox, C., editors, *The Handbook of Contemporary Semantic Theory*, pages 493–522. Wiley-Blackwell.
Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167.
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal for the American Society for Information Science*, 41(6):391–407.
Derrac, J. and Schockaert, S. (2015). Inducing semantic relations from conceptual spaces: A data-driven approach to plausible reasoning. *Artificial Intelligence*, 228:66–94.
Dreyfus, H. L. (2012). A history of first step fallacies. *Minds and Machines*, 22(2):87–99.
Dummett, M. (1981). *Frege: Philosophy of Language*. Duckworth, London, 2nd edition.
Erk, K. and Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 897–906.
Erk, K. and Padó, S. (2010). Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97.
Erk, K. and Smith, N. A., editors (2016). *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany.
Evans, V. (2009). *How Words Mean: Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford University Press.
Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1606–1611.
Gärdenfors, P. (2000). *Conceptual Space: The Geometry of Thought*. The MIT Press, Cambridge, MA.
Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. The MIT Press.
Geffet, M. and Dagan, I. (2005). The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 107–114.
Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Miffline, Boston.
Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics Volume 3: Speech Acts*, pages 41–58. Academic Press, New York.
Gutiérrez, E. D., Shutova, E., Marghetis, T., and Bergen, B. K. (2016). Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.
Hill, F. and Korhonen, A. (2014). Learning abstract concept embeddings from multimodal data: Since you probably can't see what I mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 255–265.
Jäger, G. (2010). Natural color categories are convex sets. In Aloni, M., Bastiaanse, H.,

de Jager, T., and Schulz, K., editors, *Logic, Language and Meaning: 17th Amsterdam Colloquium, Amsterdam, The Netherlands, December 16-18, 2009, Revised Selected Papers*, pages 11–20.

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

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

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

Kartsaklis, D. and Sadrzadeh, M. (2016). Distributional inclusion hypothesis for tensor-based composition. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2849–2860.

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

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

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

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

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

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

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

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

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

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.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

von Uexküll, J. (1957). A stroll through the worlds of animals and men: A picture book of invisible worlds. In Schiller, C. H., editor, *Instinctive Behavior: The Development*

of a Modern Concept, pages 5–80. International Universities Press, Inc., New York City, NY.

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

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

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

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

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