A Geometric Method for Context Sensitive Distributional Semantics

by

Stephen McGregor

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

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

September 2017

My university requires me to make the following statement:

I add this:

Abstract

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

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

This method is presented as a complement and a counterpoint to established distributional methods for generating lexically productive word-vectors. Where contemporary vector space models of distributional semantics have almost universally involved either the factorisation of co-occurrence matrices or the incremental learning of abstract representations using neural networks, the approach described in this thesis preserves the connection between the individual dimensions of word-vectors and statistics pertaining to observations in a textual corpus. The hypothesis tested here is that the maintenance of actual, interpretable information about underlying linguistic data allows for the contextual selection of non-normalised subspaces with more nuanced geometric features. In addition to presenting competitive results for various computational linguistic targets, the thesis will suggest that the transparency of its representations indicates scope for the application of this model to various real-world problems where an interpretable relationship betweendata and output is highly desirable. This, finally, demonstrates a way towards the productive application of the theory and philosophy of language to computational linguistic practice.

Glossary

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

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

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

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

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

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

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

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

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

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

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

Table of Contents

List of Figures

List of Tables

Chapter 3

Semantics in Context

This chapter is concerned with a theoretical overview of a novel distributional semantic method designed to map words into conceptually productive geometric relationships. At the heart of this approach is the idea that concepts, and, correspondingly, cognition are fundamentally contextual phenomena: by this view, concepts are process oriented, not objective, and so "instantiating a concept is always a process of activating an ad hoc network of stored information in response to cues in context," (Casasanto and Lupyan, 2015, p. 546). It follows that language, as a mechanism for manipulating and transmitting cognitive content, is then likewise contextually situated, with meaning itself crucially being determined only in the moment of language use (Austin, 1962). So, theoretically speaking, the method which will be described is based on some well travelled, if not entirely mainstream, ideas about the nature of language and mind:

1. Concepts are not stable; they are generated in response to unfolding situations in an unpredictable environment;

2. Lexical semantics are accordingly always underspecified, and always resolved in some environmental context;

3. There is no relationship of strict supervenience between language and concepts one way or the other, but instead a dynamic by which concepts invite representation and communication, and language affords conceptualisation.

These ideas, which have been outlined throughout the previous chapter, are not the standard dogma of computational linguistics, which generally, and understandably, has modelled concepts as modular, portable entities, language as a likewise stable system of representations and rules, and the relationship between the two as one of source and contingent data (see, for instance, the textbook treatment in Jurafsky and Martin, 2000, particularly Ch. 17). This structure-oriented approach to language and mind epitomises a project that Dreyfus has described as "finding the features, rules, and representations needed for turning rationalist philosophy into a research program," (Dreyfus, 2012, p. 89). As computer science and philosophy of mind increasingly interact at the vertex of cognitive modelling, culturally relative ideas about the connection between mental representations and linguistic symbols become incorporated into the very architecture of data structures, engendering a positive feedback loop by which the outputs of symbol manipulating information processing systems reinforce the premise that representations are stable entities which can be trafficked in the form of words according to the rules of a grammar.

I present the method outlined and tested in this thesis as an alternative to the foundationalist trend in computer science in general, and in computational linguistics in particular (see Rorty, 1979, for a robust philosophical criticism of the idea that concepts are stable). This project involves trading the computational and mathematical allure of dimension reduction techniques and neural modelling, which have been prevalent in distributional semantic approaches, for a theoretically robust notion of situational context selection. The methodology outlined theoretically in this chapter, and described technically in the next, has been conceived as a mechanism for the contextual generation of lexical representations that are structurally productive, in that the statistical features which make up a given representation define its geometric situation in relation to other representations in a particular context, and the geometry itself becomes semantically productive, with spatial relationships offering up interpretations of context specific word meanings.

I have no pretensions of instigating a paradigm shift in computer science. I do not claim that the methodology I will now describe represents a radical departure from the prevailing and highly productive approach to the computational modelling of lexica or knowledge; indeed, it is very much grounded in the same broadly pragmatic considerations that have been the foundation of the statistical aspect of distributional semantics: word meaning is to an appreciable extent determined by the sentential context associated with observable word use. My methodology is, rather, an attempt to build some consideration of the idea that minds are not populated by representations and that words are not static containers of meaning into the existing computational paradigm. With this in mind, my model is predicated upon four interrelated desiderata, derived generally rather than in a one-for-one way from the points enumerated above:

1. The method should generate representations that incorporate semantics directly into their structures;

2. The method should be dynamically sensitive to context;

3. The method should function in a way that is transparent and operationally interpretable;

4. The method should situate words in spaces that are likewise geometrically interpretable.

The first stipulation is a fundamental criteria for computational approaches to lexical modelling, if not to lexical semantics in general, and is to a certain extent built into the distributional semantic paradigm at the root of my methodology. The second stipulation encapsulates the theoretical premise of this work. A primary objective of my methodology is to identify a statistically tangible mechanism for choosing word co-occurrence features in a contextually relevant way. Specific mechanisms will be outlined in Chapter 4, and what counts as context will be discussed further in the course of the empirical results presented in Chapters **??** and **??**, but the general idea is that a context sensitive model needs to react dynamically to information about what's happening in some linguistic situation. The second requirement follows directly from the first: in order to pick semantic contexts *in situ*, there needs to be a way to get a handle on the data which underwrites a model. In practice this means that the scalars that form the basis for all the models which will be explored here represent literal information about co-occurrences in a large scale corpus, and the feature selections that take place in the course of delineating a contextual geometry can be traced to specific events in the underlying data.

Finally, the informationally transparent selection of contextual subspaces must result in a likewise interpretable geometry, where there is a coherent mapping between spatial features and semantic properties. This last criterion in particular will lend the methodology one of its most powerful characteristics: by contextually selecting subspaces in which a variety of geometric relationships between word-vectors and more general features of the space can be analysed, we can hope to discover a single general way of representing a variety of semantic phenomena in a particular subspace. As will be seen in Chapter 4, these subspaces will have a variety of geometric properties, including an origin, distance from an origin, and central and peripheral regions. In this regard, my methodology presents an additional point of comparison with the standard distributional semantic approaches, which typically employ normalised spaces, often in the form of a hypersphere with both positive and negative values: while these are all vector space models and are all therefore to a certain extent concerned with extracting meaning from spatial relationships, my approach is in a certain respect *more* geometrical, in that a variety of relationships, linear, angular, relative, and absolute, emerge in a given projection. This geometric richness gives a model constructed using my methodology a wealth of interpretive features, ultimately allowing for the observation of different semantic properties – for instance, similarity versus relatedness – to emerge as different geometric aspects of the same subspace.

In the following sections, each of these requirements will in turn be analysed in the context of the underlying theoretical subtext. This analysis is performed with an eye towards the immediate project of designing a statistical model for mapping word-vectors to concepts by way of semantic geometry, and each element of the profile of desirable properties will be explored with this in mind.

3.1 Modelling Lexical Semantics

This thesis is primarily concerned with the problem of semantic representations, and in this regard finds itself in good philosophical stead. Russell, for instance, was concerned with the property by which language *denotes*, meaning the way in which a word or phrase actually points to a thing in the world rather than the more elusive concept of meaning. Russell concludes that denotations can only denote in those instances where they correspond to true propositions, and moreover "that denoting phrases have no meaning in isolation" (Russell, 1905, p. 192), which is to say that things like words acquire semantics situationally. Kaplan (1979) engages with denotation again in his explication of demonstratives (words that mean what they mean relative to the situation of interlocutors), re-enforced by the intermediary development of possible world semantics (Carnap, 1947), arguing in particular that these types of denotational entities are mapped to propositions and, correspondingly, meanings in a way that in necessarily context specific. From this standpoint, Kaplan constructs a productive formalism for how words like *this*, *that*, *here*, and *now* denote particular entities, times, and places relative to the situation in which the denotation comes about. The point to extract for present purposes from this logical tangle is that there is a critical distinction to be made between a thing in the world, it's representation, and the way in which the representation acquires meaning in terms of the comportment of a linguistic agent—and this distinction occurs to a great extent *contextually.*

The idea of structurally productive lexical representations finds its roots even deeper in the tradition of the philosophy of signification, in the semiotics of Peirce, who suggests that "there must exist, either in thought or in expression, some explanation or argument or other context, showing how–upon what system or for what reason the Sign represents the Object or set of Objects that it does," (Peirce, 1932, ¶2.230). In other words, a representation denotes and means by virtue of the actual dynamics of the symbol itself as it exists in the world. There is a story to be told about how a representation comes to operate in the way that it does: as Rączaszek-Leonardi puts it, meaning-bearing symbols "can arise only from the history of a certain physical structure as a constraint on certain system's dynamics in a certain environment," (ibid, p. 309). Furthermore, a lexical representation's acquisition of its dynamics unfolds on a number of different timescales, for instance on the scale of individual cognitive development as well as on the scale of the history of cultural transmission, effectively prohibiting any attempt at an elimitivist interpretations of linguistic symbols as atomic units. Instead, language is, within the regime of environmentally situated cognition, taken to be a cognitive object which affords meaning-making and conceptualisation; as Clark suggests, there is scope to "consider language itself as a cognition-enhancing animal-built structure," (Clark, 2006, p. 370). Given this objective and even material quality of language, it seems clear that a good model of lexical semantics should traffic in symbols which are likewise susceptible to conceptually productive, open-ended manipulation.

From a cognitive linguistic perspective, then, the application of the concept of *frames* (Barsalou, 1992) as a mechanism for providing conceptually structured representations of cognitive content has proved fruitful: through many-to-many mappings of lexical representations to conceptual frames by way of *access sites* at which the cognitive and the linguistic interact, Evans (2009) proposes a way in which language gains its interactivity through close bindings with productive cognitive structures. From a slightly different perspective, but with a similar objective of providing a model of language that is sensitive to context and compositionally flexible, Pustejovsky (1995) proposes a *generative lexicon* predicated upon computable representations with multiple levels of interactive features. Pustejovsky's objective is to move beyond models of a *sense enumerative* lexicon, by which lexical forms are simply mapped to a variety of different semantic interpretations, and towards a structured mode of representation allowing for the open ended application of semantic phenomena such as *type coercion*, by which nouns take on different categorical denotations under the influence of a particular verb in a particular conceptual context.[1] So once again, the construction of interactive lexical representation affords the conceptually productive computation of semantic phenomena in a specific cognitive context.

In this thesis, I will skirt the important but also fraught question of semiotic processes in the natural world and their tortuous relationship to natural language; instead, I will simply take the philosophical insight into the structural nature of representations as a guideline towards an effective methodology for computationally modelling word meaning. My stance is that distributional semantics is the right framework for doing this, because

[1] Type coercion will form a test case for my model, explored empirically in Chapter **??**.

it provides a mechanism for building up representations that by design contain their own semantics. A similar point has been raised by Clark (2015), who notes that "once we assume that the meanings of words are vectors, with significant internal structure, then the problem of how to compose them takes on a new light," (ibid, p. 509). In my work, I'm concerned not directly with compositionality, but with the related issue of how lexical semantics are contextually specified, and I maintain that a similar approach to constructing representations with highly interpretable and interactive structures will be a productive pathway towards accomplishing this goal. Vector space models provide the setting for the mapping of statistical phenomena observed across large scale collections of textual data to geometric features which can be analysed quantitatively.

The logic of this approach is that, in a geometric model, the interrelationships between statistical features play out as spatial distortions as we move across the spectra of various semantic phenomena. The features in question will be, first and foremost, the relationships between word-vectors, and correspondingly the comparative co-occurrence profiles of words along specific dimensions—in this regard, my methodology starts at the same point as most distributional semantic approaches to lexical modelling. My proposition, though, is that standard distributional semantic approaches have not tended to take full advantage of the representational potentialities of statistical geometries. Bearing in mind that both Barsalou (2008) and Evans (2009), among others, have argued for the significance of statistics in understanding the way that lexical representations get built up cognitively, it seems like a good idea to embrace the affordances of vector space models, making decisions about dimension reduction through a situationally unfolding analysis and then considering the relationships between word-vectors and more general points in context specific subspaces. As will be seen in Chapter 4, these other points are to be constructed in such a way as to capture distributional properties of collections of dimensions, and one of the key findings of my thesis is that these more general properties, in addition to the standard technique of comparing the angles between individual word-vectors, provides statistical insight into semantic relationships.

Ultimately, then, the methodology described and explored in this thesis represents an attempt to move computational approaches to natural language processing toward the social and protean semantics of Putnam (1975), who famously quips that "meaning just ain't in the head," (ibid, p. 144), and instead suggests a rather abstractly defined system of representations which bear some of the load, so to speak, of semantics externally.[2] So, rather than thinking of meaning as a thing which is built into a lexical representation on an arbitrary and abstract level, my methodology is grounded in the idea that robust

[2] In fact, Putnam literally suggests that the his type of socially adapted representation might be thought of as a *vector*, though he surely means something a bit different than a string of co-occurrence statistics.

representations are emergent properties of complex dynamic systems, and aspects of these same dynamics are encoded, on various levels of abstraction, into the structure of a representation. This premise is, at least implicitly, built into the distributional hypothesis itself, but my proposal is to delve further into the question of how statistical analysis can afford contextualisation, and then how contextualisation can in turn provide a platform for a semantically productive geometric analysis of statistics.

3.2 Dynamic Context Sensitivity

At the heart of the technical work described in this thesis is an insight which is broadly accepted by theoretical linguists and philosophers of language: word meaning is always to some extent contextually specified. This wisdom is built into the foundations of both formal semantics (Montague, 1974) and pragmatics (Grice, 1975), and is likewise acknowledged in contemporary context-free approaches to syntax (Chomsky, 1986). As evident from the implementations of conceptual models surveyed in the previous chapter, however, the computational approach has generally relied on the idea that concepts can, at some level of composition, be cast as essentially static representations. The tendency to treat concepts as self-contained ontological entities consisting of properties that are wholly or partly transferable is built into the fabric of the formal languages used to program computers, and indeed into the mechanisms of modular data processing systems with specific compartments for the storage and processing of data.[3].

With that said, the importance of context has certainly not been ignored by statistically minded computer scientists. Indeed, Baroni et al. make a case for vector space approaches as a mechanism for "disambiguation based on direct syntactic composition" (Baroni et al., 2014a, p. 254), arguing that the linear algebraic procedures used to compose words into mathematically interpretable phrases and sentences in these types of models result in a systemic contextualisation of words in their pragmatic communicative context. Likewise, Erk and Padó (2008) outline an approach that models words as sets of vectors including prototypical lexical representations capturing information about co-occurrence statistics and ancillary vectors representing *selectional preferences* (*per* Wilks, 1978) gleaned from an analysis of the syntactic roles each word plays in its composition with other words. These composite vector sets are then combined in order to consider the proper interpretation of multi-word constructs of lexically loose or ambiguous nouns and

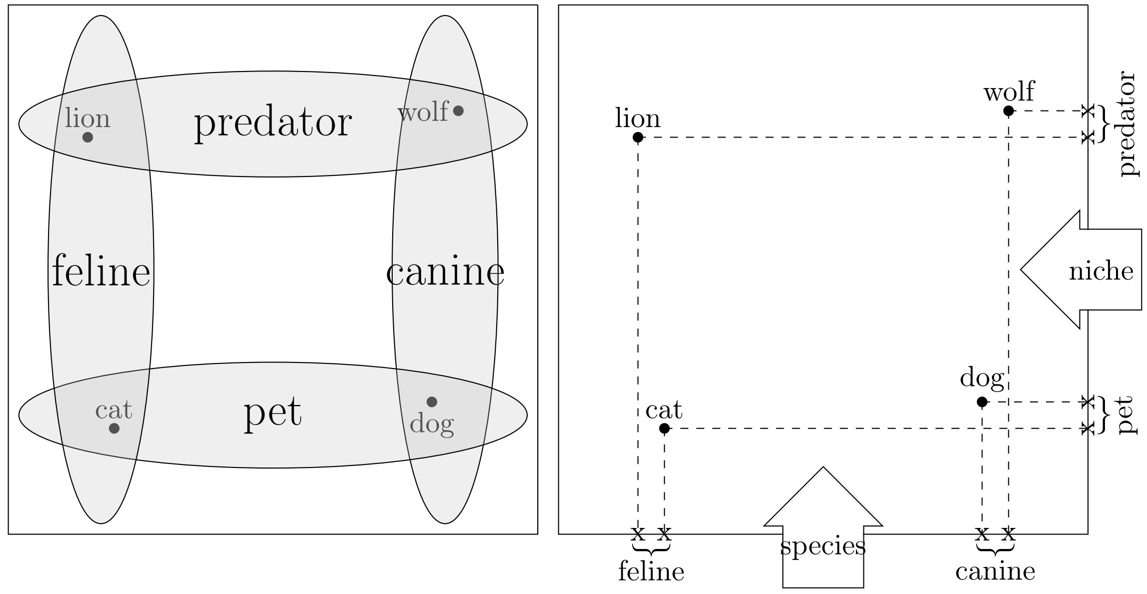
[3] It is perhaps not a coincidence that von Neumann was a seminal figure in the description of both the logic of lattice theory (Birkhoff, 1958) that has motivated more recent developments in concept modelling such as formal concept analysis (Wille, 1982) and the modular architecture of memory and processing components that defined computers in the period before the advent of highly parallel processing (von Neumann, 1945)

verbs. In subsequent work, the same authors describe a model which selects *exemplar* word-vectors from, again, composites of vectors, in this case extracted from observations of specific compositional instances of the words being modelled (Erk and Padó, 2010). In the first instance, composition is the mechanism by which word meaning is selectively derived, while in the second instance observations of composition are the basis for constructing sets of representational candidates to be selected situationally.

The model presented in this thesis is motivated by a premise similar to the one explored by Erk and Padó: there should be some sort of selectional mechanism for choosing the way that a lexical representation relates to other words in context. I would like to push this agenda even further, though. Following on Barsalou's (1993) insight into the *haphazard* way in which concepts emerge situationally, and likewise Carston's (2010) ideas regarding *ad hoc* conceptualisation, I propose that the mechanism for contextually mapping out conceptual relationships between representations of words should be as open ended as possible, ideally lending itself to the construction of novel conceptual relationships in the same way that the state space of possible word combinations offers an effectively infinite array of semantic possibilities. In particular, I will suggest that the ephemeral nature of concept formation can be modelled in terms of *perspectives* on the conceptual affordances of lexical relationships.

Figure 3.1 Illustrates this point. In a conceptual space of ANIMALS, we find subcategories such as PREDATORS and PETS, CANINES and FELINES, and we also find, not surprisingly, a degree of ambiguity which stretches the representation of these subcategories as contiguous, convex geometric regions. The distortion and overlap that occurs in Figure 3.1a is, however, resolved in Figure 3.1b by taking two different *perspectives* on the space. So, from one point of view, *dog* and cat collapse neatly into one cluster, while *wolf* and *lion* collapse into another. But through a shift in perspective, we discover another point of view from which *wolf* groups with *dog* and *lion* with *cat*. Crucially, the mechanism for achieving this trick of perspective taking is a matter of *dimensional reduction*: by aligning our own viewpoint in different ways, we can eliminate extraneous spatial information in the space and achieve a context specific interpretation of the relationships between word-points. This move of establishing a conceptual perspective by selectively reducing some of the spatial information available in our semantic space is one of the central components of my proposed methodology for semantic modelling, and much of this thesis will be spent evaluating the effectiveness of applying specific techniques for dimension reduction to higher dimensional spaces comprised of statistics about the situation of words in a large scale corpus.

Furthermore, the high dimensionality of vector space models of distributional semantics in particular should afford precisely these types of contextual viewpoints on potential



(a) a lexical space

(b) a contextual projection

Figure 3.1: In the two-dimensional space depicted in (a), the conceptual vagary of four words maps to overlapping, elongated and indeterminate spaces. In (b), two different perspectives on the lexical space, represented by the arrows labelled *niche* and *species*, offer contextualised projections in one-dimensional clusters which remit conceptual clarity.

relationships between words. Rather than depending on *a priori* disambiguation based on clustering or observations of context in the form of existing combinations of words, I propose that a technique for defining semantic subspaces *in situ* will capture the momentary and situated way in which concepts come about in the course of a cognitive agent's entanglement with the world. The way that relationships between words coalesce and then dissolve as we change our perspective on the space of this model is designed to reflect the way that concepts emerge dynamically in response to unfolding events in the world, and the ability to selectively specify the dimensional profile of a space of geometrically related semantic representations should enable just this kind of shifting of conceptual perspective. The theoretical mechanisms for making choices about multi-dimensional perspectives in semantic spaces will be discussed in the next section.

A Note on *Context* The term *context* has been used widely and varyingly by authors in both theoretical and computational linguistics, and with good reason, as various sense of the concept of context are clearly at play in any serious discussion of the interplay between language and cognition. Statistically minded computational linguists in particular, of whom I would like to count myself as one, have often used *context* to refer to the window of co-occurrence in which a word token is observed within a sample of text.

In his description of a co-occurrence statistic for measuring semantic similarity, Schütze (1992) introduced the term *context space* to refer to a space of co-occurrence dimensions, a terminology subsequently adopted by Burgess and Lund (1997) in relation to their HAL system. This notion of proximity within a text as context has persevered in the natural language processing literature.

Theoretical linguists and cognitive scientists, on the other hand, have tended to treat *context* as a much more general condition wrapped up with the entire perceptual, phenomenological aspect of existing as a cognitive agent in a complex world. So for instance Bateson says that "message material, or information, comes out of a context into a context," (Bateson, 1972, p. 404), meaning that there is an alignment between the inner context of an agent and the outer context of the world, while Grice's (1975) notion of *implicature* holds that meaning is somehow always determined in a context, with the exact nature of context remaining somewhat open-ended, and this nomenclature has been carried on by subsequent researchers interested in the idea that cognition, conceptualisation, and, correspondingly, language are always in some way specified by a situation in the world. The idea is that context is probably something that exists in large part outside of language, and almost certainly outside the informationally restrictive confines of word co-occurrences within a sentence.

Miller and Charles (1991) address this definitional vagary in the context of early work on distributional semantics, and specifically opt to use context to refer to the co-occurrences that occur on a purely lexical and sentential level. In my thesis, which seeks to address both those components of language measurable by an information processing system and the more general question of meaning as an environmentally situated phenomenon, I will endeavour to use the term *context* strictly in reference to the latter notion of the situation in which concepts and semantics emerge in tandem. With regard to words observed in proximity to one another, on the other hand, I will refer to *co-occurrence*, and so additionally to a *co-occurrence window* within which such observations are made and correspondingly a *co-occurrence statistic* as a measure of the relative frequency of such observations. Hopefully this terminological commit will serve to avoid confusion.

3.3 Literal Dimensions of Co-Occurrence

The model presented here is grounded within the paradigm of distributional semantics, which means that the conceptual geometries that it constructs are the product of observations of word co-occurrences in a large-scale corpus of textual data represented statistically. Two procedurally distinct methodological regimens have emerged from the recent study of distributional semantics. The first, and more established, approach involves tabulating word co-occurrence frequencies and then using some function over these to build up word-vector representations. With roots in the frequentist analysis described by Salton et al. (1975), recent research has typically involved matrix factorisation techniques presented as either (or both) an optimisation method (Bullinaria and Levy, 2012) or a noise reducing mechanism (Kiela and Clark, 2014).[4] A more recent approach, which has received a great deal of attention with the increasing availability of large-scale data and the corresponding advent of complex neural network architectures, involves using machine learning techniques to iteratively learn word-vector representations in an online, stepwise traversal of a corpus (Bengio et al., 2003; Collobert and Weston, 2008; Kalchbrenner et al., 2014). Baroni et al. (2014b) have described the former as *counting* and the latter as *predicting*, but it must be noted that both methods are very much grounded in observations about the co-occurrence characteristics of vocabulary words across large bodies of text.

Another important similarity between these two approaches is that they each in their own way move towards a representation of relationships between word-vectors which is to some extent optimally informative, and, by the same token, abstract. In the instance of neural network approaches, this is clearly the case due to the fundamental nature of the technique: the dimensions of this variety of model exist as basically arbitrary handles for gradually adjusting the relative positions of vectors, slightly altering every dimension of each vector each time the corresponding word is observed in the corpus. And, as far as models based on explicit co-occurrence counts are concerned, the favoured technique tends to involve starting with a large, sparse space of raw co-occurrence statistics (frequencies, or, more typically, an information theoretic type metric) and then factorising this matrix using a linear algebraic technique such as singular value decomposition. The result, in either case, is a space of vectors which exists just for the sake of placing words in a relationship where distance corresponds to a semantic property, consisting of dimensions which can only be interpreted in terms of the way that they allow the model to relate words, not in terms of their relationship to the underlying data. In fact, Levy and Goldberg (2014) have argued that recently developed neural network approaches just exactly recapitulate the process of matrix factorisation, and that a careful tuning of hyperparameters will generate commensurable results from either type of model.

A key feature of the methodology proposed in this thesis is that it maintains a base space of highly sparse co-occurrence statistics, which, despite their anchoring in the rel-

[4] Bullinaria and Levy (2012), Lapesa and Evert (2013), and Kiela and Clark (2014) have all reported that dimensional reduction techniques including SVD, random indexing, and top frequency feature selection generally do not improve results on word similarity and composition tests, with some notable parameter and task specific exceptions.

atively abstract realm of word positions in a digitised corpus, I will describe as *literal* in the sense that they can be interpreted as corresponding to actual relationships between word tokens in the world. As mentioned in the previous section, a fundamental objective of this methodology is to afford an abundance of potential perspectives on co-occurrence data. This objective is accomplished by providing a model with a corresponding proliferation of dimensions from which to make projections by way of context specific selections of subsets of dimensions. Furthermore, by maintaining the literal connection between the dimensions and the underlying data, the methodology likewise sustains a mechanism for selecting the dimensions in a way that is fundamentally interpretable, in that we can predict something about the geometric contribution of a given dimension to a subspace based on the types of words which tend to co-occur with that dimension. The co-occurrence profiles of the dimensions themselves will become an important criterion for dimensional selection, and having a very large set of such profiles to analyse will give a semantic model great scope in its capacity for adopting situational perspectives on the relationships between words.

So the proper framework for describing the model to be examined in this thesis is not so much a single space of word-vectors as a Grassmannian lattice consisting of the power set of all possible combinations of the dimensions characterising the base space. At the top of this lattice – the *join* – sits a single d-dimensional space consisting of every available one of the d co-occurrence terms observed throughout the underlying corpus. At the bottom of the lattice – the *meet* – sit d different one-dimensional spaces, each space corresponding to a single co-occurrence term. If the meet is considered layer-1 of the lattice, and the join is considered layer-d, then any given interstitial layer-j consists of every possible combination of j dimensions of co-occurrence statistics. A diagram of a very simple example of one such model is presented in Figure 3.2, illustrating the possible subspaces projected from a vastly simplified model consisting of just three co-occurrence dimensions (these particular spaces will be explored in the next section, providing the basis for the interpretable geometries illustrated in Figure 3.3).

An important distinction must be drawn, however, between the representation of my model as a lattice and the use of manifolds as an inferential mechanism. Formal concept analysis in particular has made a productive discipline out of applying lattice type structures to conceptual modelling, using the semi-hierarchical properties of lattices to capture logical relationships of entailment (Wille, 1982). That body of work takes as given that concepts are "the basic units of thought formed in dynamic processes within social and cultural environments," (Wille, 2005, p. 2). Widdows (2004) offers a broad overview of how this approach might be pursued through corpus linguistic techniques, while Geffet and Dagan (2005) and, more recently, Kartsaklis and Sadrzadeh (2016) have proposed

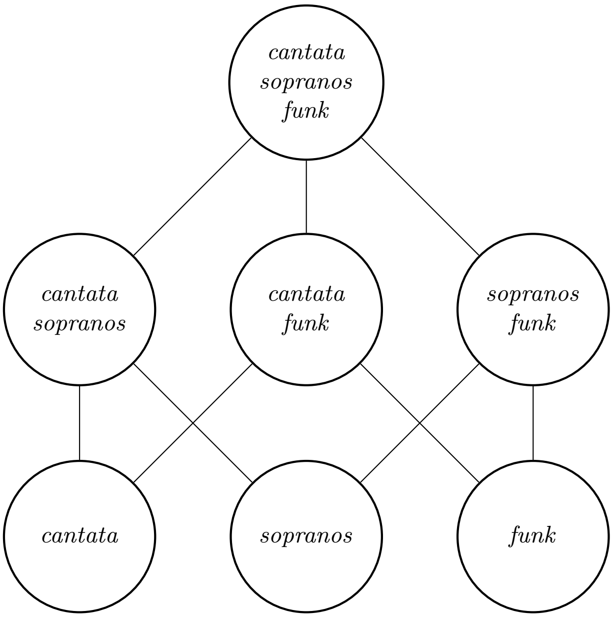


Figure 3.2: A lattice of three dimensions, including the two-dimensional subspaces which are used for analysing the conceptual geometry of a small set of word-vectors in Figure 3.3

statistical techniques using *feature inclusion* metrics to assess the potential entailment relationships between candidate words and corresponding concepts. The assumption inherent in this interesting work is that words are in some sense supervenient upon the concepts they denote, and that the statistical features of a language will by and large recapitulate the conceptual structure upon which it sits.

As Rimell (2014) has pointed out, however, it is problematic to assume that a spectrum of co-occurrence alone can indicate relationships of hyponymy and hypernymy. It stands to reason, for instance, that a word with a taxonomically specific denotation such as *bulldog* should probably have a co-occurrence profile including words omitted from the corresponding profile of a word like *lifeform*, which has an ostensibly more general extension—even excluding some of the ambiguity inherent in *bulldog*, it seems reasonable to talk about a *pet bulldog* but less so to talk about a *pet lifeform*, for instance. Rimell has proposed a measure of change in *topic coherence* as word-vectors are combined algebraically in order to detect entailment relationships. This measuring is achieved specifically through a process of dimension-by-dimensions comparison between potentially related word-vectors, in particular the *vector negation* method described by Widdows (2003), combined with topic modelling techniques to analyse the coherence of features distilled by the selectional process.

The methodology proposed in this thesis adheres to the same principle of fine-grained cross-dimensional analysis described by Rimell. In addition to the practical issues raised by Rimell, my approach is also designed to remain pointedly uncommitted to any claim that concepts are atomic or elementary to thought, or that language and concepts are involved in any kind of strictly hierarchical interrelationship. Instead, my models operate through an analytical traversal of lattices of subspaces in search of combinations of dimensions that capture conceptually *salient* profiles of co-occurrence features. If a consequence of this stance is that a model built from this methodology can't be understood in terms of nested, ordered relationships, though, then the question of how conceptual relationships do emerge situationally from the methodology remains. The next section of this theoretical overview will examine how the actual geometry of a projected subspace itself is expected to do this conceptual work.

3.4 Interpretable Geometry

It is important at this point to distinguish between two different modes of interpretability at play within the operation of the methodology I'm proposing. On the one hand, we have the process for selecting subspaces described above: this process requires a model composed of tractable dimensions of statistics that can be interpreted based on expectations generated from an analysis of some sort of contextually relevant information. Some specific mechanisms for this process will be discussed in the next chapter. Then on the other hand, once this selectional process has taken place, we find ourselves with a subset of dimensions defining a specific subspace. My claim is that, given the correct selectional criteria for performing this projection – this traversal of our lattice of vector spaces – we should be able to generate a subspace in which the projected word-vectors will be interpretable in terms of the actual geometric features of this subspace.

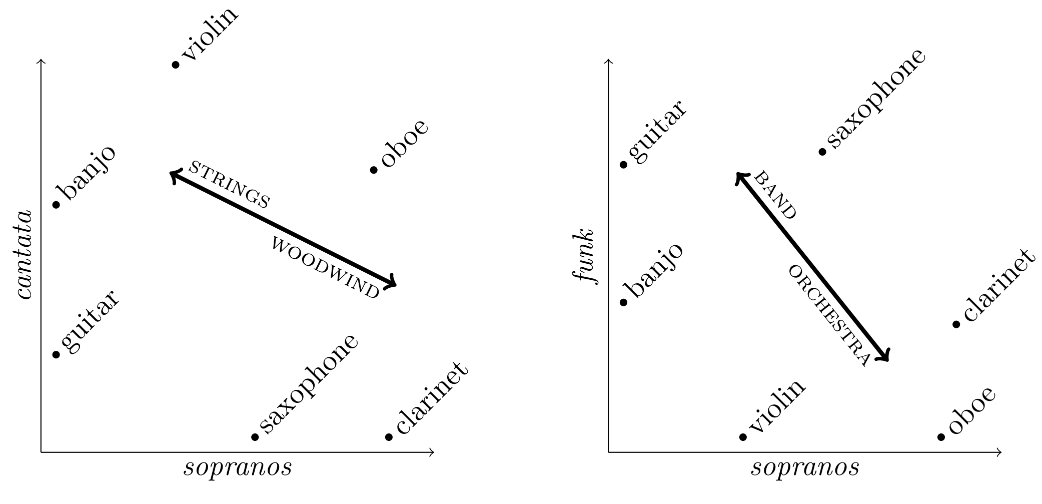
The idea of exploiting the geometry of a transformed space of word statistics is not new. Indeed, seminal work on latent semantic analysis was motivated by precisely the insight that a singular value decomposition of a high-dimensional, sparse matrix of statistical data about word co-occurrences would result in a dense lower dimensional matrix in which dimensions characterise *latent semantics* rather than literal word co-occurrences (Deerwester et al., 1990). Thus the linear algebraic methodology of generating a lower dimensional matrix of optimally informative dimensions arguably transforms a space of specific co-occurrence tendencies into a space of more general conceptual relationships. In fact, Landauer et al. have subsequently argued that the dimensional reduction by way of factorisation itself might directly mirror cognitive conditioning, modelling the way that the mind can "correctly infer indirect similarity relations only implicit in the temporal correlations of experience," (Landauer et al., 1997, p. 212).

Of course the dimensions of a factorised matrix are still not interpretable in themselves. They are, rather, an optimal abstraction of the underlying data, in which each dimension is maximally informative – and, accordingly, orthogonal – in comparison to the other dimensions. What we desire in a model, however, is a mechanism for actually interpreting directions and regions within a subspace projected by the model. This objective is motivated by Gärdenfors's (2000) insight into the inferential power of *conceptual spaces*: by building spaces in which the dimensions themselves correspond to *properties*, Gärdenfors has illustrated how features of points and regions within these spaces such as convexity and betweeness can be interpreted as corresponding to conceptual membership and can accordingly be used to reason about relationships between concepts. In more recent work, motivated by psycholinguistic insight into the significance of the *intersubjectivity* by which language facilitates the mutual ascription of cognitive content between interlocutors, Gärdenfors (2014) has proposed that semantics are derived from a communicative alignment of conceptual spaces.

A classic example of a Gärdenforsian conceptual space is the space of colours, which can be defined in terms of, for instance, hue, brightness, lightness, and colourfulness: any colour percept can be specified as a point corresponding to coordinates along each of these dimensions. Moreover, regions within the space of colours can be defined geometrically: the concept RED will correspond to a convex region within the space, and any point lying between two points known to be labelled *red* will likewise be considered RED. Jäger (2010) has devised an experiment mapping linguistic descriptions to conceptual regions precisely within the domain of colours. Taking a large set of multi-lingual data regarding colour naming conventions and treating each of 330 different colours as an initially independent dimension, Jäger demonstrated how an extrapolation of optimally informational dimensions via a principle component analysis revealed clusterings of colour names into convex regions.[5]

Similarly motivated by Gärdenfors's model of conceptual spaces, Derrac and Schockaert (2015) have built vectors of domain specific documents, associating word frequencies within documents with document labels. A multi-dimensional scaling procedure is then used to project these document-vectors into a Euclidean space in which the authors predict that properties such as *parallelness* and *betweeness* will correspond to conceptual relationships between documents. The authors demonstrate that geometry in their projected spaces does indeed afford conceptual interpretation: the vector they construct from large scale textual data for the word *bog* is found to be more or less between the vectors

[5] The cross-cultural universality of colour naming conventions presented by Kay and Maffi (1999), which Jäger takes as a basis for his research, is controversial to say the least – see Levinson (2001) for an alternative point of view – but Jäger's work remains a good example of a computational technique for extrapolating conceptual spaces from quantitative linguistic data.



(a) STRINGS vs WOODWIND

(b) BAND vs ORCHESTRA

Figure 3.3: Based on real co-occurrence data, swapping one dimension in a two-dimensional subspace reveals two different conceptual geometries.

for *heath* and *wetland*, for instance, and the vector for the film *Jurassic Park* lies in directions associated with DINOSAURS and SPECIAL EFFECTS. This work is particularly notable in that Derrac and Schockaert appreciate the significance of projecting spaces which are interpretable in terms of Euclidean distances rather than simply the cosine similarity of vectors extending from the origin of a space: Euclidean metrics provide a platform for more nuanced considerations of the relationships between points.

The type of space exemplified by the research of Jäger and Derrac and Schockaert is moving towards being a conceptual space in the way that its geometry offers itself up to semantic interpretation, but importantly these remain static spaces comprised of abstract dimensions, albeit dimensions generated in order to optimise the interpretability of the spaces they delineate. The objective of my model is to emulate the geometric interpretability of these other spaces in an extemporaneous, contextually dynamic way. To illustrate this point, consider the two spaces illustrated in Figure 3.3 (taken from real co-occurrence data, as described in the next chapter, and based on the lattice of subspaces illustrated in Figure 3.2). Here co-occurrence statistics are used to define three different dimensions, from which two different two-dimensional subspaces are selected with word-vectors plotted into each subspace. In each subspace, a particular conceptual geometry emerges, oblique to the axes of each subspace but nonetheless indicating distinct conceptual regions in which words align themselves in an interpretable way.

The first thing to note about these spaces is the way that swapping a single dimension in a two dimensional subspace can have a significant impact on the conceptual affordances of the subspace's geometry. Realigning the relationships between terms along a single axis leads to a complete shift in the groupings of terms, and, correspondingly, to the interpretation of regions and directions. If these are conceptually sound subspaces, then we might expect word-vectors found within the area of the triangle described by the points labelled *guitar*, *banjo*, and *violin* in Figure 3.3a to be the names of other string instruments, or other conceptually relevant terms. This is possibly asking too much of a subspace consisting of data regarding co-occurrences with just two terms across a large scale corpus, but as we scale up the dimensionality of the space – as we ascend the lattice of subspaces of a fully realised model – we can expect proper conceptual spaces to begin to coalesce.

The next thing to note is that the dimensions themselves are not especially interpretable. While these dimensional profiles are explicable – and indeed the ability to trace these statistics back to the corpus might turn out to be a desirable property for some applications – the dimensions themselves do not conform to Gärdenfors's (2000) notion of dimensions as representing the properties that compose a concept. It might be surprising, for instance, that the word *cantata* has a higher propensity for co-occurrence with the word *banjo* than with the word *clarinet*, given that cantatas have traditionally included parts for the latter but not the former. An examination of the underlying data, extracted, as described in the next chapter, from English language Wikipedia, reveals that the term *cantata* has been adopted, perhaps somewhat figuratively, by some bluegrass musicians, and so co-occurrences with *banjo* are indeed observed.

Rather than consider such usage as anomalous or attempt some sort of *a priori* word sense disambiguation, I propose to embrace the haphazardness of language and use it as a tool for projecting conceptually productive geometries. In fact it would be surprising if it turned out that in anything other than the most specialised cases we could simply pick dimensions based on their labels and then expect co-occurrence statistics to play out in a conceptually coherent way, as this would contradict the Relevance Theoretic thesis that language in use is always significantly underspecified. With this in mind, I suggest that we consider some set of dimensions, delineating a subspace and the corresponding geometry of word-vectors, to map precisely to a given context, and to effectively serve as the connective structure between language and conceptualisation. Under this regimen, the dimensions themselves become the constitutive substance of a context, but they do not compositionally define any context in which they participate; rather, the contextualisation is an emergent property of the combination of dimensions underwriting it, corresponding to *a way of speaking* about things.

The spaces illustrated in Figure 3.3 are the product of a survey of a lattice consisting of combinations of just three dimensions, and as such the conceptual affordances of this toy model are highly limited. As we add dimensionality to the model, however – as we observe more terms co-occurring with our vocabulary of word-vectors – we can expect an exponential growth in the combinatory possibilities of subspace construction. With enough dimensions from which to choose, and with an appreciable degree of variance between the profiles of each dimensions, there should be scope for projecting more or less any constellation of word-vectors we desire. The next question, then, is how to go about actually extracting a high dimensional base model of co-occurrence statistics from a large scale textual corpus and then explore the conceptual possibilities of this base space's inherent subspaces. The next chapter will answer this question.

Chapter 4

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.

4.1 Establishing and Analysing a Large Scale 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, Gabrilovich and Markovitch (2007) and Collobert and Weston (2008), 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 Baroni et al. (2014b), Pennington et al. (2014), and Gutiérrez et al. (2016) 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

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

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 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:

$$M_{w,c} := \log_2\left(\frac{f_{w,c} \times W}{f_w \times (f_c + a)} + 1\right) \tag{4.1}$$

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 4.1 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 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 (Levy and Goldberg, 2014, 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 (Socher et al., 2013), and also the processing of corpora, where part-of-speech and dependency tagging have been applied to positive effect (Padó and Lapata, 2007). Lapesa and Evert (2014) and Milajevs et al. (2016) 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

[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.

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.

4.2 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* (von Uexküll, 1957). In the world of physical entanglements, language presents itself with precisely the same open-ended opportunities for action as other modes of cognition (Clark, 1997; Gibson, 1979)—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, Hill and Korhonen (2014) and Bruni et al. (2014), 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 4-A, 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.

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

		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 4-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.

values across T:

$$n_h := \left\{ t \in t_h : \prod_{g=1}^{k} t_{g,i} > 0 \right\} \tag{4.2}$$

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

$$J := \left\{ f_{1...d} \in \operatorname*{argmax}_{f} \left(\sum_{g=1}^{k} \frac{M_{g,f}}{||m_g||} \right) \right\} \tag{4.3}$$

-

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:

$$I := \left\{ binB : t_{h,b} \geq \max_{d/k} t_h \right\} \tag{4.4}$$

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 4.2, 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:

$$Z := \left\{ n \in N : t_{h,n} \geq \max_{d/k} t_h \right\} \tag{4.5}$$

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 4-B, 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 4-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 Polajnar and Clark (2014), 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 Polajnar and Clark 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.

4.3 Exploring the Geometry of a Context Specific Subspace

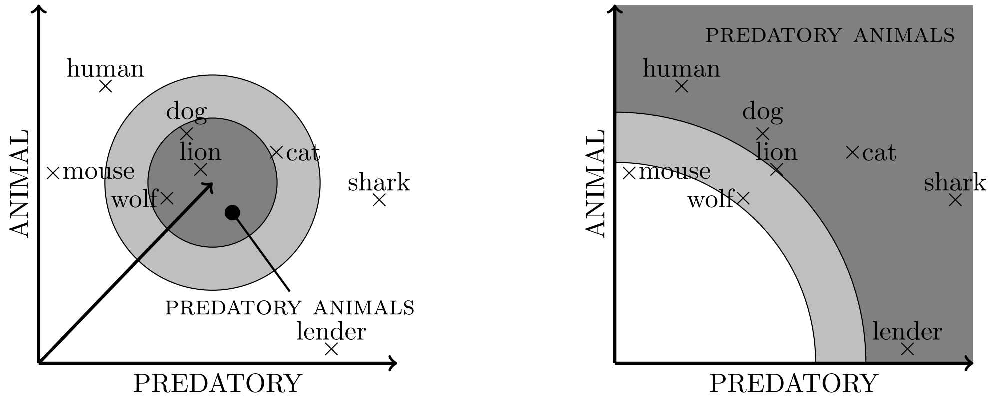
Before delving into the question of the types of geometries my method might be expected to generate, I would like to raise a point regarding the typical application of the term *geometry* to vector space models of distributional semantics in the first place. Widdows (2004) makes an enthusiastic and compelling case for the representational power of geometry, while Clark has pointed out that treating words as geometric features endows lexical representations with “significant internal structure” (Clark, 2015, p. 509) which can be applied towards modelling the meaning making compositionality of language.

Baroni et al. (2014a) 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 (Dummett, 1981) 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 3.1.

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 (Abdi and Williams, 2010). 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



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

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

Figure 4.1: 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.

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 4.3.2; first, though, I'll consider a couple basic measures for analysing word-vectors in context.

4.3.1 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 4.1. 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-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 4.1a, 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 4.1b, 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-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 Gärdenfors's (2000) notion of a region within a conceptual space, particularly in the case of the distance based metric illustrated in Figure 4.1a: 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

[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 4.1, 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 4-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*).

can be identified.

Examples of the tendencies of both norms and relative distances are explored in Table 4-C and Table 4-D, 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 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

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 4-D: The top word-vectors in subspace, as in Table 4-C but selected by input terms characteristic of PETS.

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.

4.3.2 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

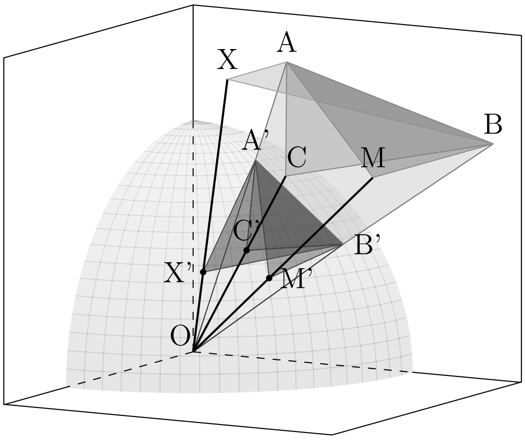


Figure 4.2: The geometric features of a subspace contextually projected based on an analysis of two input word-vectors.

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 phenomena.

Figure 4.2 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:

$$M := \{\mu(J_1),\ \mu(J_2)...\ \mu(J_d)\} \tag{4.6}$$

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

$$X := \{max(J_1),\ max(J_2)...\ max(J_d)\} \tag{4.7}$$

Finally, a generic central point C, a vector with all dimensions set to the same value, 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 4-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.

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:

$$C := \{\mu(M),\ \mu(M)...\mu(M)\} \tag{4.8}$$

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 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 4-E 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

[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.

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 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 4.2. 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 5-H. 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

[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.

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

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 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 5-H is in this regard a schematic for the computational experimentation and corresponding evaluation and analysis at the core of this thesis.

4.4 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 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 4.1, the scalars of the base model are defined by considering a ratio of frequencies approximately equivalent to a ratio of probabilities:

$$PMI(w, c) \approx \log \left(\frac{p(w, c)}{p(w) \times p(c)} \right) \tag{4.9}$$

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:

$$PMI(w_1, c) - PMI(w_2, c) \approx \log \left(\frac{p(w_1, c) \times p(w_2) \times p(c)}{p(w_2, c) \times p(w_1) \times p(c)} \right) \tag{4.10}$$

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

$$PMI(w_1, c) - PMI(w_2, c) \approx \log \left(\frac{p(c|w_1)}{p(c|w_2)} \right) \tag{4.11}$$

Next it must be noted that the geometry of the features described in Table 5-H are in large part derived from the vectors between the various points of interest – word-vectors as well as generic features – in a contextualised subspace. These vectors can now be understood as concatenations of logarithms of ratios of the pointwise conditional probabilities of the dimensions delineating a d dimensional context:

$$\overrightarrow{w_1} - \overrightarrow{w_2} \approx \left\{ \log \left(\frac{p(c_1|w_1)}{p(c_1|w_2} \right), \log \left(\frac{p(c_2|w_1)}{p(c_2|w_2)} \right) ... \log \left(\frac{p(c_d|w_1)}{p(c_d|w_2)} \right) \right\} \tag{4.12}$$

So from this perspective, the various features used to analyse the semantic situation of lexical representations in a contextualised subspace are, in fact, operations on conditional probabilities derived from observations of co-occurrence dimensions in the vicinity of target words. This then becomes a recapitulation of my hypothesis, namely, that there should be a mechanism for exploring how the semantic context in which word meaning comes about can be captured in terms of a way of talking about things, with this way of talking mapping more specifically to a set of conditional probabilities relating to the chances of finding a particular context term in the vicinity of a target word (or, indeed, the average or maximal probability of finding that context term, as with the non word-vector features of a subspace). Then the dimensional selection techniques proposed earlier in this chapter are now effectively three postulates about methods for discovering the set of co-occurrence terms which should be considered in the context of the conditions of target word-vectors and generic points in a subspace.

Furthermore, when we consider the various geometric features of a contextualised subspace as the independent variables of a model designed to classify or quantify a semantic phenomenon, we are in fact looking for weighted linear combinations of operations on conditional probabilities that maximise the correlation between those statistics and a set of dependent variables generally based on human observations. In Chapter **??**, for instance, a linear regression will be used to try to learn to predict human ratings of relatedness and similarity based on geometric features of subspaces, and in Chapter **??** a logistic regression will be used to similarly classify binary judgements of metaphoricity. At this point, the geometry of the subspaces generated by my methodology becomes not only a convenient mechanism for humans to use to visualise the relationships between various statistical spaces, but actually also a handle for an algorithm to selectively learn rather complex combinations of probabilistic features. A machine learning approach to analysing the geometry of a contextualised subspace then becomes a mechanism for iterating through inferential expressions formulated as operations on conditional probabilities, and an effective model will extrapolate an interpretable treatment of these probabilities directly from the geometry of a subspace.

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.

4.5 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, Schwartz et al. (2015) have developed a syntactic heuristic for identifying semantic similarity as compared to relatedness in particular, and Gutiérrez et al. (2016) 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 involves factorising my base space using singular value decomposition (SVD), abstracting the space into a smaller set of abstract dimensions representing axes of maximum variance between PMI values. 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.

4.5.1 Static Interpretations of the Base Space

Using the dimension reduction techniques described by, for instance, Deerwester et al. (1990) in the context of latent semantic analysis, it is possible to directly transform the same base spaces used for my context sensitive projections into a static model consisting of word-vectors defined along dimensions abstracted away from co-occurrence statistics in order to instead represent maximal axes of variance across the underlying data. The mathematical technique applied here is a low rank approximation of a singular value decomposition of the full blown co-occurrence matrix. To revisit this linear algebraic procedure, a $c \times d$ co-occurrence matrix M can be decomposed into three separate matrices, two orthonormal matrices U of shape $c \times r$ and V of shape $d \times r$ and a diagonal $r \times r$ matrix σ of eigenvalues, where r is the rank of M, such that M is the product of the decomposition:

$$M = U \Sigma V^T \tag{4.13}$$

In order to find an approximation of the variance between word vectors, a k dimensional matrix $U\hat{\Sigma}U'$ can be derived by setting all but the top k values in Σ to zero in $\hat{\Sigma}$. Since the highest eigenvalues in Σ will correspond to the orthonormal decomposition of dimensions with the highest variance between word-vectors, the resulting lower dimensional matrix will contain maximal information about interrelationships between word-vectors. Some authors, including Deerwester et al. and, more recently, Turney and Patel (2010) have argued that the dimensions of such an approximation can be understood to correspond to conceptual axes across the data.

Of course, as mentioned in Chapter 3.3, the matrices derived through such a process of factorisation and recomposition effectively abstract away from any interpretability in terms of their dimensions, which now just represent orthogonal axes of maximal variance, and so they are insusceptible to my methodology for contextual dimensional reduction. My case is that, when it comes to deriving spaces where the conceptual underpinnings of semantics play out in terms of geometric relationships between lexical representations, the geometries necessarily must be supplied in a context specific, online manner. Gauging the difference in performance between the SVD decomposition of my base spaces and the contextualised subspaces generated using the dimension selection techniques described above will provide a basis for comparing the extent to which each approach really does manage to extract conceptually significant relationships from the underlying co-occurrence data.

Because my base spaces are sparse and positive, the dense matrix resulting from the operation of an SVD approximation is skewed from the centre of the resulting lower dimensional space. To compensate for this, I take a final step in order to facilitate the calculation of semantic relationships between words in terms of the angular situations of the corresponding word-vectors: I translate and then scale the matrix by performing mean zero, standard deviation one normalisation across all dimensions of the reduced matrix. This means the reduced space resembles something very much like the hyperspheres derived from the neural network approach to distributional semantics which will be described in the next section, and, as will be seen in the experiments carried out over the next three chapters, it has an interesting impact on model output.

4.5.2 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 Mikolov et al. (2013a) 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 (Erk and Smith, 2016). 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 (Kiela et al., 2015) to analogy completion (Mikolov et al., 2013c), and have also been applied to multimodal tasks such as image labelling (Kottur et al., 2016).

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 Mikolov et al., 2013b, 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.

4.6 A Proof of Concept

sec:pof In this section, I present a preliminary experiment performed using my contextually dynamic distributional semantic model. This experiment, conceived as a proof of concept, involves using multi-word phrases as input and evaluating my methodology's capacity for building subspaces where words associated with the conceptual category denoted by the input term can be reliably discovered. The experiment expands upon the notion of proto-conceptual spaces outlined in Section 4.3.1, examining 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 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 Agres et al. (2015). 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 van der Velde et al., 2015). 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 4, but without any rigorous evaluation.

Reviewer responses to a subsequent journal article (McGregor et al., 2015), 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 (Fellbaum, 1998). 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. There is precedent for the construction of *ad hoc* datasets from WordNet, with for instance Baroni et al. (2012), Riedl and Biemann (2013), and Melamud et al. (2014) all mining the extensive lexical taxonomy for gold standard entailment relationships. My 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.*[6]

[6] 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.

4.6.1 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. The 12 synset labels are itemised in Table 4-H.

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 4. 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 4.3.1: 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 4.5.2. 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 4.5.2. 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.

		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-G: 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.

4.6.2 Results and Analysis

Results for the set-up described in the previous section can be found in Table 4-G, 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 measure generated relevant but not always categorical terms such as *wild*, *giant*, and *golden*. To give an example from the data used for this experiment, top-50 results from the JOINT distance measure returned for the input (*body, part*) include words like *portion*, *upper*, *shape*, and *whole*, while the results from PHYSICAL PROCESS include *method*, *complex*, and *affect*—so, terms that are conceptually relevant to the target domain but are not strictly part of the category BODY PART. We might characterise this trend 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.

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 4.1a 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.

Table 4-H presents accuracy rsults for each of the 12 conceptual categories targeted by this experiment, focusing on the two measures applied to JOINT type subspaces as well as the bag-of-words version of the `word2vec` methodology. It's particularly pleasing to see my methodology handling the ambiguity inherent in the inputs (*body, part*) and (*physical, process*) so well as it finds the relevant terms very far from the origin, while, as discussed above, the distance measure falls short here, presumably because it is finding

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

Table 4-H: Item-by-item accuracy results for the entailment experiment run on WordNet synsets, reported for the norm and distance metrics using the joint technique as well as `word2vec's` bag-of-words method.

terms that are related to the input rather than terms that are entailed by it. On the other hand, the distance measure does quite well for inputs such as (*psychological, feature*) and (*human, action*). A pitfall for the norm measure and the bag-of-words method is that they both seem to have identified a region of psychological [thriller] feature [film], yielding outputs such as *slasher*, *offbeat*, and *blockbuster*, so there is clearly still scope for ambiguity here even with a degree of context. It's interesting to observe how the norm measure manages to recover from this category error as it returns more results, whereas the bag-of-words method evidently wanders further off topic. That said, the bag-of-words results are impressive, at least in the top 50 outputs, for the inputs (*social, group*) and (*taxonomic, categories*), arguably instances where the context is already somewhat evident with one of the two inputs.

These are, on the whole, promising results for my methodology. They illustrate its ability to delineate a context specific subspace based on a conceptually targeted input and then discover regions within this space that evidence a degree of conceptual inclusion. Furthermore, the regions discovered seem to be relatively well defined, with a lesser degree of dithering away from the top or centre of the regions compared to a standard static semantic model. On the other hand, the outputs from these regions are marked by an different kind of ambiguity than polysemous word senses: there is a confusion between words which denote entities entailed by the input, and words which simply relate to the input. The next section will expose the methodology to a group of datasets that have already been broadly reported in the computational linguistic literature, with the objective of establishing precisely the ability of context sensitive models to make distinctions between similarity and relatedness.

Chapter 5

Relatedness and Similarity

In Chapter 3, I laid out the theoretical groundwork for statistical context sensitive models of lexical semantics, and in Chapter 4 I described the actual methodology for building such models, accompanied by a preliminary proof of concept involving conceptual entailment. In this chapter, I will present the first set of experiments designed to evaluate the utility of this methodology. These experiments are intended to probe the productivity of a context sensitive, geometric approach to building a computational model of lexical semantics based on statistics about word co-occurrences. Beyond testing my models' performances on some well-travelled datasets, this will provide an opportunity to explore whether different components of the methodology and, moreover, different aspects of geometric output lend themselves to modelling related but distinct semantic phenomena.

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

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

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

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

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

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

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

5.1 An Experiment on Relatedness

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

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

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

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

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

[3] The standard approach in the empirical literature on word relatedness and similarity has been to report Spearman's correlations rather than Pearson's correlations, and I will follow suit here. The presumption is, perhaps, that word similarity is always relative—more on this in Section 5.4.

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

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

5.1.1 Relatedness: Methodology and Model

In order to test the ability of my statistical methodology to model relatedness, I build JOINT, INDY, and ZIPPED subspaces using each of the 353 word pairs in the WordSim data as input. I project subspaces of 20, 50, 200, and 400 dimensions, extrapolated from base spaces built using 2x2 and 5x5 word co-occurrence windows. For each subspace, I extract the geometric features listed in the previous chapter in Figure 4.2 and Table 5-H. I normalise each feature across all word pairs to have a standard normal distribution, and then I use these normalised features as the independent variables of a least squares linear regression, taking the WordSim rating of each word pair as the dependent variable. The relatedness ordering of word pairs inherent in the scores assigned by the regression are then compared to human WordSim ratings in terms of Spearman's correlations, as is standard practice in the NLP literature. Results from my model are compared with results from singular value decompositions of my base space using comparable parameters, as well as `word2vec` skip-gram and bag-of-words models, again using commensurable parameters.

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

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

Table 5-A: Spearman's correlations for word ratings output by a linear regression model of the WordSim data for various subspace types and model parameters, compared to the correlations for cosine similarities output by static models using comparable parameters.

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

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

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

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

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

5.1.2 The Geometry of Relatedness

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

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

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

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

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

So it seems that there is a substantial risk of overfitting the data given the quantity of information being extracted from the geometry of my subspaces. In order to get a sense of what's actually happening in these models, I produce Spearman's correlations between the WordSim data and each of the features of different subspaces independently. The top five features for 400 dimensional JOINT, INDY, and ZIPPED spaces generated using 2x2 word co-occurrence windows are reported in Table 5-B. The first thing to note here is that angular measures are significantly predictive for all three dimensional selection techniques—but not the angles that may have been expected based on static distributional semantic models. Where the SVD and `word2vec` results reported in Table 5-A are based on cosine similarity between word-vectors, in my subspaces, the angles at the vertexes of the generic vectors C and M in particular seem to be predictive for all dimension selection techniques, with the measure $\angle AOB$, corresponding to cosine similarity, only figuring as the fifth most predictive feature for INDY type subspaces. All correlations here are positive, which means that words are more likely to be related as their corresponding word-vectors move closer to one another relative to their relationship to the points C and M.

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

Beyond the angular relationships between word-vectors and generic vectors, in the case of JOINT subspaces in particular, and also to a lesser extent ZIPPED subspaces, the mean norm of the word-vectors $\mu(A, B)$ correlates positively with relatedness, both alone and as the numerator of fractions where the norms of generic vectors are denominators. This corroborates the findings regarding the relationship between conceptual entailment and word-vector norm presented in Chapter 4.3.1: in an appropriately contextualised subspace, distance from the origin is indicative of conceptual pertinence. This result can be interpreted as meaning that, in subspaces constructed from dimensions containing co-occurrence information about both words being analysed, mutually high PMI scores are indicative of higher degrees of relatedness. In other words, words that tend to have the same terms at the high end of their co-occurrence profiles also tend to be related. It is interesting, then, that this measure isn't more predictive for INDY type subspaces as well, where we might expect that the independent selection of dimensions that are informative about one word and happen to be informative about another word would indicate a strong degree of relatedness and also result in word-vectors with large norms. But these results clearly indicate that, in subspaces delineated by the concatenation of independently derived dimensions, it is the relative situation of word-vectors on these dimensions and correspondingly angular measures that point to relatedness.

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

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

Table 5-C: A comparison of Spearman's correlations returned by various models, including my optimal $\angle ACB$ measure.

subspaces. It would actually make more sense to see this effect in the INDY spaces, where the potential for selecting dimensions with unusual profiles based on a single input word, potentially leading to geometric strangeness, is higher. On the other hand, it may be the case that there is a more dynamic interaction between the various features of these spaces. This supposition will be addressed with regards to semantic similarity in particular in the next section, and then will be examined comparatively in terms of similarity and relatedness in Section 5.3.

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

5.2 An Experiment on Similarity

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

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

In this section, I apply my own methodology to the SimLex data in order to investigate the extent to which context specific subspaces of word-vectors can accurately represent the similarity between words. As with the previous experiment exploring word relatedness, a primary objective here is to test the extent to which the geometric features of my subspaces both collectively and independently align with human ratings. In addition to performing a linear regression mapping the full sets of geometric features generated for various combinations of parameters and likewise comparing the correlation between individual features and human similarity ratings, here I will also attempt to extract a set of features which optimally predict similarity while avoiding collinearity and without overfitting the resultant model. This approach will offer a mechanism for interpreting the dynamics at play between different features of contextualised statistical subspaces.

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

5.2.1 Similarity: Methodology and Model

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

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

[6] The role of frames in word association judgements will be discussed in more detail in Section 5.4.

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

Table 5-D: Spearman's correlations for word ratings output by a linear regression model of the SimLex data for various subspace types and model parameters, compared to the correlations for cosine similarities output by static models using comparable parameters.

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

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

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

5.2.2 The Geometry of Similarity

Next, as with the relatedness data in the previous experiment, in order to escape overfitting and explore the particular statistical geometry of similarity in context specific co-occurrence subspaces, I consider the predictive capacities of independent geometric features. Table 5-E reports the Spearman's correlations of the five most predicative features for each dimensional selection technique used to pick 400 dimension from a 2x2 word co-occurrence window base space. The features that independently emerge are strikingly similar to those found to be most predictive of relatedness: for the INDY subspaces, a number of different cosine measures, including angles of the vetors converging at the vertexes of generic vectors and the normalised versions of these angles, as well as the cosine similarity between the word-vectors, all correlate positively with similarity, meaning that as these angles grow smaller, the words in question tend to be more similar. Angles are also seen to be predictive of similarity in the JOINT and ZIPPED subspaces, though here the distance from the norm inherent in fractions involving $\mu(A, B)$ as the numerator are even more strongly predictive than before.

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

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

Table 5-E: Independent Spearman's correlations with SimLex data for top five features of each subspace type for 2x2 word co-occurrence window, 400 dimension subspaces.

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

But the most important thing to note here is that the correlation scores for these independent features are significantly lower than the scores achieved by the multi-variable linear models reported in Table 5-D. This is in contrast to the relatedness results, where the difference in correlation with human ratings achieved by the top feature and the linear model learned from all 34 features were so close that the difference was statistically insignificant. This serves first of all to reiterate a point that has already been made: where judgements of general relatedness can be extrapolated in a fairly straightforward way from a comparison of co-occurrence statistics, the more particular quality of similarity does not yield as readily to the direct quantification of co-occurrence. The critical question, then, is whether there is a combination of geometric features which, in an appropriately contextualised subspace, will reliably indicate semantic similarity between the terms used to generate that subspace—and, if so, whether we can interpret that combination of features in a way which is theoretically productive.

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

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

Table 5-F: The optimal combination of seven non-correlated features for a linear regression modelling SimLex data for 2x2 word co-occurrence window, 400 dimensional subspaces projected using each dimensional selection technique.

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

Rather than exhaustively searching the state space of combinations of features, I treat the discovery of feature combinations as a beam search problem, returning the top 1,000 performing combinations, in terms of Spearman's correlation, for each number of features progressively and then exploring the contribution of adding each of the remaining features to each of these optimal combinations. The top combinations of seven features for each dimensional selection technique, projecting 400 dimensional spaces based on the 2x2 word window base space, are detailed in Table 5-F (leave-one-out Spearman's correlations with human ratings level out with more than seven features). The Spearman's correlations reported here are once again based on a leave-one-out cross-validation, and, unlike with the relatedness data, reveal a marginally significant improvement over the best performing independent features ($p = .166$ in the case of the combined feature score for the INDY type subspaces versus $\angle ACB$ alone, the top feature reported in Table 5-E). These scores are, on the other hand, substantially lower than the scores derived from the coefficients of determination of a linear model trained on all features ($p = 0.049$). So this process of feature combination discovery reveals that, on the one hand, there is something to be gained by considering the overall statistical geometry of a subspace, and, on the other hand, there is a degree of overfitting at play in the full blown linear model.

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

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

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

Table 5-G: A comparison of Spearman's correlations with SimLex data reported for various models, including my optimal INDY technique.

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

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

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

First, and finally as far as this experiment on word similarity is concerned, Table 5-G offers a comparison between a sampling of results from the literature (and it should be noted that, due to it's relatively newness, the SimLex data has not yet received as much attention as the WordSim data, though there is a growing body of relevant work emerging). Clearly approaches involving the application of heuristics, such as Schwartz et al.'s (2015) trick of mining syntactic patterns specifically indicative of similarity, Banjade et al.'s (2015) construction of a regression based on the output of a variety of models, or Recski et al.'s (2016) recourse to a structured knowledge base do significantly better than my methodology. But again, as with the relatedness experiment described in the previous section, my interest here is not merely in pursuing quantitatively strong results but also in exploring the ways in which models derived from raw word co-occurrence data can be mapped to semantic phenomena and used to explore their cognitive underpinnings (more on that in the next section). If anything, the results here indicate that similarity is clearly a complex phenomenon requiring a great deal of nuance for detection through statistical means, and an expansion of the features used to explore the words that humans deem to denote things that are alike may be in order in future work.

5.3 Comparing the Two Phenomena

The results for correlations between independent geometric features and ratings of relatedness or similarity presented in Tables 5-B and 5-E would at first pass seem to largely refute the hypothesis presented at the beginning of this chapter: the same angular and norm features predict both phenomena in similar ways in similar subspaces. Furthermore, the predictions are substantially more reliable for relatedness than they are for similarity, suggesting that these statistics reflect co-occurrence tendencies that are primarily indicative of a general pattern of semantic association and then only incidentally indicative of similarity to the extent that being similar is a special case of being related, meaning that word pairs that are similar will necessarily tend to receive higher ratings than word pairs that are unrelated. The combinations of non-correlated features obtained in Table 5-F, however, tell a slightly different story. While the best way to bluntly predict similarity based on a single statistical feature might be to guess that words that are related might also be similar, there seems to be a meaningful combination of features that collectively indicates similarity in a way not independently obvious in any of its constituents. The question, then, is whether there is a similarly dynamic and at the same time distinct combination of features indicative of relatedness.

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

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

Table 5-H: Comparison of most predictive features for relatedness and similarity in both JOINT and INDY type 2x2 word window, 400 dimensional subspaces, with models optimised for leave-one-out cross-validation.

to discover the combination of seven non-collinear features that achieve the highest Spearman's correlation for the WordSim data. The results are reported in Table 5-H. In the end, angles play an important role in predicting both phenomena, with the angle between vectors $\angle AOB$ being especially indicative of similarity: word-vectors with a similar ratio of PMI values across the set of dimensions they choose are more likely to be considered similar. The offsetting of the positive correlation with the angle $\angle ACB$, formed by the points corresponding to the word-vectors at the vertex of point C, for relatedness by the negative correlation for the angle $\angle A'C'B'$ by normalised versions of the same points suggests that related word-vectors tend to be close to one another relative to their distance from C but at the same time on either side of the central line defined by C. A similar effect can be observed for similarity, where word-vectors tend to pass on either side of the line defined by M, which can be thought of as a kind of weighted centre line, but on the same side of the potentially less central line defined by X.

The really interesting thing to note here, though, is that, outside of angular measures, the two different semantic relationships tend to be associated with different sets of geometric features. Relatedness is strongly associated with ratio type features, with the negative correlation with $\overline{A'C'} : \overline{B'C'}$ indicating that one related word tends to be significantly closer to the centre line than the other in INDY subspaces (this is also supported by the observation above regarding the negative correlation with $\angle A'C'B'$). Returning to the mathematical analysis of Chapter 4.4, the ratios involve a fraction of the norm of a vector of differences between PMI values: so, the likewise negatively correllated ratio $\overline{AM} : \overline{BM}$ involves the difference between scalars of word vectors and mean values of corresponding dimensions, so $\overline{AM} = \sqrt{\sum(A_i - M_i)^2}$ for all dimensions i in a given subspace. The difference $A_i - M_i$ is, in turn, per Equation 4.11, can be understood as a logarithm of a ratio of probabilities, in this case the conditional probability of the term associated with i co-occurring with the word associated with A versus the average of all such probabilities across i. Because the values are squared, it doesn't matter which probability is the numerator and which the denominator; the important thing here is that relatedness correlates with a larger differential in the ratio of the conditional probabilities of each selected dimension co-occurring with each word and the average conditional probabilities of co-occurrence across all these dimensions. This is all to say that related words tend to choose subspaces where one of the words is considerably closer to an average co-occurrence profile than the other, which suggests that the relatedness models may be picking up on situations where an exemplar is judged related to a prototype, or a component is considered related to a whole.

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

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

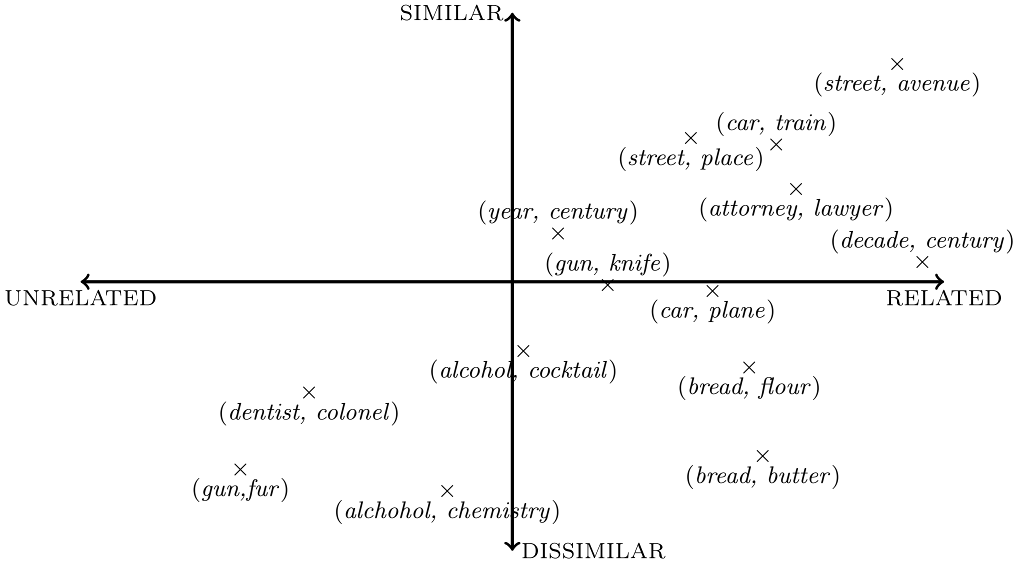


Figure 5.1: Noun pair scores along axes of relatedness and similarity as returned by a model built from features of 2x2 word co-occurrence window, 400 dimensional, INDY type subspaces.

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

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

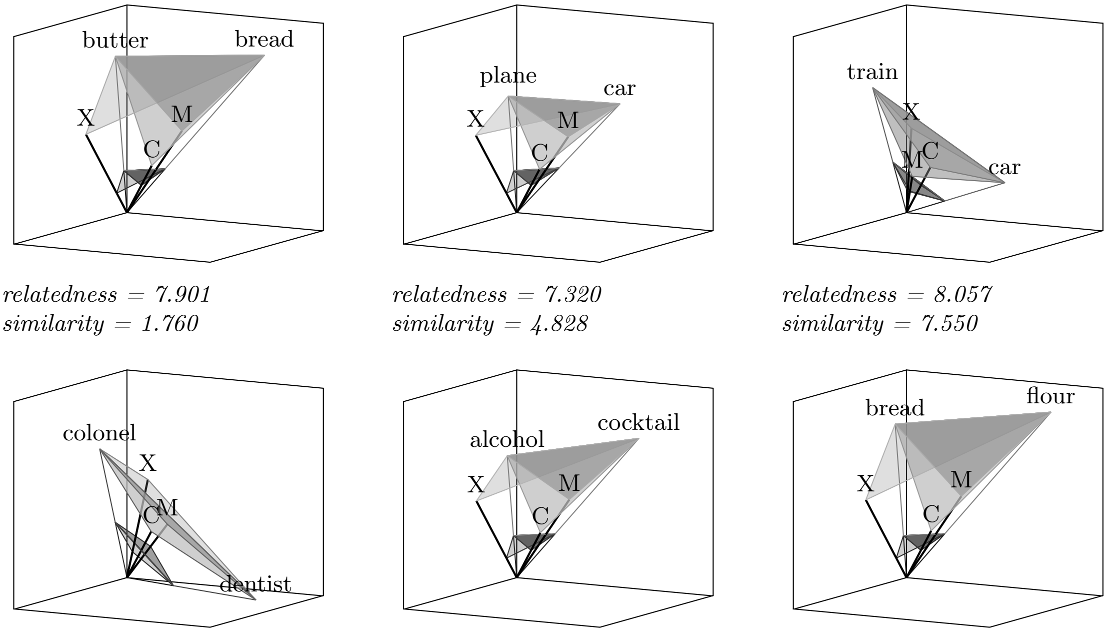
[7] Levy et al. (2015) have similarly proposed that success in distributional semantic models capturing entailment relationships is in fact down to their ability to identify *prototypical hypernyms* that are simply more likely to be identified as categorically containing some other unseen word—but those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

features are predictive of relatedness versus similarity.[8] This is illustrated in Figure 5.1, where a selection of word pairs from both the WordSim and SimLex datasets are projected along axes of relatedness and similarity based on the outputs of the respective models learned based on the geometric features of 2x2 word window, 400 dimensional INDY subspaces. So, for instance, *bread* is considered fairly related but not at all similar to *butter*; *flour* is rated as being about equally related to *bread* as *butter*, but somewhat more similar. Similar trends are observed in the progress from (*car, plane*) to (*car, train*) and (*alcohol, chemistry*) to (*alcohol, cocktail*). Meanwhile, and perhaps less explicably, *year* and *decade* are about equally similar to *century*, but *decade* is modelled as being considerably more related. The emptiness of the upper-left region of the field in this selection is characteristic of the models overall: words that are similar are in general *de facto* related to one another, but *relatedness* does not conversely predict similarity.

Figure 5.2 presents an assortment of renderings of three dimensional projections of 400 dimensional subspaces chosen from across the spectrum of both similarity and relatedness ratings as returned by the INDY technique operating on the 2x2 word window base space. The projection to three dimensions preserves the distance of the word-vectors and the generic vectors from the origin, as well as the angles between each vector, keeping the centroid vector C in the centre of the positive region of the space. It should also be noted that the norm of the vector X is scaled by a factor of 0.5 for the sake of visibility. The objective of these renderings is to offer an impression of the shifts in the overall comportment of the statistical geometry of subspaces moving along axes of both relatedness and similarity.

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

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



relatedness = 7.901
similarity = 1.760

relatedness = 7.320
similarity = 4.828

relatedness = 8.057
similarity = 7.550

relatedness = 2.646
similarity = 2.940

relatedness = 5.129
similarity = 3.714

relatedness = 7.744
similarity = 3.406

Figure 5.2: Subspaces, including word-vectors and generic features, derived from word pairs with an assortment of relatedness and similarity scores.

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

5.4 Frames of Similarity

Tversky (1977), in his psychologically motivated reflections on the geometry of similarity, observes that relationships of similarity are fundamentally not symmetric: there tends to be a preference to consider the specific more similar to the general, and the peripheral more similar to the prototypical, than the other way around. So, to use Tversky's own example, *ellipse* is more similar to *circle* than *circle* is to *ellipse*; we might extend this conjecture to predict that *wolf* is more similar to *dog*, *radiologist* more similar to *doctor*, and *limping* more similar to *walking* than the converse propositions. Indeed, the frequentist axiom extrapolated through the geometric analysis of the previous section, stating that more common words denote things that are more likely to generally be a component of a similarity relationship, is broadly in line with this observation. Tversky makes the point that the conventional conditions of geometric relationships – *minimality*, *symmetry*, and *the triangle inequality* – do not pertain in the case of similarity judgements, a point which if taken seriously serves to foil the project of a vector space model of word similarity.

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

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

These types of conceptual contexts are evident in the relatedness and similarity datasets which have been explored in this chapter. In the SimLex data, for instance, (*dentist*, *colonel*) is rated as one of the least similar word pairs at 0.40, while (*attorney*, *lawyer*) is, at 9.25, considered one of the most similar pairs. The difference seems reasonable enough in terms of a comparison between the two pairs, but the low rating of (*dentist*, *colonel*) leaves little room for either dentists or colonels to be even less similar to, say, gorillas, or electricity, or democracy, and so forth. What seems to be happening here is that human evaluators are identifying an implicit conceptual frame in which each word pair is to be evaluated: in the case of attorneys, lawyers, dentists, and colonels, the frame is something like PROFESSION, and so the professional activities of colonels and dentists are judged to be more or less orthogonal, while attorneys and lawyers pursue very similar careers. The inclusion of some additional comparison, for instance (*dentist*, *grandparent*), would suggest a broadening of the conceptual frame to something like HUMAN, and a corresponding drawing together of words denoting professions in particular.

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

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

Faruqui et al. (2016) raise a number of issues with relatedness and similarity datasets, among them the uncertainty surrounding specific semantic phenomena and the lack of applicability of quantified word pair scores to practical NLP tasks. Those authors ultimately propose that quantitative evaluations of vector space models of word meaning should avoid claims of generality, instead treating particular models as task specific implementations. There is something to be said for this approach, and even more to be said in support of the effort to apply statistical NLP techniques to activities in other fields where heterogeneous data and contextual complexity present potentially confounding factors to the relatively abstract and rigid representational structures of distributional semantic models. All the same, I maintain that word association tasks, particular a battery of tasks spanning a variety of semantic phenomena, can be a productive tool for exploring the capabilities of a methodology, and present the work that has been described in this chapter as a case in point.

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

The INDY dimensional selection technique in particular would lend itself to this type of programmatic extension of research into semantic relatedness, as it facilitates the open-ended concatenation of dimensions from an analysis of an arbitrarily large set of constituent word-vectors (the JOINT and INDY techniques, on the other hand, would presumably return increasingly uninteresting dimensions with universally non-zero values as the set of input words expands). A subspace built using the INDY technique based on an analysis of a set of words denoting, for instance, constituents of the concept PROFESSIONALS would acquire co-occurrence dimensions specifically salient to each of the input terms, and the construal of other word-vectors in the space along the collective profile of dimensions would, I forecast, be indicative of their conceptual situation according to the various properties of being a professional. In such a space, we might predict that we would find, for instance, *surgeon* somewhere in the vicinity of the region between *barber* and *butcher*

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

International Conference on Machine Learning, pages 160–167.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics Volume 3: Speech Acts*, pages 41–58. Academic Press, New York.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Processing.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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