Stage Two Report: A Geometric Method for Context Sensitive Distributional Semantics

by

Stephen McGregor

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

Department of Electronic Engineering
Queen Mary, University of London
United Kingdom

July 2017

Abstract

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

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

from large-scale textual corpora.

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

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

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

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

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

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

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

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

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

Table of Contents

Chapter 1

Semantics on the Fly

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). 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 (Evans, 2009). 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 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. This structure-oriented approach to language and mind epitomises a project that Dreyfus (2012) has described as "finding the features, rules, and representations needed for turning rationalist philosophy into a research program," (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 language 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 three interrelated desiderata, derived generally rather than in a one-for-one way from the points enumerated above:

1. The method should be dynamically sensitive to context;

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

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

The first 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 2, 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 2, 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 context. 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. First, however, I will briefly consider the nature of semantic representations.

1.1 Lexical Semantics

This thesis is primarily concerned with the problem of semantic representations, and in this regard finds itself in good philosophical stead. **?**, 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" (p. 192), which is to say that things like words acquire semantics situationally. **?** engages with Russell's banner 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 (**?**), 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 work described in my thesis finds its roots even deeper in the tradition of the philosophy of signification, in the semiotics of **?**, who suggested 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," (¶230).

not only that symbols like words can represent things in the world, but that this representation unfolds through the actual dynamics of the symbol itself, though the

Ultimately, then, the methodology described and explored in this thesis represents an

attempt to move computational approaches to natural language modelling toward the

semantics of **?**, who

1.2 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 taken into account 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.[1].

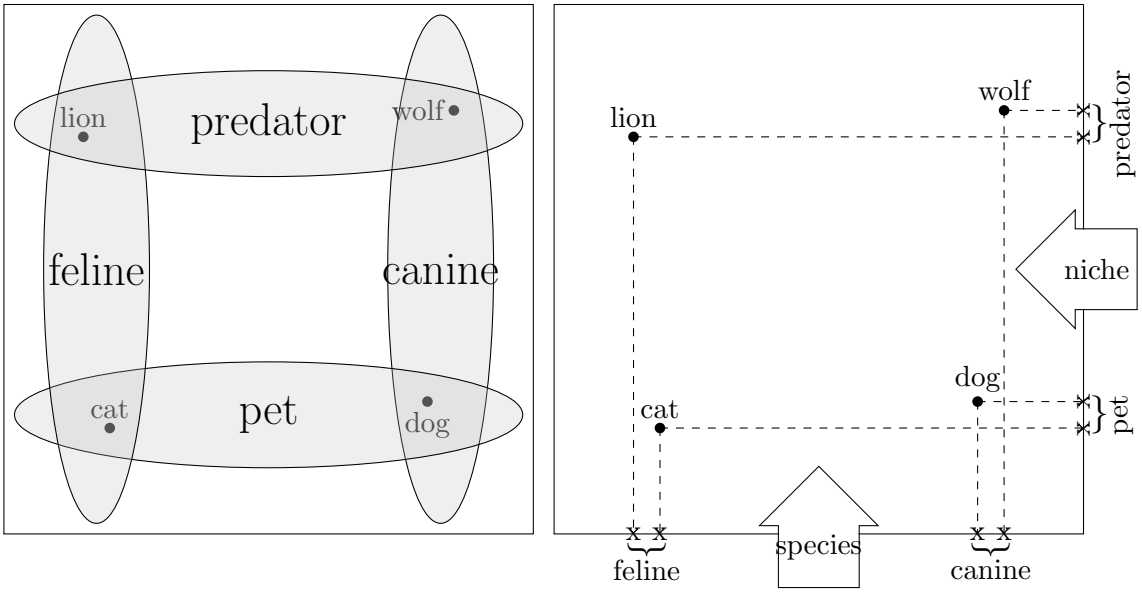
With that said, the importance of context has certainly not been ignored by statistically minded computer scientists. Indeed, Baroni, Bernardi, and Zamparelli (2014) make a case for vector space approaches as a mechanism for "disambiguation based on direct syntactic composition" (p. 254), arguing that 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. Erk and Padó (2008) outlines an approach that models words as sets of vectors including prototypical lexical representations capturing information about co-occurrence

[1]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)

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 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 word 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 suggest that the ephemeral nature of concept formation can be modelled in terms of *perspectives* on the conceptual affordances of lexical relationships. Figure 1.1b Illustrates this point. From one point of view, *dog* and *cat* refer to exemplars of the conceptual category PETS, while *wolf* and *lion* are typical of the category PREDATORS. From a more taxonomically aligned point of view, though, *dog* and *wolf* group naturally in the CANINE category, while *cat* and *lion* clearly both belong to the category of FELINES.

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 1.1: 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 variously 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, **?** 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 (1972) says that "message material, or information, comes out of a context into a context," (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.

In this 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.

1.3 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, Wong, and Yang (1975), recent research has typically involved matrix factorisation techniques presented as either (or both) an optimisation technique (Bullinaria and Levy, 2012) or a noise reducing mechanism (Kiela and Clark, 2014).[2] 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, Ducharme, Vincent, and Jauvin, 2003; Collobert and Weston, 2008; Kalchbrenner, Grefenstette, and Blunsom, 2014). Baroni, Dinu, and Kruszewski (2014) 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

[2]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 specific exceptions.

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 system: 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 relatively 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 words 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 1.2, 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 1.3).

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

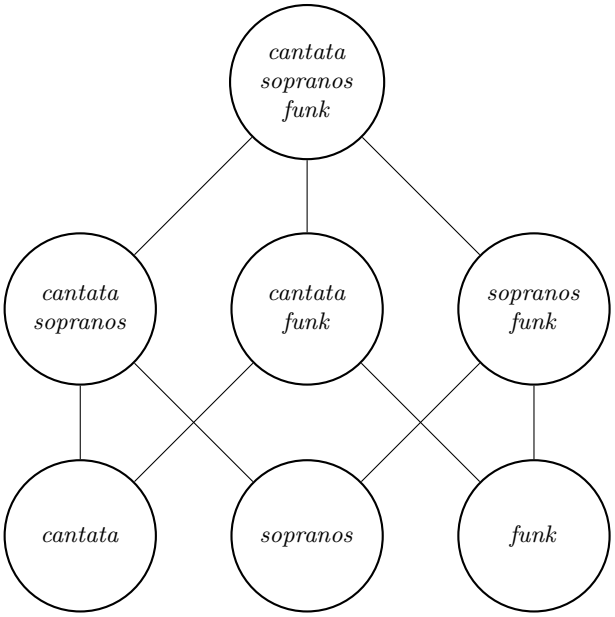


Figure 1.2: A lattice of three dimensions, including the two-dimensional subspaces which are used for analysing the conceptual geometry of a small set of word-vectors in Figure 1.3

proposed 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 referent 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. 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.

1.4 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 mechanism for selecting subspaces described above: this mechanism 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 section. 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, Dumais, Furnas, Landauer, and Harshman, 1990). Thus the linear algebraic methodology of generating a lower dimensional matrix of optimally informative dimensions arguably transforms a space of specific co-occurrence events into a space of more general conceptual relationships. In fact, Landauer, Laham, Rehder, and Schreiner (1997) 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," (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 (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.

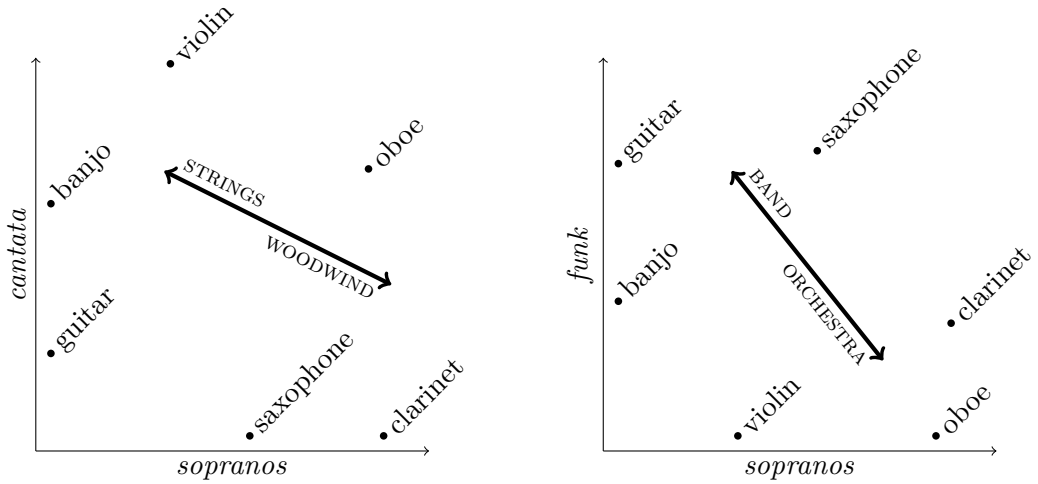
A classic example of a Gärdenforsian conceptual space is the space of colours, which can be defined in terms of, for instance, hue, brightness, lightness, and colourfulness:

any colour 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 color names into convex regions.[3]

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 word *bog* is found to be more or less between *heath* and *wetland*, for instance, and the vector for the film *Jurassic Park* lies in a direction 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

[3]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 1.3: Based on real co-occurrence data, swapping one dimension in a two-dimensional subspace reveals two different conceptual geometries.

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 1.3 (taken from real co-occurrence data, as described in the next section, and based on the lattice of subspaces illustrated in Figure 1.2). 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 cluster 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 clusterings 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 1.3a 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 section, 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 1.3 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 2

A Computational Implementation of 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, all to be used as a point of comparison in the empirical results which will be discussed in subsequent chapters.

2.1 Establishing and Analysing a Large Scale Textual Corpus

The first step in a corpus based approach to natural language processing is the selection of the data which will provide the basis for our model. I've picked the English language portion of Wikipedia as my data source, a choice which is in accordance with a good deal of work done in the field. For instance, 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. (2014), Pennington, Socher, and Manning (2014), and Gutiérrez, Shutova, Marghetis, and Bergen (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 periphrasitc. This should supply a source of linguistic data in which, to revisit the central dogma of the distributional hypothesis, words which occur in a specific 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 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. The cleaned corpus contains nearly 1.1 billion word tokens, consisting of almost 7.5 million unique word types. The distribution of these types is predictably Zipfian: over 10 million occurrences of the top nine word types are observed, while the least frequent 4.27 million words – more than half of all types – only occur once. The top end of this distribution is populated by conjunctions, prepositions, and pronouns, while the bottom end is characterised by obscure place names, one-off abbreviations, unicode representing non-Latin alphabet spellings, and a good many spelling errors.

As is generally the case with data cleaning, these measures are prone to error: for instance, due to the removal of punctuation, the contraction *we're* will be considered identical to the word *were*. One of the strengths of the subspace projection technique that my methodology uses is its resilience to noise. So, for instance, misspellings will be categorised as highly anomalous co-occurrence dimensions and are therefore unlikely to be contextually selected – or, if they are regularly encountered enough to be contextually significant, there may well be useful information in the co-occurrence profile of such mistakes – and, 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 – that is, occurrences of any word of any type

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

– by selecting all instances of all vocabulary words equals the total number of word types – that is, unique word forms – 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 based 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 of w, 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:

$$M_{w,c} = \log_2 \left(\frac{f_{w,c} \times W}{f_w \times (f_c + a)} + 1 \right) \tag{2.1}$$

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 2.1 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 is observed less often than the balance of the mean values for all occurrences of w and all co-occurrences with c – as being equivalent to zero, or no co-occurrence at all (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 areas for variation in deriving co-occurrence statistics include the nature of the co-occurrence window itself, where some researchers have taken weighted samples (**?**)r considered word order, and also the actual representation of tokens within

[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 the combination 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.

the corpus, where part-of-speech and dependency tagging (Padó and Lapata, 2007) have been applied to positive effect. Lapesa and Evert (2014) and Milajevs, Sadrzadeh, and Purver (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 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 by the same token 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.

2.2 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—and, in the case of language, the action afforded is meaning making. In practice, however, for the purposes of my methodology, context will be defined lexically, as 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 the idea that there should be a set of words which collectively selects a set of dimensions that are conceptually relevant to some conceptual context, and the geometry of the word-vectors of my model vocabulary as projected into the subspace delineated by this set of dimensions should reveal the semantics of this context.

So, notwithstanding interesting work on multi-modal approaches to distributional semantics from, for instance, Hill and Korhonen (2014) and Bruni, Tran, and Baroni (2014), with regard to the present technical description, I will treat *context* 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

		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 2-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.

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 2-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 respective co-occurrence profiles more in line with one another, facilitating the selection of a subspace which is jointly characteristic of the input terms.

In the cases of the INDY and ZIPPED techniques, the selectional process is more straightforward, since mean values between 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. 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. 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 dimensions selection techniques, a preliminary and intuitively motivated case study of dimension selection is outlined in Table 2-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 animals 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 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

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 2-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.

Cowardly Lion character from *The Wizard of Oz*) and *crouching* (indicating 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.

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 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 numbers 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 theoretically 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.

2.3 Exploring the Geometry of a Context Specific Subspace

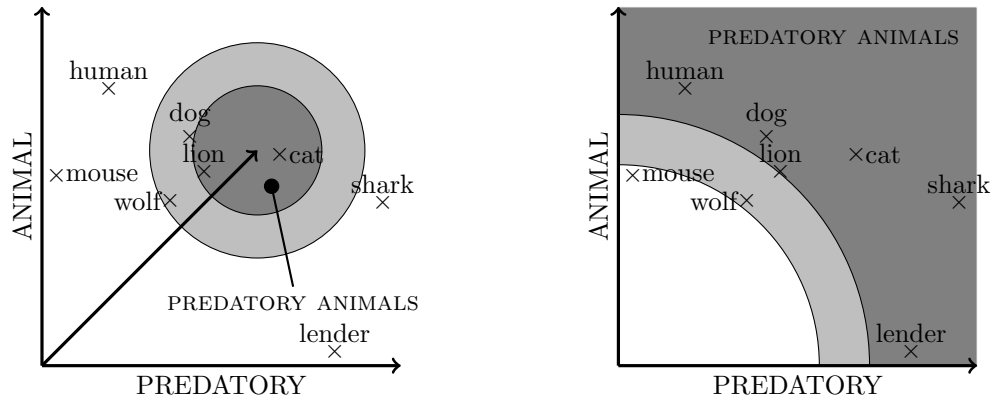
Before delving into the question of the types of geometries my method might be expected to generate, I would like to raise a point regarding the typical application of the term *geometry* to vector space models of distributional semantics in the first place. Widdows (2004) makes an enthusiastic and compelling case for the representational power of geometry, while Clark (2015) has pointed out that treating words as geometric features endows lexical representations with "significant internal structure" (p. 509) which can be applied towards modelling the meaning making compositionality of language. Baroni et al. (2014) 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.

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 eigenvectors are used as dimensions of maximal variance, there is no meaningful interpretation of sign 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 the 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 pro-

jected onto the selected number of eigenvectors, and cosine similarity is in practice the measure used to determine the similarity between two word-vectors. And in the case of models built using neural networks, there is no meaningful interpretation of dimensions to begin with, so the resulting space is a *de facto* hypersphere of word vectors that are only relative in terms of their relationship to one another, not their relationship to any objective features of the space.

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

The second effect is that there is a notion of centre and periphery in my subspaces. Since all values are positive, a word-vector with high scores across all or most dimensions in a subspace will be far from the origin and in the central region of the space. A further consequence of the positivity of these subspaces is that word-vectors with mainly low or null PMI values will be far from the centre, so in the end two word-vectors may be both close to the centre of a subspace, or at the periphery of a subspace but close to one another, or at the periphery and far from each other, at two different edges of the positively valued space, and each of these situations can be predicted to have a particular semantic interpretation. The third effect, which follows from the first two points, is that a subspace can be characterised in terms of a set of key points based on an analysis of the collective profiles of the dimensions delineating the subspace, by which I mean some



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

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

Figure 2.1: 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.

straightforward assessments of the statistical distribution of each dimension involved.

2.3.1 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 2.1. 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. In practice, the central point will be defined as the mean point between the input word-vectors used to generate the subspace, but for the purposes of Figure 2.1a, a point along the line extending from the origin through the centre of the two dimensional space has been chosen. 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,

[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 2.1, conceptual

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 2.1b, 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 2.1a: here a clear sense of convexity as a criterion for a conceptual region exists, and likewise of

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 2-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, word-vectors closest to the mean point between the input word-vectors, and also the smallest angle with this mean vector regardless of actual position in the subspace.

betweeness as an indicator of conceptual inclusion. Importantly, though, these spaces as they stand lack the dimensional interpretability that characterises Gärdenfors's spaces, in that it is not possible to say that there is a dimension of size, or strength, or ferocity, or so forth along which a boundary for inclusion in the concept of PREDATORY ANIMAL can be identified.

Examples of the tendencies of both norms and relative distances are explored in Table 2-C and Table 2-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

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 2-D: The top word-vectors in subspace, as in Table 2-C but selected by input terms characteristic of PETS.

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

Notably, in the case of the norm measure, word-vectors that are exemplary of the conceptual category suggested by the intersection of the input terms seem to rise to the top of the subspace, so to speak: for both dimension selection techniques for the WILD ANIMAL type inputs, a list of wild animals, some rather exotic, are returned. A similar outcome is observed for the norm measure in the case of the pet inputs, with some admittedly disputable admissions such as *rats* coming up in the JOINT output; jerboas, which are indicated in the INDY output, are apparently a somewhat popular pet, and *akc* presumably refers to the American Kennel Club, so, not a pet, but an institution related to pet keeping. An interesting side effect of the INDY technique in particular is that it returns 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 space 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 animals* 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 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.

2.3.2 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. 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. The presumption is that these word vectors are to be analysed in terms of their semantic relationship; the precise nature of the relationship being analysed could be more or less anything, and in the next two chapters this method will be applied to the assessment of lexical similarity, relatedness, metaphor, and metonymy. The objective of this analytic method will be first to test the hypothesis that the geometry of contextually projected subspaces should be semantically informative, and second to compare the aspects of the geometry that are most informative for different semantic phenomena.

Figure 2.2 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

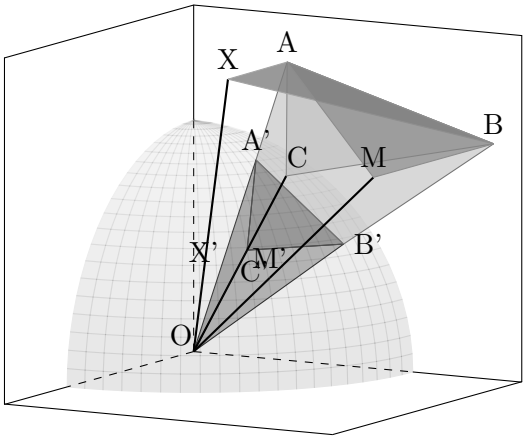


Figure 2.2: The geometric features of a subspace contextually projected based on an analysis of two input word-vectors.

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:

$$M = \{\mu(J_1),\ \mu(J_2)...\ \mu(J_d)\} \tag{2.2}$$

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

$$X = \{max(J_1),\ max(J_2)...\ max(J_d)\} \tag{2.3}$$

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

	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 2-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.

$$C = \{\mu(M),\ \mu(M)...\mu(M)\} \tag{2.4}$$

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 point here is that, due to the frequentist components of the information theoretic co-occurrence statistics used to build the base space described here, different dimensions have different distributional profiles. To demonstrate this point, Table 2-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

[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.

clearly quite frequent words, and this is to be expected, given that the high frequency of independent observations of the word will drive PMI scores down for that word across the board. The emergence of relatively infrequent words at the top end of the spectrum is then also to be expected. 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. 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 2.2. 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 angle $\angle AOB$ might still be very small—and, even then, the angle $\angle A'C'B'$ 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 2-F. The points to be found in the space are broken down into three types, names, 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 (for instance, A'). The relationships between these points are construed across five categories as follows:

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

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

Means The average 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 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 lower frequency dimensions are jointly selected by a pair of words, a (more correction) expectation will be that this indicates a high degree of conceptual overlap between the words'

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

referrents, and therefore a high degree os 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 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 2-F is in this regard a schematic for the computational experimentation and corresponding evaluation and analysis at the core of this thesis.

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 points of comparison for my methodology.

2.4 Comparing to Alternative Approaches

In order to evaluate the effectiveness of my methodology, it will naturally be necessary to compare the preformance 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 empirical 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 same data towards which I apply my methdology. 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, Reichart, and Rappoport (2015) have developed a syntactic heuristic for identifying semantic similarity as compared to relatedness in particular, and Gutiérrez et al. (2016) develop a model that generates compositional adjective-noun representations geared towards metaphor detection. One of the key features of my models is that they are intended to be *general*: the geometries generated by my methodology are expected to be replete with semantic interpretability, allowing for the same potential for diverse and often surprising conceptualisation corresponding to the infinitely combinatory characteristic of natural language in use. For this reason, it is desirable to have a base case of a generic model that can be compared across the board to all the different tasks handled by my methodology.

With all this in mind, I propose two different points of comparison that, in addition to

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

2.4.1 Static Interpretations of the Base Space

In cases where the geometry being explored involves just target word-vectors and generic points of a space – so, for all the features described in Table 2-F – it is computationally tractable to treat the sparse base matrix from which subspaces are projected as a semantically interpretable space in its own right. This is because universal generic points (the mean point for all dimensions, the maximum point for all dimensions, and the central point based on the average value of the mean point) can be discovered through a one-off caluculation, and the word-vectors themselves will be relatively sparse. In the most honerous case of comparing two orthogonal word-vectors using the INDY technique, the total number of scalars involved in the computations of geometric features would be the sum of the number of non-zero dimensions for each word-vector, so something on the order of thousands to tens of thousands of values—not that bad, computationally speaking.

Of course, the norms of the generic points in such a general space will be extremely high compared to any given actual word-vector, since these generic points will have non-zero values in several million dimensions. With this in mind, a second and more typical approach to building a general and computable distributional semantic space out of my base space of co-occurrence statistics is to through matrix factorisation: using singular value decomposition, I project the base space onto the top most informational

eigenvectors up to a dimensionality to match the parameters tested using my context specific dimensional selection techniques.[5] Because of computability constraints, I take the top 100 to 50,000 most frequent word types as the vocabulary for this model, and consider the top 10,000 most frequent co-occurrence terms as the dimensions of the matrix to be factorised. This means that 1,979 of the 1,998 word tokens in the SimLex999 dataset (Hill and Korhonen, 2014, discussed in detail in Chapter **??**) are included in the vocabulary, and almsot 90% of the co-occurrence observations tabulated in the base space are represented in the decomposed model.

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

$$d'_i = |d_i - \underset{d \in D}{\operatorname{argmax}}(\mu\{d_w : w \in W\} - d)| \tag{2.5}$$

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

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

2.4.2 A Model Trained Using a Neural Network

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

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

In addition to the size of the co-occurrence window, model parameters include the

number of iterations of the corpus, the architecture of the single-layer network connecting input to output vectors, and, in the case of the skip-gram model, a rate of negative sampling by which random sets of words are taken as instances of non-co-occurrences and used to push the corresponding word-vectors away from the input word-vector. The skip-gram model, with its sensitivity to word order, has been reported to perform particularly well on analogy completion task involving semantic similarity, so for instance in discovering the relationship *king:queen::man:woman*. The CBOW model, on the other hand, has performed better on what the authors have described as *syntactic*[6] analogies such as *good:better::bad:worse*.

A point of note regarding the `word2vec` models is that word similarity in their spaces is measured in terms of cosine similarity, which means these are treated as effectively normalised vector spaces in which the Euclidean distances between points are

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.

[6]Following **?** and, in the context of computational linguistics, **?**, I would prefer the distinction *paradigmatic* versus *syntagmatic* in place of syntax versus semantics, though it should be noted that in a very clear sense all distributional semantic models are primarily concerned with both semantics and paradigms, using syntax and syntagm as the mechanism for discovering these semantic relationships.

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.

References

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

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

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

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

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

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

Birkhoff, G. (1958, 05). Von neumann and lattice theory. *Bulletin of the American Mathematical Society 64*, 50–56.

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

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

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

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

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

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

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

Collobert, R. and J. Weston (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 160–167.

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

Derrac, J. and S. Schockaert (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* (2nd ed.). London: Duckworth.

Erk, K. and S. Padó (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, pp. 897–906.

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

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

Gabrilovich, E. and S. Markovitch (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International*

Joint Conference on Artifical Intelligence, pp. 1606–1611.

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

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

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

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

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

Hill, F. and A. Korhonen (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*, pp. 255–265.

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

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

Kartsaklis, D. and M. Sadrzadeh (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*, pp. 2849–2860.

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

Kiela, D. and S. Clark (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*, Gothenburg, pp. 21–30.

Landauer, T., D. Laham, B. Rehder, and M. E. Schreiner (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*, pp. 412–417.

Lapesa, G. and S. Evert (2013, August). 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)*, Sofia, Bulgaria, pp. 66–74. Association for Computational Linguistics.

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

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

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

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

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for

word representation. In *Conference on Empirical Methods in Natural Language Processing*.

Polajnar, T. and S. Clark (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*, pp. 230–238.

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.

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

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

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

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

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

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*, pp. 136–143.

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

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 I. Rival (Ed.), *Ordered Sets*, Dordrecht/Boston, pp. 445–470. Reidel.

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