

**Stage Two Report:**  
**A Statistical Language Model for the**  
**Generation of Figurative Language**

by  
Stephen McGregor

An outline of 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

September 2015

# Abstract

MIMO (Multiple Input Multiple Output) technology has been regarded as a practical approach to increase the wireless channel capacity and reliability.

Abstract goes here...

# Table of Contents

<b>Abstract</b>	<b>i</b>
<b>Table of Contents</b>	<b>1</b>
<b>1 Semantics on the Fly</b>	<b>2</b>
1.1 Dynamic Context Sensitivity . . . . .	6
1.2 Literal Dimensions of Co-Occurrence . . . . .	8
1.3 Interpretable Geometry . . . . .	13
References . . . . .	19

# 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 ? 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 which 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 ??, 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 ??, 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 three 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.

## 1.1 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 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.<sup>1</sup>

With that said, the importance of context has certainly not been ignored by statistically oriented computer scientists. Indeed, Baroni et al. (2014a) 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 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

---

<sup>1</sup>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 (Neumann, 1945)



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 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

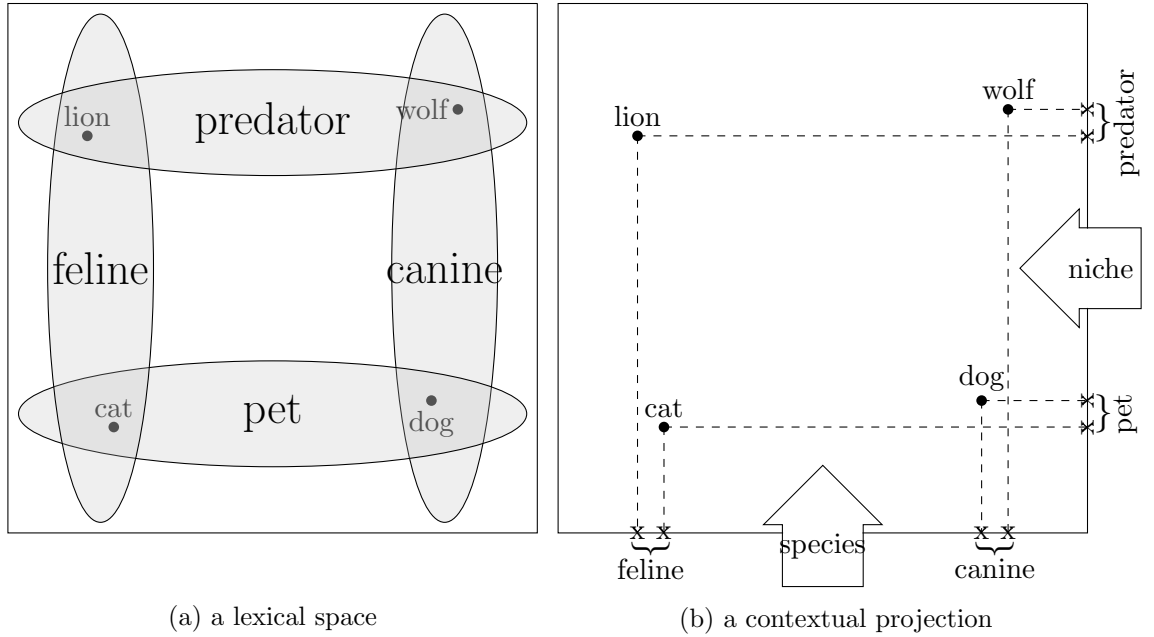


Figure 1.1: In the two-dimensional space depicted in (a), the conceptual vagary of four words maps to overlapping, elongated and indeterminate spaces. In (b), two different perspectives on the lexical space, represented by the arrows labelled *niche* and *species*, offer contextualised projections in one-dimensional clusters which remit conceptual clarity.

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.

This description of a process of perspective taking has thus far been quite high-level: that high dimensional spaces afford different vantage points on data is evident, but this begs the question of the actual mechanism for defining perspectives in such spaces. The next section will focus on answering this question by outlining a method for

## 1.2 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 obser-

variations of word co-occurrences in a large-scale corpus of textual data represented statistically. Two procedurally distinct methodological regimens have emerged from the recent study of distributional semantics. The first, and more established, approach involves tabulating word co-occurrence frequencies and then using some function over these to build up word-vector representations. With roots in the frequentist analysis described by Salton et al. (1975), recent research has typically involved matrix factorisation techniques presented as either (or both) an optimisation technique (Bullinaria and Levy, 2012) or a noise reducing mechanism (Kiela and Clark, 2014).<sup>2</sup> 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 non-linear regression techniques to iteratively learn word-vector representations in an online, stepwise traversal of a corpus (Bengio et al., 2003; Collobert and Weston, 2008; Kalchbrenner et al., 2014). Baroni et al. (2014b) have described the former as *counting* and the latter as *predicting*, but it must be noted that both methods are very much grounded in observations about the co-occurrence characteristics of vocabulary words across large bodies of text.

Another important similarity between these two approaches is that they each in their own way move towards a representation of relationships between word-vectors which is to some extent optimally informative, and, by the same token, abstract. In the instance of neural network approaches, this is clearly the case due to the fundamental nature of the 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, a mutual information type metric) and then factorising

---

<sup>2</sup>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.

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 networks approaches just exactly do 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.

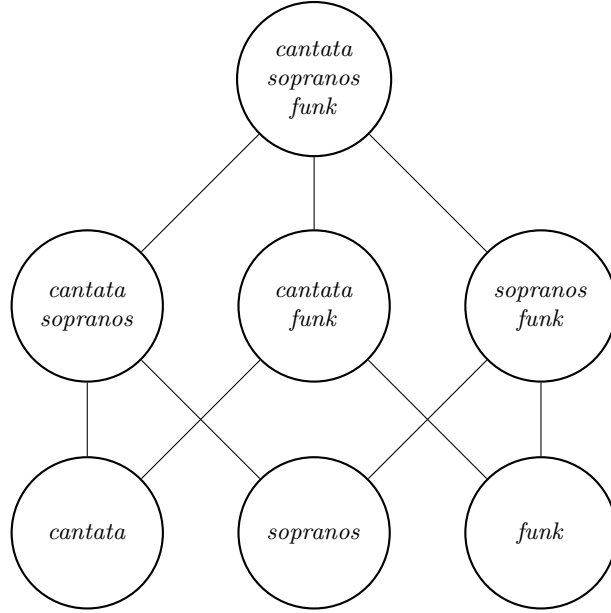


Figure 1.2: A lattice of three dimensions, including the two-dimensional subspaces which are used for analysing the conceptual geometry of a small set of word-vectors in Figure 1.3

At the top of this lattice – the *join* – sits a single  $k$ -dimensional space consisting of every available one of the  $k$  co-occurrence terms observed throughout the underlying corpus. At the bottom of the lattice – the *meet* – sit  $k$  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* –  $k$ , 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 proposed statistical techniques using *feature inclusion* metrics to assess the potential entailment relationship 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 model 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 for 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 methodological overview will examine how the actual geometry of a projected subspace itself is expected to do this conceptual work.

### 1.3 Interpretable Geometry

It is important at this point to distinguish between two different modes of interpretability at play within the operation of the model 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 (LSA) was motivated by precisely the insight that a singular value decomposition of a high-dimensional, sparse matrix of statistical data about word co-occurrences would result in a dense lower dimensional matrix in which dimensions characterise *latent semantics* rather than literal word co-occurrences (Deerwester et al., 1990). Thus the linear algebraic methodology of generating a lower dimensional matrix of optimally informative dimensions arguably transforms a space of specific co-occurrence events into a space of more general conceptual relationships. In fact, Landauer et al. (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 betweenness 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 psycholinguist 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 colors, 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*. ? 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, 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.<sup>3</sup>

---

<sup>3</sup>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



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 *betweenness* 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 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.

---

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.

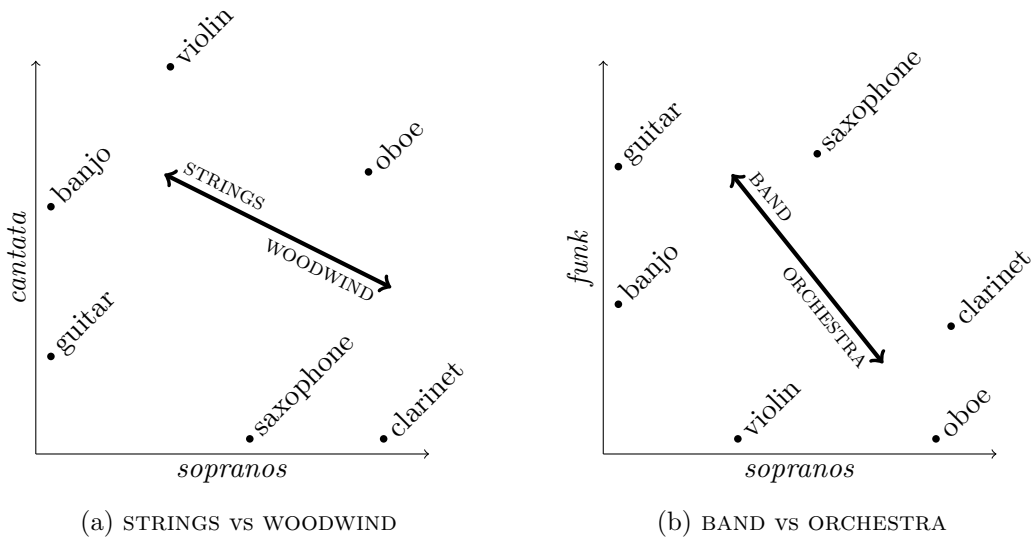


Figure 1.3: Based on real co-occurrence data, swapping one dimension in a two-dimensional subspace reveals two different conceptual geometries.

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 that we 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 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 combinatorial 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. The next section will answer this question.

## References

- Baroni, M., Bernardi, R., and Zamparelli, R. (2014a). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9:241–346.
- Baroni, M., Dinu, G., and Kruszewski, G. (2014b). Don’t count, predict! In *ACL 2014*.
- Barsalou, L. W. (1993). *Theories of Memory*, chapter Flexibility, Structure, and Linguistic Vagary in Concepts: Manifestations of a Compositional System of Perceptual Symbols. Lawrence Erlbaum Associates, Hove.
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.
- Birkhoff, G. (1958). Von neumann and lattice theory. *Bulletin of the American Mathematical Society*, 64:50–56.
- Bullinaria, J. A. and Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior Research Methods*, 44(3):890–907.
- Carston, R. (2010). Metaphor: Ad hoc concepts, literal meaning and mental images. In *Proceedings of the Aristotelian Society*, volume CX, pages 297–323.
- Casasanto, D. and Lupyan, G. (2015). All concepts are ad hoc concepts. In Margolis, E. and Laurence, S., editors, *The Conceptual Mind: New Directions in the Study of Concepts*. MIT Press, Cambridge, MA.
- Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origins, and Use*. Praeger, New York, NY.
- Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal for the American Society for Information Science*, 41(6):391–407.

- Derrac, J. and Schockaert, S. (2015). Inducing semantic relations from conceptual spaces: A data-driven approach to plausible reasoning. *Artificial Intelligence*, 228:66–94.
- Erk, K. and Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08*, pages 897–906.
- Erk, K. and Padó, S. (2010). Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97.
- Evans, V. (2009). *How Words Mean: Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford University Press.
- Gärdenfors, P. (2000). *Conceptual Space: The Geometry of Thought*. The MIT Press, Cambridge, MA.
- Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. The MIT Press.
- Geffet, M. and Dagan, I. (2005). The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 107–114.
- Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics Volume 3: Speech Acts*, pages 41–58. Academic Press, New York.
- Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.
- Kartsaklis, D. and Sadrzadeh, M. (2016). Distributional inclusion hypothesis for tensor-based composition. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2849–2860.
- Kay, P. and Maffi, L. (1999). Color appearances and the emergence and evolution of basic color lexicons. *American Anthropologist*, 101(4):743–760.
- Kiela, D. and Clark, S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models*

- and their Compositionality (CVSC) @ EACL 2014, pages 21–30, Gothenburg.
- Landauer, T., Laham, D., Rehder, B., and Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pages 412–417.
- Lapesa, G. and Evert, S. (2013). Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 66–74, Sofia, Bulgaria. Association for Computational Linguistics.
- Levinson, S. C. (2001). Yéli dnye and the theory of basic color terms. *Journal of Linguistic Anthropology*, 10(1):3–55.
- Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.
- Montague, R. (1974). English as a formal language. In Thompson, R. H., editor, *Formal Philosophy: selected papers of Richard Montague*. Yale University Press, New Haven, CT.
- Neumann, J. v. (1945). First draft of a report on the edvac. Technical report.
- 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., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. In *Proceedings of the 12th ACM SIGIR Conference*, pages 137–150.
- Widdows, D. (2003). Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 136–143.
- Widdows, D. (2004). *Geometry and Meaning*. CSLI Publications, Stanford, CA.
- Wilks, Y. (1978). Making preferences more active. *Artificial Intelligence*, 11(3):197–

223.

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

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