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

Chapter 1

A Model for Constructing Concepts on the Fly

This chapter is concerned with a theoretical overview and a technical description of a novel distributional semantic model designed to map words into conceptually productive geometric relationships. Theoretically speaking, the model 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 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. It 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 using 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 model should be dynamically sensitive to context;

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

3. The model should be situate words in spaces which are likewise geometrically interpretable.

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, and each element of the profile of desirable properties will be explored with this in mind. Then finally, in a fourth section, the fundamental implementation of the model will be described in technical detail.

1.1 Dynamic Context Sensitivity

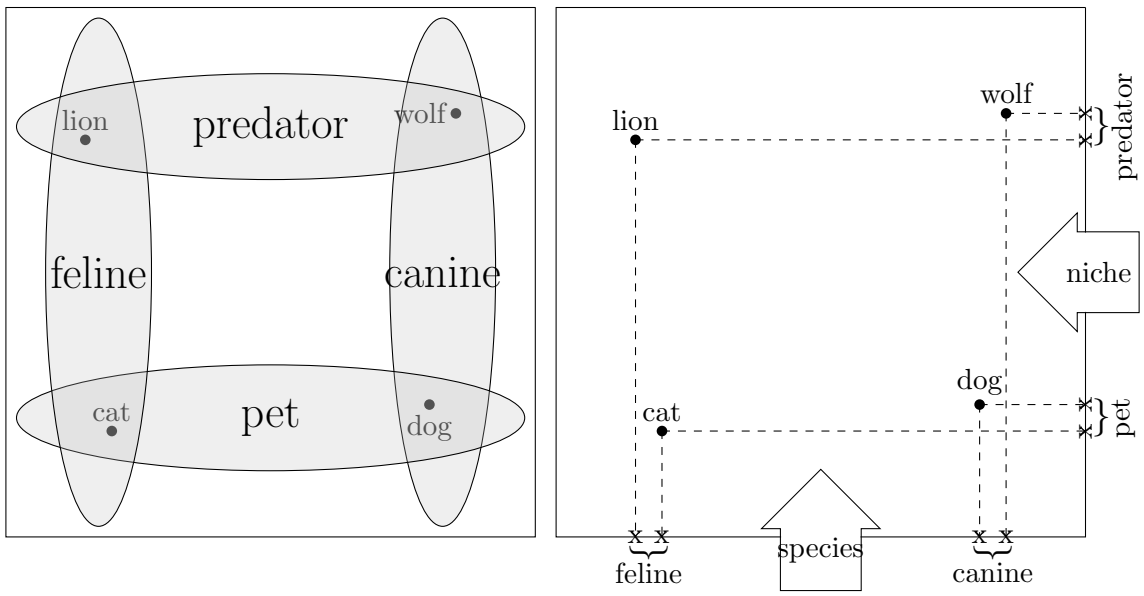
At the heart of the technical work described in this thesis is an insight which is broadly accepted by theoretical linguists and philosophers of language: word meaning is always contextually specified. This wisdom is built into the foundations of both formal semantics (Montague, 1974) and pragmatics (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.[1]

With that said, the importance of context has certainly not been ignored by statistically oriented computer scientists. Indeed, Baroni, Bernardi, and Zamparelli (2014) make a case for vector space approaches to "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

[1]It is perhaps not a coincidence that **?** was a seminal figure in the description of both the logic of lattice theory that has motivated more recent developments in concept modelling such as formal concept analysis (**?**) and the modular architecture of memory and processing components that defined computers in the period before the advent of highly parallel processing.

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* (*a la* 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 (Erk & Padó, 2010), 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. 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 a even further, though. Following on **?**'s (**?**) insight into the *haphazard* way in which concepts emerge situationally, and likewise **?**'s (**?**) proposition 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 linguistic 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



(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.

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 vistas 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 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

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 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 & Levy, 2012) or a noise reducing mechanism (Kiela & 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 non-linear regression techniques to iteratively learn word-vector representations in an online, stepwise traversal of a corpus (Bengio, Ducharme, Vincent, & Jauvin, 2003; Collobert & Weston, 2008; Kalchbrenner, Grefenstette, & 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 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

[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.

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 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 model proposed in this thesis is that it maintains a 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. I don't wish to claim that there is scope for completely or even mainly recapitulating a nuanced conceptual model from the data available in a purely textual environment; to do so would be to move towards claims that intentionality can emerge from rule-based operations on symbols, and the problems with this have been explored by Searle's (1980) Chinese room argument and a subsequent generation of philosophers (Preston & Bishop, 2002). But I would like to suggest that by building a base model that maintains the accessibility of unreduced co-occurrence information, we likewise maintain the ability to manipulate this base model extemporaneously, in reaction to the ongoing emergence of new contextual information. The idea is that such a base model would

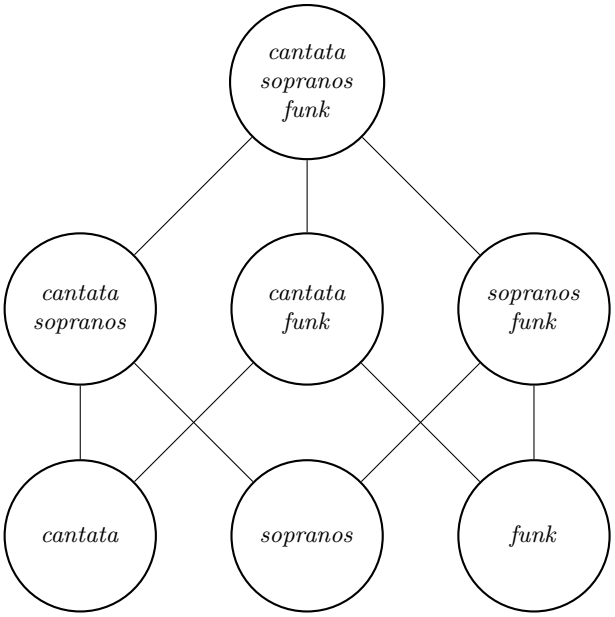


Figure 1.2: A lattice of three dimensions, including the two-dimensional subspaces which are used for analysing the conceptual geometry of a small set of word-vectors in Figure 1.3

essentially represent the superset of all possible dimensions available for *ad hoc* selection in the course of a

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 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, as 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 extent. 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 model 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 idea that

concepts are atomic or elementary to thought, or that language and concepts are involved in any kind of strictly hierarchical relationships. Instead, the model operates through an analytical traversal of a lattice of subspaces in search for a combination of dimensions that captures a conceptually *salient* profile of co-occurrence features. If a consequence of this stance is that the model can't be understood in terms of nested ordered relationships, though, then the question of how conceptual relationships do emerge situationally from the model 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, Dumais, Furnas, Landauer, & 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 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.

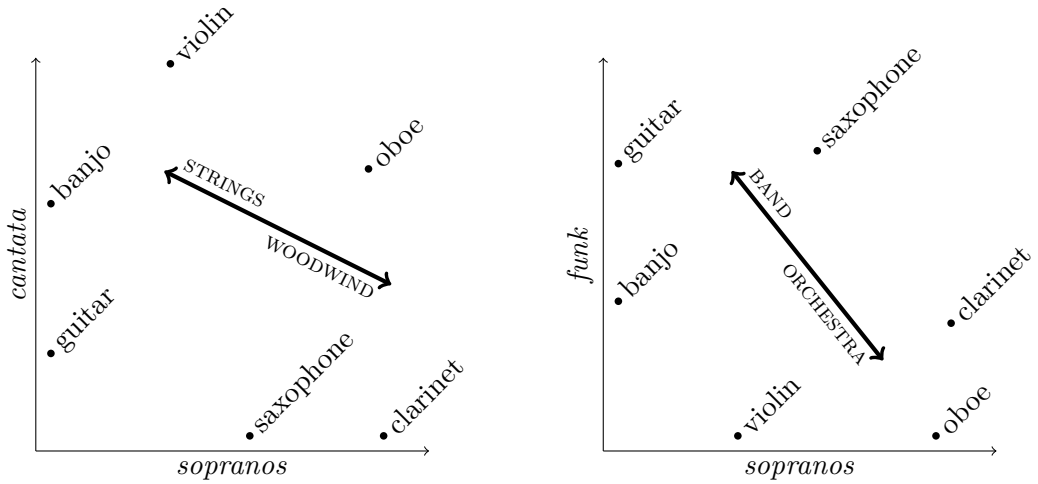
A classic example of a Gärdenforsian conceptual space is the space of colors, which can be defined in terms of, for instance, hue, brightness, lightness, and colourfulness: any colour can be specified as a point corresponding to coordinates along each of these dimensions. Moreover, regions within the 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.[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 **?** 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 Fig-

[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.

ure 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 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 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. The next section will answer this question.

1.4 A Computational Process

In this final section of this chapter, a technical implementation of the model described throughout the preceding three sections will be explained in detail.

1.4.1 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. Some authors

In the case of the model used throughout this thesis, the November 2014 dump of English language Wikipedia has been used.[4] 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 from being considered as unique types. Finally, the arti-

[4] Accessible at XXX

cles *a*, *an*, and *the* are removed as they can distort co-occurrence windows (consider, for instance, how these terms affect the proximity of the other words in the phrase "a mouse, an owl, and a dog sat on the moon"). The cleaned corpus contains about

-WORD, SENTENCE COUNTS

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 model 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 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, the 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 – included 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.

- human vocabulary size

1.4.2 Mutual Information of Word Co-Occurrences

The critical event in the

Here, following the example of almost all 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 model, 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, including the possible co-occurrence of a word with itself ("a *rose* is a *rose* is a *rose*", for instance) 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 dimension of co-occurrence.

In this last respect, my model diverges from the typical approach, which usually seeks to limit not only the vocabulary but also the dimensionality of the underlying co-occurrence matrix. This has typically involved a curtailing of the number of co-occurrence terms at both ends of the frequency spectrum, based on the assumption that both high frequency so-called function words (the prepositions, conjunctions, and so forth) and low frequency terms such as obscure proper names will muddy a model with either general flattening or highly topical skewing. In the case of my model, however, these problems are irrelevant, as dimensions will be selected on a case-by-case, context specific basis, and there is no good reason to discard information which may in some possibly unforeseen circumstance prove relevant. The result is a 200,00 by $\approx$ $7.5 million matrix \mathrm{M} where as calar corresponding to co-occurrences between w and c is defined in terms of t$

$$M_{w,c} = \log_2 \left(\frac{n_{w,c} \times W}{n_w \times (n_c + a)} + 1 \right) \tag{1.1}$$

Here $n_{w,c}$ represents the total number of times that that c is observed as co-occurring in a sentence within k words on either side of w, n_w is the independent frequency of occurrences of w, and 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 1.1 is equivalent to the joint probability of w and c co-occurring. The application of a logarithm to this ratio, again a common practice, is in the spirit of **?**'s (**?**) information theory, and is

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.[5]

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 is again a departure from standard practice, where, in word counting models, a *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 than often than balance of the mean values for all occurrences of w and all co-occurrences with c – as being equivalent to 0, or no co-occurrence at all. 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

[5] 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*—my primary supervisor's email handle. 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.

such information, and at the same time, as will be seen, data regarding relatively unlikely co-occurrences can sometimes also be quite informative. In support of my technique, it is worth mentioning that the vast majority of potential co-occurrences will never be observed, and, at the same time, a comprehensive language model should maintain at least the possibility of any co-occurrence

?

so there seems to be wisdom in the idea of not throwing away information about even relatively unlikely linguistic events.

1.4.3 Dimensional Selection Techniques

Having established a base model of co-occurrence statistics, the

$$w_i^j = \frac{w_i^j}{\sqrt{\sum_{k=1}^{b} \left(w_i^k\right)^2}} \tag{1.2}$$

$$w_i^j = \frac{w_i^j}{\sum_{k=1}^{b} abs\left(w_i^k\right)} \tag{1.3}$$

$$\mu_c = \frac{1}{n} \sum_{w=1}^{n} N_{w,c} \tag{1.4}$$

$$M_{w,c} \Rightarrow S_{w,c'} \tag{1.5}$$

1.4.4 Extracting Semantics from Geometric Features

References

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

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

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

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior Research Methods*, *44*(3), 890–907. Retrieved from `http://dx.doi.org/10.3758/s13428-011-0183-8` doi: 10.3758/s13428-011-0183-8

Chomsky, N. (1986). *Knowledge of language: Its nature, origins, and use.* New York, NY: Praeger.

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

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

Derrac, J., & Schockaert, S. (2015). Inducing semantic relations from conceptual spaces: A data-driven approach to plausible reasoning. *Artificial Intelligence*, *228*, 66–94. Retrieved from `http://dx.doi.org/10.1016/j.artint.2015.07.002` doi: 10.1016/j.artint.2015.07.002

Erk, K., & 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* (pp. 897–906). Retrieved from `http://dl.acm.org/citation.cfm?id=1613715.1613831`

Erk, K., & Padó, S. (2010). Exemplar-based models for word meaning in context. In *Proceedings of the acl 2010 conference short papers* (pp. 92–97). Retrieved from

http://dl.acm.org/citation.cfm?id=1858842.1858859

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., & Dagan, I. (2005). The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 107–114). Retrieved from https://doi.org/10.3115/1219840.1219854 doi: 10.3115/1219840.1219854

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

Kalchbrenner, N., Grefenstette, E., & 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., & 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* (pp. 2849–2860). Retrieved from http://aclweb.org/anthology/C/C16/C16-1268.pdf

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

Kiela, D., & 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* (p. 21-30). Gothenburg.

Landauer, T., Laham, D., Rehder, B., & 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* (p. 412-417).

Lapesa, G., & Evert, S. (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) (pp. 66–74). Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W13-

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

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2177–2185). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5477-neural-word-embedding-as-

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.

Preston, J., & Bishop, M. (2002). *Views into the chinese room: New essays on searle and artificial intelligence.* Clarendon Press.

Rimell, L. (2014). Distributional lexical entailment by topic coherence. In *Proceedings of the 14th conference of the european chapter of the association for computational linguistics.* Gothenburg.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. In *Proceedings of the 12th acm sigir conference* (p. 137-150).

Searle, J. R. (1980). Minds, brains, and programs. In M. A. Boden (Ed.), *The philosophy of artificial intelligence* (p. 67-88). Oxford University Press.

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). Retrieved from http://dx.doi.org/10.3115/1075096.1075114 doi: 10.3115/1075096.1075114

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* (pp. 445–470). Dordrecht/Boston: Reidel.

Wille, R. (2005). Formal concept analysis as mathematical theory of concepts and concept hierarchies. In (pp. 1–33). Retrieved from http://dx.doi.org/10.1007/11528784_1