A Geometric Method for Context Sensitive Distributional Semantics

by

Stephen McGregor

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

First Supervisor: Prof. Geraint Wiggins
Second Supervisor: Dr. Matthew Purver

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

September 2017

I, Stephen McGregor, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third partyâĂŹs copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Date:

My university has required me to make the above statement. To this I add the following:

I hereby grant permission to anyone to do anything they so please with the text of this thesis and any information they derive from it or meaning they find in it.

Details of collaboration and publications:

McGregor, S., Jezek, E., Purver, M., Wiggins, G.: A Geometric Method for Detecting Semantic Coercion. 12th International Workshop on Computational Semantics. Montpellier (2017).

2016 McGregor, S., Purver, M., Wiggins, G.: Words, Concepts, and the Geometry of Analogy. Proceedings of the Workshop on Semantic Spaces at the Intersection of NLP, Physics and Cognitive Science. Glasgow (2016).

McGregor, S., Purver, M., Wiggins, G.: Process Based Evaluation of Computer Generated Poetry. Proceedings of the INLG Workshop on Computational Creativity in Natural Language Generation. Edinburgh (2016).

Agres, K., McGregor, S., Rataj, K., Purver, M., Wiggins, G.: Modeling Metaphor Perception with Distributional Semantics Vector Space Models. Proceedings of C3GI at ESSLLI 2016.

McGregor, S., Agres, K., Purver, M., Wiggins, G.: From Distributional Semantics to Conceptual Spaces: A Novel Computational Method for concept creation. Journal of Artificial General Intelligence 6(1) (2015).

Agres, K., McGregor, S., Purver, M., Wiggins, G.: Conceptualising Creativity: From Distributional Semantics to Conceptual Spaces. Proceedings of the 6th International Conference on Computational Creativity. Park City, UT (2015).

McGregor, S., Purver, M., Wiggins, G.: Metaphor, Meaning, Computers, and Consciousness. Proceedings of the 8th AISB Symposium on Philosophy and Computing. Canterbury (2015).

McGregor, S., McGinty, M., Griffiths, S.: How Many Robots Does It Take? Creativity, Robots, and Multi-Agent Systems. Proceedings of the AISB 2015 Symposium on Computational Creativity. Canterbury (2015).

McGregor, S., Purver, M., Wiggins, G.: Computational Creativity: A Philosophical Approach, and an Approach to Philosophy. Proceedings of the 5th International Conference on Computational Creativity. Ljubljana (2014).

McGregor, S.: Considering the Law as an Evaluative Mechanism for Computational Creativity. Proceedings of the 50th Anniversary Convention of the AISB. London (2014).

Abstract

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

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

This method is presented as a complement and a counterpoint to established distributional methods for generating lexically productive word-vectors. Where contemporary vector space models of distributional semantics have almost universally involved either the factorisation of co-occurrence matrices or the incremental learning of abstract representations using neural networks, the approach described in this thesis preserves the connection between the individual dimensions of word-vectors and statistics pertaining to observations in a textual corpus. The hypothesis tested here is that the maintenance

of actual, interpretable information about underlying linguistic data allows for the contextual selection of non-normalised subspaces with more nuanced geometric features. In addition to presenting competitive results for various computational linguistic targets, the thesis will suggest that the transparency of its representations indicates scope for the application of this model to various real-world problems where an interpretable relationship betweendata and output is highly desirable. This, finally, demonstrates a way towards the productive application of the theory and philosophy of language to computational linguistic practice.

Glossary

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

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

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

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

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

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

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

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

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

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

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

Table of Contents

List of Figures

List of Tables

Chapter 2

Background

In this chapter, I will undertake the daunting task of outlining the scholastic background to my own research. I say this task is daunting because of the ambitious scope of my project: I intend to present a system which is both technically innovative and theoretically robust, and so I am faced with the double responsibility of providing an overview of the theory of concepts, representations, and semantics as well as a survey of ongoing work in the highly productive computational linguistic domain of distributional semantics. By achieving the right balance between theory and practice, I hope to lay the groundwork for a project that is suited for and enhanced by application to tasks developed within the field of natural language processing, but at the same time provides an empirical basis for making further theoretical commitments about the plausible operations of linguistic agents.

The theoretical background for my project in particular will lead to the development of an inventory of what Gallie (1956) has called *essentially contested concepts*, words and corresponding ideas that are more likely to invite debate and academic dissent than to offer resolution. As Deacon (2011) has put it in his biologically grounded account of the emergence of goal-directed behaviour, "Such concepts as information, function, purpose, meaning, intention, significance, consciousness, and value are intrinsically defined by their fundamental incompleteness," (ibid, p. 23). But, as Gallie points out in the context of the social sciences, these words are nonetheless important and can be useful components of a productive discourse, just so long as we are not overambitious in our claims to have arrived at some sort of conclusion about their objective definitions. Instead, I propose that the ideas of *information*, *meaning*, *creativity*, *representations*, and *concepts* should be viewed as boundary conditions for the empirical work that will be the primary focus of this thesis, delineating the conceptually territory from which my approach arises and

to which it ultimately seeks to contribute. Rather than claiming to offer any particularly visionary insight into the complex and, in general, ancient questions that foment at the perimeter of my technical work, I hope to illustrate that my research is in communication with a robust philosophical tradition and could in principle provide an empirical basis for future contributions to this discourse. Sections 2.1, 2.2, and 2.3 will deal with this.

Then, with the theoretical apparatus of my research in place, Section 2.4 will outline the technical background for the computational implementation of lexical semantic modelling that I have developed. One of my goals in this Chapter is to map out a robust correspondence between the theory of language and mind to the practice of statistic semantic model building. As will be seen both in this chapter and later in my thesis when I offer more detailed background on the experiments I use to test my methodology, there has already been appreciable thought given on the part of computational linguists to the theoretical background supporting existing models and systems described in the literature, with cognitive linguists in particular providing a useful basis for conceptual modelling, and it is not my intention to suggest that my own contribution is in some sense conceptually superior. I do, however, believe that there are some novel and valuable connections made in this manuscript, in particular with the philosophical discourse surrounding matters of representation and intentionality as well as the pragmatic approach to conceptualisation. What we will finally reach by the end of the chapter is a starting point, situated in the familiar computational linguistic domain of distributional semantics, for considering how to apply theoretical insight into the contextuality of semantics to computationally tractable lexical representations.

2.1 Meaning Making

At its heart, this thesis is about the emergence of meaning from data, and in this regard it sits atop a tradition of analytic enquiry into the nature of being itself. The very question of how meaningfulness can come about in a material universe has been arguably the unifying theme of modern Western Philosophy, spanning from the *cogito* of Descartes (1911) to the phenomenology of Husserl (2001) and Heidegger (1962), by way of empiricism (Hume, 2000; Locke, 1997), transcendental idealism (Kant, 1996), pure idealism (Hegel, 1989), and intentionality (Brentano, 1995), to delineate just one of the countless pathways through the rich tradition of ideas about minds. Broadly speaking, I intend to present a philosophically motivated, empirically oriented project that, without making controversial commitments or overambitious overtures, sits comfortably with Wittgenstein's (1967) idea that "only the act of meaning can anticipate reality," (ibid, ¶188), which I will interpret to suggest that meaning is somehow properly in the world, not

only in some immaterial, nominally mental space—but also that there really is such a thing as meaning, that it is not merely a convenient fiction of an otherwise behaviouralist ontology.

With this in mind, the project I describe here is broadly conversant with Floridi's (2011) pursuit of a *theory of strongly semantic information*, by way of which he arrives at a quantitative model of meaning.[1] The idea that observable data can be computational transformed into information is underwritten by the Information Theory of Shannon and Weaver (1949), which seeks without making any philosophical claims about knowledge or beliefs to formalise the measurement of what can be known in terms of the unexpectedness associated with sets of observations (see Pierce, 1980, for a thorough treatment). An early attempt to import technical insight from signal processing into the study of meaning can be found in Carnap and Bar-Hillel's (1952), who use Shannon-type metrics as the basis for quantifying the inferential properties associated with the semantic content of sentences, followed by Dretske (1981), who describes the formation of meaningful concepts in terms of the development of internal semantic structures that evolve to indicatively correspond with quantifiable informational situations in an environment. Subsequent forays in a *situation logic* designed to model semantic information content in a way which is simultaneously measurable and context specific (Barwise and Perry, 1983) have contributed to the resolution of computationally amendable formalisms, both in the tradition of Shannon and the semantics that have followed from Montague (1974), with the environmentally grounded approach to cognition which will be discussed presently.

At the more ambitious extent of the spectrum, the likes of Koch (2004) and Tononi (2008) have put forward theories attempting to quantify consciousness itself, generally in terms of the differentiable components of complex dynamic systems. *Consciousness*, however, is one of the aforementioned essentially contested terms, so instead of taking a stance here, I will take the easier route of simply acknowledging that there is a *hard problem* to be solved, to use the jargon of Chalmers (1996), and it should be perfectly possible to do good empirical work without necessarily taking sides in the fraught debate over the computability of the subjective experience of existence—or rather, perhaps an effective empirical approach comes about precisely from recognising the intractability of the debate in the first place. So here I will propose to use the notion of *creativity* as a kind of representative for the entire idea that being a cognitive agent has something to do with the production of meaning in reaction to the rampant stimulus provided by a dynamic and unpredictable cognitive *umwelt* (von Uexküll, 1957). In the spirit of Koestler (1964), then, and his model of creativity as "a new synthesis of previously unconnected matrices

[1] Unlike Fredkin (2003) and, more popularly, Bostrom (2014), Floridi navigates a middle way towards a computational model of semantics without committing to outright digital ontology.

of thought," (ibid, p. 182), I will offer a general definition of creativity as the act of meaning making in a universe of heterogeneous environmental data, and I will further assert that modelling this type of cognitive activity is, in a general sense, the target of my research.[2]

This then pushes my research into the broad domain of *computational creativity*, a field outlined in the seminal work of Boden (1990) and subsequently formalised in terms of "behaviour exhibited by natural and artificial systems, which would be deemed creative if exhibited by humans," (Wiggins, 2006, p. 206). The thrust of this work and the theory and practice that have sprung up around it involves treating creativity in terms of state spaces of combinatory components susceptible in the most productive cases to transformational transgressions of the rules for traversing the space, resulting in artefacts (and, arguably, processes) which can be evaluated in terms of their novelty and value (see Colton, 2008; Jordanous, 2012; Ritchie, 2007, among others for interesting theoretical work on the evaluation of computational creativity). If meaning making is to be construed in terms of creativity, and creativity is in turn modelled as a process of combination and composition, then at the root of the computational application of a theory of data, information, and meaning we encounter another essentially contested concept, namely, that of *representation*.

Representations have played a roll in philosophy of mind certainly since Descartes (1911) and Hobbes (1651), and by any but the most abstracted interpretation at least since Plato (1892)—perhaps they are a necessary passage in any movement towards a robust theory of mind (if, in fact, such a theory is even desirable—*cf* Rorty, 1979). The recent trend in philosophy, however, not to mention in empirically fastidious fields such as cognitive science and psychology, has been towards a resolute materialist reductionism, to such an extent that Rowlands (2010) reports that in the current cognitive scientific milieu, "even the word 'Cartesian' is often used as a term of abuse," (ibid, p. 12). This has been bad news for representations which, when applied to a theory of mind, can degrade into a homuncular regression that Dennett (1991) has described as the *Cartesian theatre*: if something is being represented, and something is doing the representing, who or what is at the receiving end of the process? The embodied and enactivist school of thought instigated by Maturana and Varela (1987) and pursued by, for instance, Haugeland (1993) and Thompson (2007), has led to the reanimation of discourse regarding the nature of mind from a perspective that does not take the *explanatory gap* (Levine, 1983) between what is subjectively experienced and what is objectively described for granted. Subsequently

[2] Creativity is itself, as Colton et al. (2014) have pointed out specifically in the context of computational approaches, an essentially contested concept, but, in the spirit of Gallie (1956), I will presume that there is significant value in identifying creativity as a boundary condition of sorts for the range of activities that I wish to explore without reaching a conclusive definition of the concept.

Gelder (1995) has outlined the premise of a mathematically tractable model of non-representational cognitive systems described in terms of dynamically coupled differential equations, while the emergentist system theory of biosemioticians like Kauffman (1995), Hoffmeyer (1997), and Pattee (2001) have provided fertile material for the sophisticated and evolutionarily plausible cognitive model of Deacon (2011).

But these anti- or post-representationalist approaches to cognition tend to unravel a bit when it comes to saying anything about language. In this particularly well travelled domain, the type theory of Whitehead and Russell (1927) and Church (1940) still holds a certain sway, with the subsequent formalisms of intensional semantics (Fox and Lappin, 2005) treating language as an ineluctably symbolic phenomenon. As such, there is an overt representationalism that is more or less necessarily at play in the symbolic commitments made by any sustainable theory of semantics, particularly in the context of natural language. Regardless of whether the representations in question are strictly in the mind, a theory promoted by Fodor (2001), or are in some sense in the world in line with the philosophy of Putnam (1975), it becomes difficult to imagine an operational model of semantics which doesn't fall back on structures which are to some extent extracted from the reality that they denote.

McGregor et al. (2014) have presented something of a start towards addressing or, perhaps more to the point, avoiding this issue (and the issue has been subsequently explored by Coeckelbergh (2016), in both cases specifically with reference to computational creativity). The idea put forward there is that, in the context of computational creativity in particular, it should be acceptable to take seriously the evident efficacy of talking about representations when talking about cognitive processes without necessarily making a commitment to the fundamental reality of such representations. I will stick to this position in the work presented in this thesis: by starting with the assumption that representations are a useful, maybe even necessary, component when talking about semantics and meaning, I maintain that we might eventually arrive at a more satisfying resolution of why this kind of structure has held such sway over the modern Western tradition of analytic philosophy in particular, and whether this influence is fundamental or just incidental. I don't claim to come close to actually answering this hard question, but I do think that there will be apparent merit in taking my methodology seriously as an empirical tool for gaining some sort of theoretical traction in this regard. So, in summary, in the following chapters, I will be describing a methodology which traffics in a particular theoretically motivated variety of meaning bearing representation, without making any commitment as to the essentialism of that device; the desideratum of these representations is that they be susceptible to the environmental situatedness that is clearly an important component of any effective cognitive or linguistic model. My contention here

is that sound theoretical grounding based on insight from cognitive science should grant my models a degree of at least temporary immunity from accusations of dualism.

2.2 Concepts

As Searle (1983) points out, representations have intentional content: they have to be about things, whether or not they take the form of materially or abstractly transportable entities like words or icons. The intentionality of representations invites the addition of another term to our growing catalogue of essentially contestable concepts, this time the word *concept* itself, which I will take to refer to the cognitive aspects of the things indicated by representations. The idea that concepts are interactive structures of the mind (Fodor, 2008; Margolis and Laurence, 2007) has been productive in aligning cognitive science with computational modelling (Boden, 2006). If concepts can be modelled as rule bound composite symbolic entities, then a symbol manipulating, constraint satisfying device should provide the right kind of architecture for simulating productive interactions between conceptual representations. This type of modelling has proven practically effective in, for instance, the structured ontology of Lenat (1995) and the graph theoretical work of Sowa (2006).

There is discord afoot, however, amongst researchers interested in modelling concepts, parallel to a certain extent to the debate over mental representations outline in the previous section. The net result of this tension has been the generation of a kind of negative space: where philosophers like Fodor and Pylyshyn (1988) have made a convincing case against treating concepts as associationist networks, more recent cognitive scientific research from the likes of Hutto (2001) and Chemero (2009) offers a likewise compelling rebuke to any theory of mind that falls back on a framework of symbolic conceptual representations. What remains is a clearly developed picture of what cannot constitute a concept in a cognitive model, but a much more murky impression of what positively does count as a thought or a perception and so forth. A remedy of sorts is offered by Gibson (1979), with his view of cognition in terms of the direct perception of environmental *affordances* of opportunities for action in a situation. Clark (1997) has expanded upon this to arrive at a notion of *action-oriented representations* which outsource much of the computational load of conceptualisation to the physical and spatial domain of a cognitive agent's environment.

Here Kant (1996) has proved to be, perhaps not surprisingly, especially profound: the Kantian notion of a domain of *conception* that is supervenient upon an underlying field of emphintuition which is in turn grounded in the essentially geometric nature of reality

provides a philosophically robust starting point for a spatial model of conceptualisation. By positioning conceptual models geometrically, the components of concepts which give them the composability that symbolic models afford while at the same time maintaining some degree of contact with the potentially physical context of space. The work of Gärdenfors (2000) is particularly germane here, and will serve as a primary point of reference for the methodology that I present in this thesis. By modelling concepts in terms of convex regions within conceptual spaces defined by interpretable dimensions representing attributes of the concepts themselves, Gärdenfors provides a plausible intermediary between the low-level stimulus to which a cognitive agent is exposed in an environment and the high-level symbols that become the representational currency of thought and communication: stimuli provide the data which becomes the values defining the points in a symbolically realisable conceptual space. More recent work has explored the way that a conceptual space model can be applied to lexical semantics in order to provide a geometric grounding for the categorical nature of language composition (Gärdenfors, 2014).

The environmental grounding of a conceptual model further provides a mechanism for understanding the important role of *context* in cognition. Here Barsalou's (1992) work modelling concepts in terms of *frames* offers a valuable perspective on the way that particular conceptual schemes are activated in response to situations in the world. Barsalou's approach facilitates notions of prototypicality and periphery that emerge in the course of online, context sensitive conceptualisation, once again at least hinting at a spatial component of this cognitive framework. Also of note is the *conceptual blending* approach of Fauconnier and Turner (1998), which makes use of a spatial theory of mind to develop a framework of conceptualisation as integration between frames of representation. This approach has been applied in the domain of computational creativity in particular, to the generation of language in the case of Veale (2012) and to automatic software generation by Znidarsic et al. (2016). And it is also worth mentioning the *global workspace* framework proposed by Baars (1988), which models cognition as a multi-agent system in which functional components compete and collaborate to forge a situated cognitive gestalt: this approach has been adopted by Shanahan (2010) in his work on cognitive robotics and by Wiggins (2012) again in the domain of computational creativity. A common and significant theme here is the dynamism and distribution inherent in all these approaches, contravening conceptual models that resort to static and hierarchical representational regimes.

Ultimately, I think we have to take seriously Davidson's (1974) case against the idea of conceptual models in the first place. Davidson's point is not so much that there is no such thing as a concept – that would be a fatuous claim – as that concepts are an

artefact of the way that cognitive, and in particular linguistic, agents use meaning bearing representations to structure thought and communicate about experience. At first glance this view of concepts might appear as facile as the denial of the existence of concepts is fatuous: obviously concepts have something to do with having thoughts, and it is probably impossible and certainly pointless to imagine a universe in which there are concepts but there are not cognitive agents. But the subtlety of Davidson's point is that there is a dynamic between conceptual models and representational structures which belies any kind of relationship of supervenience and complicates attempts to explain cognition in terms of levels of materialistic abstraction—as, in their own distinguished and insightful ways, Floridi (2011) and Deacon (2011) have each done. This dynamic turn invites a consideration of language as a concept supporting structure, and so sets us up for the next section of this survey of the established theory and practice surrounding my own work.

What we are then left with is the impetus for a computational approach which should be situationally dynamic and contextually sensitive. With this in mind, the methodology that is the focus of this thesis will be characterised by semantic representations that are designed to be understood as conceptually productive, contextually generated perspectives on spaces defined in terms of statistical data about language use. By using quantitative data to project representations into spaces that can be manipulated in an open ended way in response to a context which in principle can be arbitrarily defined, I will seek to mirror a theory of situated cognition permitting for the emergence of concepts in the course of the dynamics between agent and environment. As with my treatment of semantic representations themselves, I don't claim to be describing a methodology for conceptual modelling which is necessarily plausible on the level of physical or biological processes; instead I take certain assumptions about conceptual spaces for granted, and so there is an element of abstraction necessarily at play here. Once again, though, my stance is that allowing for some *a priori* assumptions about what is conceptually permissible provides a sound basis for getting on with the practical work of designing data driven experiments based on conceptual models and then turning around to apply the experimental outcomes to a productive reconsideration of theoretical assumptions.

2.3 Words

What has come to be known as the Cognitive Revolution finds its origin in, among other things, Chomsky's (1959) pointed denouncement of Skinner's (1957) attempt to apply psychological behaviouralism to the study of language. Chomsky's point is that language can only properly be understood as a specialised faculty that is in some way, more than

just a mode of stimulus and response, internal to the cognition of a linguistic agent: in order to effectively model language, we have to build some sort of notion of minds populated by cognitive content and attendant intentionality into the equation. For Chomsky and some of his acolytes, the logical extension of this view has been the development of a programme founded on the idea that language is itself an inborn characteristic peculiar to human cognition, certainly neurologically specific and quite possibly genetically encoded (Chomsky, 1986; Fodor, 2001; Pinker, 1994). A significant component of this project has been the development of various formulations designed to systematically encapsulate the conditions generally determining the parameters of natural languages, but for every attempt to categorically describe the particulars of human communication, linguistic anthropologists such as Levinson (2001) and Everett (2005) turn around and discover a group of language users who provide the exception which in the case of a scientific approach to language really does disprove the rule.

The movement against Chomskyan nativism has tended to swing towards what is arguably an even more fundamentally cognitive theory of language, often characterised by interpretations of Sapir (1970) and Whorf (2012) as a jointly declaring that language is, to a greater or lesser extent, actually the foundation upon which thought and attendant cultural spheres are built. More generally, the field of cognitive linguistics has emerged in response to the mainstream linguistic stance supporting theories of universal grammars, and a battery of interrelated linguistic models have emerged from the idea that language is, along with various other aspects of human behaviour, broadly wrapped up in and symptomatic of the general condition of having a mind rather than a compartmentalised cognitive faculty (Croft and Cruse, 2004). Of particular relevance here is the *cognitive grammar* of Langacker (1987), which proposes to overcome the divide between syntax and semantics by treating phonological and morphemic components of language as inextricably intertwined with semantics in ways that supersede evident distinctions across what Langacker calls *grammatical classes* (conventionally, parts of speech, basically). Also of note are the *image schema* of Lakoff (1987) and Johnson (1990), who, by focussing their analysis on the way that preposition usage in particular suggests distinct culturally specific embodied models of the world, developed environmentally and biologically grounded frameworks for productive semantic composition.

A general methodological commitment of cognitive linguistics is the qualitative analysis of instances of language use applied to the development of critically rich models of how conceptual and linguistic representations interface in the course of situated cognition. It should not be presumed, however, that cognitive linguists take semantic and conceptual representations to be identical or even isomorphic, and in fact Evans (2009) argues specifically that it is the nebulousness of the relationship between these domains

that gives language its particular qualities of looseness and ambiguity by which lexical representation can be deployed in context specific ways to achieve an open-ended expressivity. This aspect of semantics is particularly evident in the phenomenon of figurative language, and the study of metaphor has been an especially successful pursuit here, with a valuable compendium of the productive era from the late 1970s through the 1980s assembled by Ortony (1993). Exemplary theoretical work grounding the seemingly unlimited generative capacity of figurative language in a robustly cognitive approach to linguistics includes the *interaction* view of Black (1955,7) and the *reconstructivist* stance of Ortony (1975). It is the *cognitive metaphor* approach of Lakoff and Johnson (1980), however, which stands out most of all here, not least because it has provided the most consequential material for latter day computational research into metaphor classification and interpretation (Shutova, 2015). The description of metaphor in terms of isomorphic mappings between conceptual domains lends itself to precisely the type of symbolic manipulation of information structures that have characterised traditional AI, and, as it turns out, can also provide a theoretical grounding for sophisticated statistical modelling of lexical semantics (Shutova et al., 2012).

Statistical approaches to lexical semantic modelling will be surveyed in more detail in the following section, but a brief overview of information processing applications of the theory surrounding metaphor seems appropriate here. Some early computational approaches to metaphor maintained an essentially formal character: van Genabith (2001) proposed a type theoretical model for describing metaphor. Information processing approaches have, though, been by and large data-driven, understandably utilising the processing power of symbol manipulating machines—and these data-driven approaches have generally had some sort of connection with the cognitive linguistic stances on metaphor. So, for instance, Thomas and Mareschal (1999) describe an information processing network which selectively projects features, inspired by the previously mentioned interaction view of metaphor developed by Black (1977). In terms of theoretical grounding, Shutova (2010) identifies the *selectional preference violation* approach of Wilks (1978) as especially influential, perhaps because it was formulated specifically as an information processing mechanism. A notable early effort from Fass (1991) is derived from this theoretical background, with correspondences in the selectional preference of the arguments of verbs used to detect metonymic versus metaphoric uses of language.

The mainstream of metaphor modelling has subsequently been characterised by symbol manipulating approach and, in the spirit of the conceptual metaphor model, has involved mapping between conceptual schemes (Indurkhya, 1997), often domain specific, with the underlying assumption that mappings between domains correlates with the conceptual metaphor model (Narayanan, 1999). Typical symbolic approaches to

metaphor modelling involve the construction of an ontology defined by features which can be mapped between elements. The ATT-Meta system (Lee and Barnden, 2001), with its faculty for backchaining inferences across conceptual domains, is exemplary, and has furthermore been expanded into a metaphor generating system employing a combination of distributional semantic and incremental grammar techniques (Gargett and Barnden, 2013). ATT-Meta is particularly notable in that it applies systems of logic in the specific conceptual context of a metaphor it is handling (Barnden and Lee, 1999), and in this respect is a symbolically grounded response to some of the same theoretical concerns that have motivated my own research. Other symbolic approaches are notable for their recourse to pre-formulated knowledge bases such as WordNet (Veale et al., 2015), or the web at large in conjunction with other resources (Veale and Hao, 2007).

Symbolic approaches have tended to focus on the interpretation of metaphor by way of models of trans-conceptual mappings, but in another aspect of computational work, that of metaphor identification, statistical approaches have proved particularly effective.[3] An early example is the TroFi model of Birke and Sarkar (2006), which uses a clustering algorithm trained on a set of tagged sample sentences to disambiguate between literal and non-literal verb use, followed by Utsumi (2011), who explores clustering in the context of distributional semantics. Indeed, many of these statistical approaches (see Turney et al. 2011, Dunn 2013 for a comparison of distributional semantic and symbolic models, Shutova et al. 2012 for an overview of statistical models in particular) have employed the techniques of distributional semantics, which will be discussed in the next section: here Kintsch's (2000) model of metaphoric interpretation as a contextually selective traversal of the space between word-vectors is seminal. A notable recent instance of a statistical model for metaphor identification involving an application of compositional distributional semantics is described by Gutiérrez et al. (2016), of particular note here as the dataset presented by those authors will be used to evaluate the model at the heart of this thesis (see Chapter 6 for a more detailed description). Returning to the cognitive linguistic foundations of computational approaches to metaphor, Tsvetkov et al. (2014) go so far as to propose that their results derived from the statistical construction of what they construe as conceptual features associated with lexical representations "support the hypothesis that metaphors are conceptual, rather than lexical, in nature," (ibid, p. 248).

There is another theoretical twist which must be mentioned here, however, and it comes once again from Davidson (1978), this time by way of his controversial claim that the meaning of metaphoric propositions should always be taken at face value. Part of Davidson's point is that there is a pragmatic distinction to be drawn between what

[3]Shutova (2013) suggests that computational identification and interpretation of metaphor, in line with psychological analysis, should be considered a joint task.

the metaphor means, which is to some extent in the language, and what the metaphor communicates, which is on the other hand in the world.[4] The presumption in both conventional semantic views of metaphor such as Searle's (1979) as well as the more strongly cognitivist stances discussed above is that metaphor necessarily involves the projection of some aspect of meaning from one conceptual domain to another, but the point that Davidson raises is that there is a limit to the cognitive content that can be propositionally conveyed by language, and metaphor often reveals that limit. To borrow a popular example from the discourse surrounding relevance theory (Carston, 2012; Jr and Tendahl, 2006, for example), there is a lurking breakdown in interpretation when we try to apply any sort of transference view of metaphor to a statement such as "my boss is a bulldozer": presuming a small degree of contextual knowledge, we might easily understand that the speaker means the boss in question is inappropriately insensitive or aggressive in dealing with employees—but it is hardly clear what actual properties of BULLDOZER are transferred to BOSS, particularly in a situation which might very well not even be physical.

To address this issue, Carston (2010) proposes that metaphor necessarily involves the generation of *ad hoc concepts* that come about in the process of making a lexical mapping from one domain of encyclopaedic knowledge to another. Drawing on Barsalou's (1993) notion that language produces concepts in a way that is inherently *flexible* and *haphazard*, *ad hoc* concepts offer a relevance theoretical account of the way in which language always pragmatically, situationally specifies the semantic content of an utterance (Sperber and Wilson, 1995). This accommodates the *deflationary* view of metaphor put forward by Sperber and Wilson (2012), which holds that metaphor merely occupies an especially inferential extent of a spectrum of meaning making and interpreting activities. At stake here is the idea that language is not so much a system for codifying propositions about the world as a mechanism for achieving optimal communication of cognitive content, with the important proviso that cognition itself is primed for a perpetually unfolding contextualisation of the environmental stimuli available to an agent. This ultimately means that metaphor is able to be more than just a highly efficient way of encoding propositions about concepts; it can, even in relatively mundane instances, extend itself into domains bordering on the phenomenological, a stance eloquently summed up by Reimer (2001) in her apologetic exegesis of Davidson: "For the goal of the metaphor-maker is not to get the hearer to see that something is the case, to grasp some deep and subtle truth, but to see something in a certain way, and seeing something in a certain way is simply not the sort of thing that can be given literal expression," (ibid, p. 150).

[4] Davidson's account, which is famous or perhaps notorious amongst theoretical linguists, is notable in its absence from the computational literature, though it has recently been acknowledged at least in passing by Veale (2016).

With all this in mind, we arrive at a further specification for the boundary conditions of our computational semantic model: in addition to being a representational system with a capacity for summoning context specific relationships between lexical semantic entities, it should also be able to generate new conceptual representations in an *ad hoc* manner. This implicates the modelling of conceptual spaces that are not merely invoked by the process of specification inherent in communication, but actually generated in the course of lexical dynamics. And the situated, even arbitrary production of conceptual relationships in turn suggests, beyond just the activation of existing or implicit networks of association between semantically tractable entities, the online creation of entirely new connections and correspondingly of new ideas: put simply, the open-ended generation of conceptual spaces is the machinery of meaning-making. It seems more or less impossible to imagine a regime of strictly symbolic representations which could fulfil these requirements, because symbols necessarily come with the logic and extent of their combinatory potentials, setting the constraints for the state space of their potential for interactive conceptualisation, more or less built in. Instead, I propose that a statistical approach, in which lexical semantic representations are defined in terms of observations of symbols in use rather than rules applied directly to symbols, will offer the right kind of flexibility and dynamism for modelling the situated nature of concepts and the rampant looseness inherent in the relationship between words and objects of the mind.

2.4 Data

Finally, arriving at the technical background for the instantiation of the system of context sensitive, semantically productive representations outlined above, the research described in this thesis is grounded in recent and ongoing success in the paradigm of *distributional semantics*. The tradition of word-counting in order to predict sequences in language traces its roots back to the fastidious work of Andrei Markov, who tabulated co-occurrences of characters in Pushkin's *Eugene Onegin* by hand (Basharin et al., 2004), and Shannon and Weaver (1949) propose a comparable application in their seminal work on information theory. The idea of applying co-occurrence statistics to semantic applications is central to Harris's (1954) work examining "meaning as a function of distribution," (p. 155); the various consequent formulations of the *distributional hypothesis* have been outlined by Sahlgren (2008), with Pantel's (2005) asseveration that "words that occur in the same contexts tend to have similar meaning," (ibid, p. 126) being representative.[5]

[5]Scholars frequently cite Firth's (1959) quip "you shall know a word by the company it keeps," (ibid, p. 179) as being foundational in the field. I contend that Firth was referring in this passage specifically to the study of idiomaticity, particularly the way that idioms ossify culturally through repeated use, and this in the context of a larger proposal for a heterogeneous approach to the study of linguistics more in

Theoretically speaking, computational linguists have ambitiously sought to ground distributional semantics in the formal semantics of Frege (Baroni et al., 2014a) or indeed in the pragmatics of Wittgenstein (Grefenstette and Sadrzadeh, 2011).

Rather than indulge in speculation of what Wittgenstein might have done with a computer, I will propose a perhaps even less likely candidate as the philosophical forbearer of word-counting as a productive applied linguistic practice: the semiotics of Peirce (1932), which maintain that the very physiognomy of meaning bearing structures, or *signs* in Peirce's parlance, are semantically productive by way of their very physiognomy, and that they gain this productive structure through their ongoing contact with their environment. From his own analysis of Peirce, Eco (1976) extrapolates a notion of *unlimited semiosis* by which signs participate in an infinite regression of semantic productivity, with one sign becoming the substrate for the constitution of a subsequent sign. This begins to look, in an abstract way, a bit like the distributional semantic regime, where the sentential context in which words are found becomes the substance of interactive lexical semantic representations. Another historical touchpoint is, as Miller and Charles (1991) have pointed out, the *salva veritate* of Leibniz, by which, in terms of logical formalisms, terms are considered to be synonymous if they can be universally interchanged in logical expressions without changing the truth values of the expressions. Exporting this notion to the domain of computational linguistics, we arrive at the central dogma of distributional semantics, namely, that words can be modelled in terms of observations of their co-occurrence tendencies across large scale corpora, and furthermore that words with similar profiles can be interpreted as being likewise semantically associated.

Practically speaking, early work from, for instance, Salton et al. (1975) suggested that the information content of documents could be effectively indexed by representing them as points in a vector space whose dimensions correspond to weighted measures of word frequency within a given document. Schütze (1992b) extends this insight to represent words as vectors defined by the frequencies with which they are observed to co-occur with other words in a corpus, and uses angular measures from the consequent vector space as grounds for disambiguating the senses of polysemous words. An important result of modelling words in terms of their co-occurrence profiles is that two words which have never been observed in proximity to one another might nonetheless turn out to be very close in the model and therefore very similar to each other: so, for instance, we can imagine a language in which the words for CAT and DOG are prohibited from ever being used in the same sentence, but we might still discover a semantic correspondence between the concepts because their signifiers tend to have similar patterns of usage. The

line with the comprehensive emergent view of MacWhinney (1998) rather than anything that could be construed in terms of a computational, word-counting practice. All the same, the quote has a nice ring to it and, taken out of context, serves its purpose.

conversion of raw word counts into weighted statistics, perhaps most basically through the application of term-frequency, inverse-document-frequency type metrics (Salton and Buckley, 1988) but more typically in more recent applications with information theoretical functions (Turney, 2001), has produced particularly productive co-occurence based lexical semantic representations. The geometric efficacy of passing co-occurrence statistics through logarithmic functions will be discussed in Chapters 4.4 and **??**. The end product of this type of approach is fundamentally that words are mapped into spatial relationships with one another, where the geometry of the space itself is to a greater or lesser extent semantically productive, and authors such as Landauer and Dumais (1997) have explored some of the psychological and philosophical ramifications of this.

The vector space approach to distributional semantics has subsequently evolved into a productive computational programme. The distributional semantic methodology usually involves the selection of a corpus, the traversal of this corpus in order to tabulate the counts of co-occurrence terms within a certain proximity of target words (typically defined in terms of a window of k words around each observation of a target word), the application of a weighting function to the resulting co-occurrence matrix, and the projection of the weighted vectors into a space (see Turney and Patel 2010 and, more recently, Clark 2015 for comprehensive overviews). Bullinaria and Levy (2012) have reported comparative results based on a variety of weighting schemes, most notably *positive pointwise mutual information* (PPMI), an information theoretical metric designed to build sparse matrices capturing the most semantically salient co-occurrence features of word-vectors. Where PPMI simply disregards co-occurrences that are observed at a frequency below the overall corpus average, Levy et al. (2015a) explore a slightly more subtle techniqe of shifting their co-occurrence statistics to avoid massively negative logarithms; a similar metric will be the basis for my own methodology. The construction of distributional semantic models also often involves an additional step of dimensional reduction by way of, for instance, principal component analysis, with a particularly notable technique involving singular value decomposition described by Deerwester et al. (1990).

Distributional semantic models have evolved out of the practical requirement for effective and efficient document retrieval based on textual queries, but the linguistic tasks subsequently tackled have included entailment (Baroni et al., 2012; Geffet and Dagan, 2005; Rimell, 2014), word sense disambiguation (Kartsaklis and Sadrzadeh, 2013; Schütze, 1998), and sentiment analysis (dos Santos and Gatti, 2014; Malandrakis et al., 2013), among other things. A particularly interesting development has been the use of linear algebraic operations on representations to facilitate language composition (Mitchell and Lapata, 2010). By treating, for instance, nouns as word-vectors and adjectives as tensors, Baroni and Zamparelli (2010) describe a model for projecting adjective-noun phrases

into a vector space in which these compound linguistic entities can be compared using the same approaches applied to word-vectors. Borrowing from the mathematical arsenal of quantum mechanics, Coecke et al. (2011) conceive a correspondence between distributional semantics and formal semantics, modelling syntactic elements as vectors and tensors based on observations across a corpus that map to category theoretical components of a grammar, pushing whole sentences into vector spaces allowing for comparison between sentences and the assignment of truth values. The import of all of this is, once again, that the modelling of semantic units using high dimensional representations provides a productive and computationally tractable grounding for a variety of linguistic phenomena.

The development of high powered computers and the related advent of massive corpora of digitised textual data has facilitated another turn in the distributional semantic programme: the application of neural networks to data driven semantic modelling. Bengio et al. (2003) is an early proponent of this approach, demonstrating that the application of iteratively learned word-vectors consisting of abstract features is an effective mechanism for language modelling, followed by Collobert and Weston (2008), who use a convolutional neural network to build a vector space model suited to learning to perform a number of supervised and semi-supervised linguistic tasks including semantic modelling, language modelling, and sentence parsing. And the contribution of Mikolov et al. (2013c), dubbed `word2vec`, has been one of the most widely discussed developments in the field in recent years, offering up a highly generalisable set of models with particularly remarkable capacities for modelling the semantically significant phenomenon of analogy, which will be discussed in more detail in Chapter 7.

The dichotomy between co-occurrence statistic based models, almost always complemented with some dimension reduction technique such as a principal component analysis, and neural network approaches has led to a productive tension in the field, summarised by Baroni et al.'s (2014b) in terms of *counting* to derive statistically defined word-vectors versus *predicting* what have sometimes been called *word embeddings* using a neural network—though it should be noted that both methodologies necessarily act on observations of word co-occurrences made in the course of the traversal of a corpus, and both types of model have been successfully configured for the kind probabilistic output involved in, for instance, language modelling. And, where Baroni et al. ultimately decide that neural network based approaches offer a more robust extrapolation of semantic representations from corpus data, Levy and Goldberg (2014b) have argued that the superficial differences between the two broad methodlogies can be understood in terms of decisions regarding the tuning of the extensive range of hyperparameters inevitably associated with either type of model. Along similar lines, one of the main findings of

this thesis, and a motivation for the methodology I've developed, is that, once a layer of removal from the data has been applied to statistical models through for instance singular value decomposition, they, like neural network models, become immune to context specific manipulation, because their dimensionality becomes abstract and uninterpretable.

One consequence of the collegial arms race between the two approaches has been the development of increasingly task specific systems, often coupling distributional semantic models with heuristics involving the identification of syntactic patterns or the extraction of information from pre-formulated knowledge bases. In response to this, Baroni and Lenci (2010) have described an ensemble of vector space models packed into a tensor space of potential relationships between lexical entities—a model of models of sorts, capable of selectively activating the appropriate component of its representational hyperspace based on an assessment of the task at hand. This is well motivated, and I have sought to develop a similarly generalisable methodology, but in the case of my research the generalisability arises from the ability of my models to selectively project an astronomical range of context specific semantic subspaces rather than from an extra layer of model specification. In practice, my methodology will be tested against a battery of existing tests designed by fellow researchers in the field of computational linguistics, including word relatedness (Finkelstein et al., 2002), word similarity (Hill et al., 2015), metaphor classification (Gutiérrez et al., 2016), semantic type coercion (Pustejovsky et al., 2010), and analogy completion (Mikolov et al., 2013c).

So finally we arrive at something like a way forwards towards the computational modelling of context sensitive lexical semantics. Distributional semantics provides a mechanism for the production of dynamically interactive representations based on observations of large scale textual data, offering up a malleable lexicon suited to the rampant contextualisation indicated by theoretical insight into concept production. To chart a passage through the territory mapped throughout this chapter, then, statistics reflecting the co-occurrences of words in a large scale corpus will serve as the data substantiating the informational character of dynamic lexical semantic representations which, in their interactions, will be projected into conceptually interpretable spaces that are in turn reflective of the evidently representational character of meaning making. With this apparatus in place we can now move on to the task of a theoretical description of my own methodology in Chapter 3, followed by a technical description of the consequent computational in Chapter 4.

Chapter 3

Semantics in Context

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

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

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

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

These ideas, which have been outlined throughout the previous chapter, are not the standard dogma of computational linguistics, which generally, and understandably, has modelled concepts as modular, portable entities, language as a likewise stable system of representations and rules, and the relationship between the two as one of source and

contingent data (see, for instance, the textbook treatment in Jurafsky and Martin, 2000, particularly Ch. 17). This structure-oriented approach to language and mind epitomises a project that Dreyfus has described as "finding the features, rules, and representations needed for turning rationalist philosophy into a research program," (Dreyfus, 2012, p. 89). As computer science and philosophy of mind increasingly interact at the vertex of cognitive modelling, culturally relative ideas about the connection between mental representations and linguistic symbols become incorporated into the very architecture of data structures, engendering a positive feedback loop by which the outputs of symbol manipulating information processing systems reinforce the premise that representations are stable entities which can be trafficked in the form of words according to the rules of a grammar.

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

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

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

2. The method should be dynamically sensitive to context;

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

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

The first stipulation is a fundamental criteria for computational approaches to lexical modelling, if not to lexical semantics in general, and is to a certain extent built into the distributional semantic paradigm at the root of my methodology. The second stipulation encapsulates the theoretical premise of this work. A primary objective of my methodology is to identify a statistically tangible mechanism for choosing word co-occurrence features in a contextually relevant way. Specific mechanisms will be outlined in Chapter 4, and what counts as context will be discussed further in the course of the empirical results presented in Chapters **??** and 6, but the general idea is that a context sensitive model needs to react dynamically to information about what's happening in some linguistic situation. The second requirement follows directly from the first: in order to pick semantic contexts *in situ*, there needs to be a way to get a handle on the data which underwrites a model. In practice this means that the scalars that form the basis for all the models which will be explored here represent literal information about co-occurrences in a large scale corpus, and the feature selections that take place in the course of delineating a contextual geometry can be traced to specific events in the underlying data.

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

pretive features, ultimately allowing for the observation of different semantic properties – for instance, similarity versus relatedness – to emerge as different geometric aspects of the same subspace.

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

3.1 Modelling Lexical Semantics

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

The idea of structurally productive lexical representations finds its roots even deeper in the tradition of the philosophy of signification, in the semiotics of Peirce, who suggests that "there must exist, either in thought or in expression, some explanation or argument or other context, showing how–upon what system or for what reason the Sign represents the Object or set of Objects that it does," (Peirce, 1932, ¶2.230). In other words, a representation denotes and means by virtue of the actual dynamics of the symbol itself

as it exists in the world. There is a story to be told about how a representation comes to operate in the way that it does: as Rączaszek-Leonardi puts it, meaning-bearing symbols "can arise only from the history of a certain physical structure as a constraint on certain system's dynamics in a certain environment," (ibid, p. 309). Furthermore, a lexical representation's acquisition of its dynamics unfolds on a number of different timescales, for instance on the scale of individual cognitive development as well as on the scale of the history of cultural transmission, effectively prohibiting any attempt at an elimitivist interpretations of linguistic symbols as atomic units. Instead, language is, within the regime of environmentally situated cognition, taken to be a cognitive object which affords meaning-making and conceptualisation; as Clark suggests, there is scope to "consider language itself as a cognition-enhancing animal-built structure," (Clark, 2006, p. 370). Given this objective and even material quality of language, it seems clear that a good model of lexical semantics should traffic in symbols which are likewise susceptible to conceptually productive, open-ended manipulation.

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

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

[1] Type coercion will form a test case for my model, explored empirically in Chapter 6.

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

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

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

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

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

3.2 Dynamic Context Sensitivity

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

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

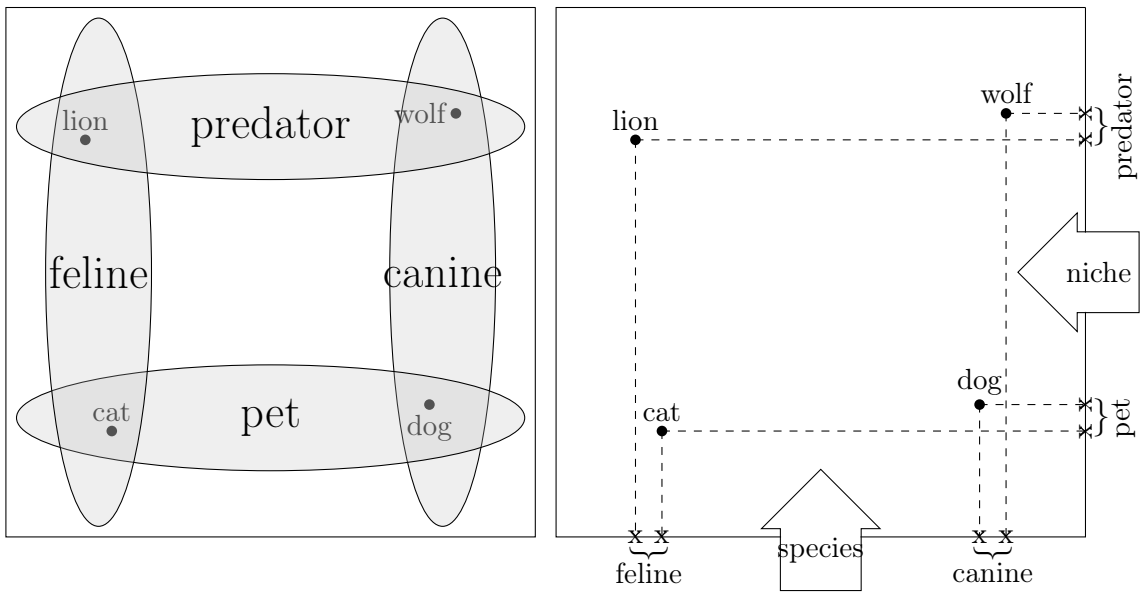
[3]It is perhaps not a coincidence that von Neumann was a seminal figure in the description of both the logic of lattice theory (Birkhoff, 1958) that has motivated more recent developments in concept modelling such as formal concept analysis (Wille, 1982) and the modular architecture of memory and processing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

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

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

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

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



(a) a lexical space

(b) a contextual projection

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

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

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

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

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

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

3.3 Literal Dimensions of Co-Occurrence

The model presented here is grounded within the paradigm of distributional semantics, which means that the conceptual geometries that it constructs are the product of observations of word co-occurrences in a large-scale corpus of textual data represented statistically. Two procedurally distinct methodological regimens have emerged from the recent

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

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

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

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

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

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

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

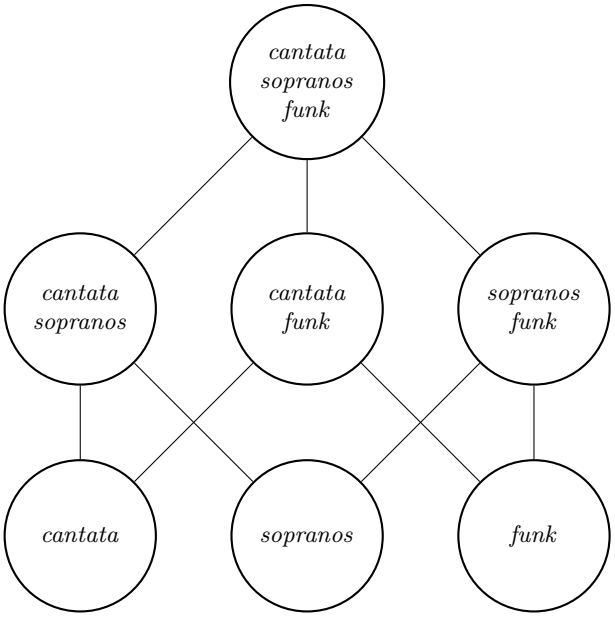


Figure 3.2: A lattice of three dimensions, including the two-dimensional subspaces which are used for analysing the conceptual geometry of a small set of word-vectors in Figure 3.3

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

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

The methodology proposed in this thesis adheres to the same principle of fine-grained cross-dimensional analysis described by Rimell. In addition to the practical issues raised

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

3.4 Interpretable Geometry

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

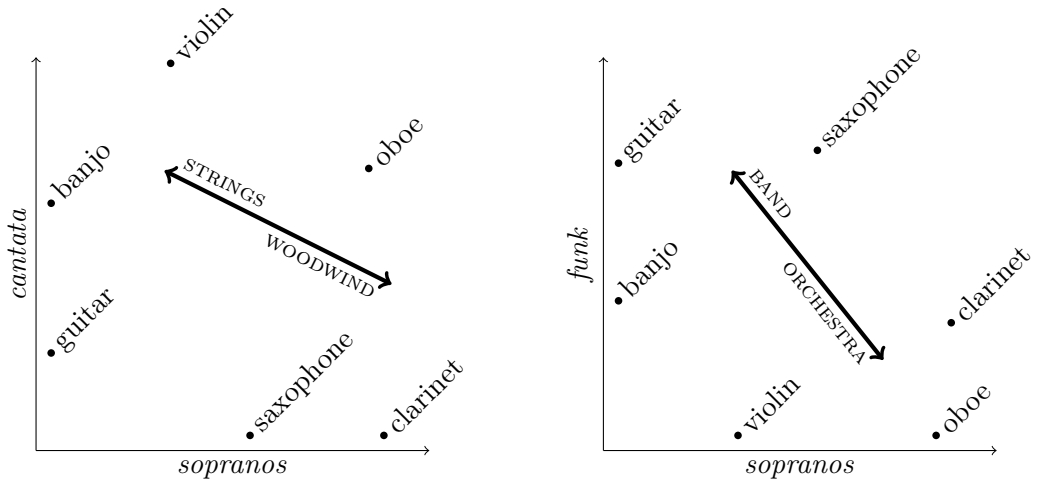
The idea of exploiting the geometry of a transformed space of word statistics is not new. Indeed, seminal work on latent semantic analysis was motivated by precisely the insight that a singular value decomposition of a high-dimensional, sparse matrix of statistical data about word co-occurrences would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

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

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

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

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



(a) STRINGS VS WOODWIND (b) BAND VS ORCHESTRA

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

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

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

The first thing to note about these spaces is the way that swapping a single dimension in a two dimensional subspace can have a significant impact on the conceptual affordances

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

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

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

The spaces illustrated in Figure 3.3 are the product of a survey of a lattice consisting of combinations of just three dimensions, and as such the conceptual affordances of this

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

Chapter 4

Context Sensitive Distributional Semantics

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

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

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

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

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

4.1 Establishing and Analysing a Large Scale Corpus

The first step in a corpus based approach to natural language processing is the selection of the data which will provide the basis for our model. I've picked the English language portion of Wikipedia as my data source, a choice which is in accordance with a good deal of work done in the field. For instance, Gabrilovich and Markovitch (2007) and Collobert and Weston (2008), to name just a couple, use Wikipedia as their base data for training distributional semantic models designed to perform tasks similar to the ones explored in subsequent chapters, while Baroni et al. (2014b), Pennington et al. (2014), and Gutiérrez et al. (2016) use amalgamated corpora that include Wikipedia as a major component. Wikipedia provides a very large sample of highly regular language, meaning that we can expect a certain syntactic and semantic consistency as well as language which, if not always overtly literal, is likewise not typically abstruse or periphrastic. This should supply a source of linguistic data in which, to revisit the central dogma of the distributional hypothesis, words which occur in a particular syntactic and lexical setting can be expected to be semantically similar.

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

As is generally the case with data cleaning, these measures are prone to error: for instance, due to the removal of punctuation, the contraction *we're* will be considered identical to the word *were*. One of the strengths of the subspace projection technique that my methodology uses is its resilience to noise. So, for instance, misspellings will be

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

categorised as highly anomalous co-occurrence dimensions and are therefore unlikely to be contextually selected – or, if they are encountered regularly enough to be contextually significant, there may well be useful information in the co-occurrence profile of such mistakes – while, at the other end of the spectrum, essentially ubiquitous words are unlikely to provide context specific information, so the ambiguity between *we're* and *were* is unlikely to be drawn into any of the subspaces actually projected by the model.

From the cleaned corpus, a model's vocabulary is defined as the top 200,000 most frequently occurring word types. This cut-off point is very close to the point where the total number of word tokens included by selecting all instances of all vocabulary words equals the total number of word types excluded. Given the Zipfian distribution of word frequencies as observed throughout the corpus, this means that more than 95% of the co-occurrence data available from the corpus will be taken into account by the model, while the number of word-vectors used to express this data represents less than 5% of the potential vocabulary—a fairly efficient way of extrapolating statistics from the corpus. The selection of this as a cut-off point means that the least frequent words in the vocabulary occur 83 times throughout the corpus.

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

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

Here $f_{w,c}$ represents the total number of times that c is observed as co-occurring in a sentence within k words on either side of w, f_w is the independent frequency of occurrences of w, and f_c is likewise the overall frequency of c being observed as a co-occurrence term throughout the corpus. W is the overall occurrence of all words throughout the corpus– and it should be noted that, excluding the term a, the ratio in Equation 4.1 is equivalent to the joint probability of w and c co-occurring. The term a is a skewing constant used to prevent highly specific co-occurrences from dominating the analysis of a word's profile, set for the purposes of the work reported here at 10,000.[2] Finally, the entire ratio is skewed by 1 so that all values returned by the logarithm will be greater than 0, with a value of zero therefore indicating that two words have never been observed to co-occur with one another.

This last step of incrementing the ratio of frequencies in order to avoid values tending towards negative infinity in the case of very unlikely co-occurrences is again a departure from standard practice, where, in word counting models, a *positive pointwise mutual information* mechanism involving not skewing the ratio and instead treating any ratio of frequencies less than 1 – that is, any co-occurrence that occurs with a lower probability than the combined joint probability of independently observing w and c – as being equivalent to zero (Levy and Goldberg, 2014b, have considered a more general variable ratio shifting parameter). The motivation for this more typical technique is again to avoid incorporating unnecessary and potentially confounding information into a model, but, again, in the case of my model, the dimensional selection process will tend to ignore such information, and at the same time, as will be seen, data regarding relatively unlikely co-occurrences can sometimes also be quite informative. Other variations on the distributional semantic approach include alternative treatments of the co-occurrence window, where some researchers have taken weighted samples or considered word order (Socher et al., 2013), and also the processing of corpora, where part-of-speech and dependency tagging have been applied to positive effect (Padó and Lapata, 2007). Lapesa and Evert (2014) and Milajevs et al. (2016) offer comparative overviews of the effects of parameter variations on the performance of distributional semantic techniques.

The net result of my methodology is a matrix of weighted co-occurrence statistics, where higher values indicate a high number of observations of word w co-occurring with word c relative to the overall independent frequencies of w and c. Values of zero indicate

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

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

4.2 Selecting Dimensions from a Sparse Co-Occurrence Matrix

Context has thus far remained a somewhat abstract concept in this thesis. In principle, the context in which conceptualisation occurs for a cognitive agent is its environment with all its affordances, linguistic and semantic but also more generally perceptual: in a word, the agent's *umwelt* (von Uexküll, 1957). In the world of physical entanglements, language presents itself with precisely the same open-ended opportunities for action as other modes of cognition (Clark, 1997; Gibson, 1979)—and, in the case of language, the action afforded is meaning making. In practice, however, context will be specified lexically, in terms of a word or set of words which are fed to a model, analysed in terms of their co-occurrence profiles, and then used to generate a subspace of conceptually relevant co-occurrence dimensions. The intuition behind this approach is that there should be a set of dimensions which collectively represent a semantic tendency which can be mapped to a context, and this tendency should be discoverable in an analysis of the co-occurrence statistics of words which are exemplary of this way of talking about things.

So, notwithstanding interesting work on multi-modal approaches to distributional semantics from, for instance, Hill and Korhonen (2014) and Bruni et al. (2014), with regard to the present technical description, I will treat *contextual input* as meaning some set of words T which have been selected for the purpose of performing some type of semantic evaluation and act as input to a context sensitive distributional semantic model. The exact mechanisms for specifying T will be discussed in subsequent chapters with regard to each of the individual experiments to be performed using my methodology; for now, I offer a general outline. Each component of T points to a word-vector in the matrix M described in the previous section, and the collection of word-vectors corresponding to T serve as the basis for an analysis leading to the projection of a context specific subspace S. I propose three basic techniques for generating these projections, with the model parameter d indicating the specified dimensionality of the subspace to be selected:

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

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

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

These techniques are used for the purpose of analysis, and, once this analysis has been performed, the subset of dimensions returned is used to project the entire model vocabulary onto a d dimensional subspace. The JOINT technique requires the greatest finesse, as there is an element of cross-dimensional comparison at play. As such, for the purposes of this technique, the word-vectors selected by T are merged, dimensions with non-zero values for any of the word-vectors are discarded, and the resulting truncated word-vectors, each consisting of an equal number of non-zero dimensions, are normalised. This ensures that certain elements of T won't dominate the analysis: because the frequency of each word in T applies a deflationary pressure on the PMI values associated with the corresponding word-vectors, very infrequent words would be liable to dominate the analysis with the associated high PMI values in their profile. This effect is illustrated in Table 4-A, where PMI values for the top dimensions selected using the JOINT type subspace by the words *guitar*, which at 88,285 occurrences is ranked 1541 in frequency, are compared with those for the word *dulcimer*, which occurs 516 times and is ranked 62,313 (the base model here was constructed using a 5x5 word co-occurrence window). Among the dimensions with non-zero values for both words, normalisation brings the high end of the respective co-occurrence profiles more in line with one another, facilitating the selection of a subspace which is jointly characteristic of the input terms.

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

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

Table 4-A: The top five and bottom five dimensions by PMI value for the words *guitar* and *dulcimer*, out of all the dimensions with non-zero values for both words, with scores tabulated independently for each word.

values across T:

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

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

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

-

In the cases of the INDY and ZIPPED techniques, the selectional process is more straightforward, since mean values between features of word-vectors are not being considered. Where the JOINT technique is intended to discover subspaces that represent an amalgamation of the input terms, the INDY technique is expected to produce a subspace where individual conceptual characteristics of the input terms, captured as collections of co-occurrence dimensions, are distilled into distinct geometric regions. So the set of d dimensions I returned by the INDY technique will delineate a subspace in which the relative geometry of contextual input word-vectors will reflect the degree to which the independent co-occurrence profiles of those word-vectors overlap. So, given the set B of all dimensions and the input word-vectors $\{t_1, t_2...t_k\}$, I can be selected from this base set of dimensions:

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

The ZIPPED technique might be seen as something of a hybrid of the JOINT and INDY techniques, since it used the INDY approach to make selections from the intermediary space of non-zero dimensions available to the JOINT technique. Here we know there will be some information about every co-occurrence dimensions for each word-vector associated with the contextual input, and so we might expect a subspace that offers a more nuanced interpretation of semantic relationships between the contextual input in particular. The set of dimensions Z delineating this space is selected from the same set N described in Equation 4.2, in this case simply selecting the dimensions with the highest values for each input word-vector, as they have non-zero values for all the input word-vectors:

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

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

In order to offer a sense of what's happening with these dimension selection techniques, a preliminary and intuitively motivated case study of dimension selection is outlined in Table 4-B, again derived from a base space generated through observations made within a 5x5 word co-occurrence window over the course of the corpus. The top dimensions selected by each technique are presented for two different three term sets of input words: *lion*, *tiger*, and *bear*, on the one hand, which are taken to represent in their union exemplars of wild animals, and on the other hand *dog*, *hamster*, and *goldfish*, which are prototypical pets. The dimensions selected by the JOINT technique in response to the WILD ANIMAL type input include the names of other wild animals, as well as *paw*, a component of many wild animals, *mauled*, an activity performed by wild animals, and, interestingly, *mascot*, presumably because many sports teams take these types of ani-

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

Table 4-B: The top 10 dimensions returned using three different dimensional selection techniques, featuring one set of input terms collectively referring to wild animals and another set collectively referring to pets.

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

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

be explored in detail in relation to specific semantic applications in subsequent chapters.

A very broadly similar approach to distributional semantics has been proposed by Polajnar and Clark (2014), who describe a *context selection* methodology for generating word-vectors, involving building a base space of co-occurrence statistics and then transforming this space by preserving only the highest values for each word-vector up to some parametrically determined cut-off point, setting all other values to zero. Setting the cut-off point relatively stringently – generating a base space of more sparse word-vectors, followed by various dimension reduction techniques – led to improvements in results on both word similarity and compositionality tests. This suggests that allowing word-vectors to shed some of their more obscure co-occurrence statistics leads to a more sharply defined semantic space, and indeed there may be an element of disambiguation at play here, as well, with vectors dropping some of the features associated with less frequent alternate word senses.

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

4.3 Exploring the Geometry of a Context Specific Subspace

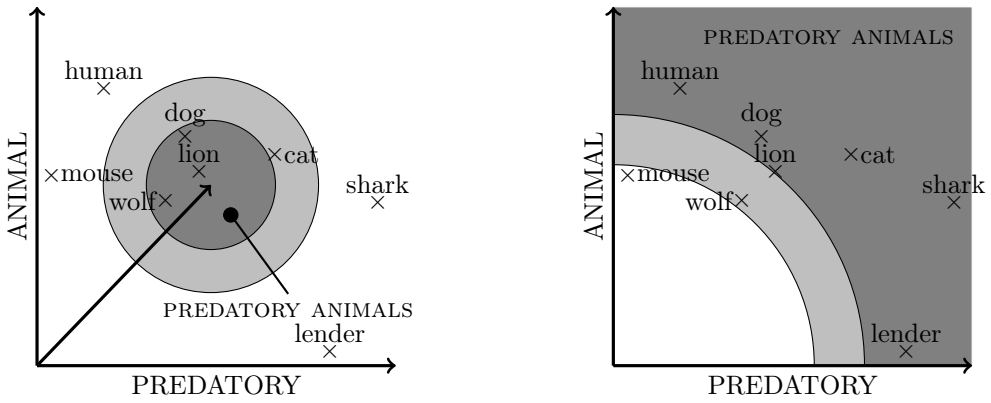
Before delving into the question of the types of geometries my method might be expected to generate, I would like to raise a point regarding the typical application of the term *geometry* to vector space models of distributional semantics in the first place. Widdows (2004) makes an enthusiastic and compelling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

Baroni et al. (2014a) go so far as to suggest that their distributional semantic model effectively instantiates the abstract principles of Frege's work on the logic of natural languages (Dummett, 1981) in a geometric mode. These are powerful points touching on the essence of semiotics, and the idea that representations that map from data to interpretable features in a space are core to my own methodology, as discussed in Chapter 3.1.

The point I would like to make now, though, is that there are different degrees of geometry that can be in principle accessed in a vector space of real valued dimensions. The great majority of approaches surveyed here, taken to be representative of the historical and ongoing trend in the field, present models consisting of spaces of normalised word-vectors, in which there is a monotonic correlation between the distance and the angle between two word-vectors. In the case of models built using a principal component analysis, this is because when the eigenvectors of a matrix factorisation are used as dimensions of maximal variance, there is no meaningful interpretation of the actual values along these dimensions; in fact, mean values along a dimension will tend towards zero and the signs of values along any dimension discovered through a singular value decomposition can be reversed without any degradation of the information available from analysis (Abdi and Williams, 2010). So, while Euclidean distance is strictly meaningful in such a dimensional reduction, there is no sense of a centre of the space other than the centre of gravity of the data as projected onto the selected number of eigenvectors, and cosine similarity is in practice the measure used to determine the similarity between two word-vectors. And in the case of models built using neural networks, there is no meaningful interpretation of dimensions to begin with, so the resulting space is a *de facto* hypersphere of word vectors that are only relative in terms of their relationship to one another, not their relationship to any objective features of the space.

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

The second effect is that there is a notion of centre and periphery in my subspaces. Since all values are positive, a word-vector with high scores across all or most dimensions in a subspace will be far from the origin and in the central region of the space. A further



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

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

Figure 4.1: Co-occurrence statistics for a small vocabulary construed along two hand-picked dimensions. Darker regions are expected to be more conceptually prototypical for the context captured by these dimensions.

consequence of the positivity of these subspaces is that word-vectors with mainly low or null PMI values will be far from the centre, so in the end two word-vectors may be both close to the centre of a subspace, or at the periphery of a subspace but close to one another, or at the periphery and far from each other, at two different edges of the positively valued space, and each of these situations can be predicted to have a particular semantic interpretation. The third effect, which follows from the first two points, is that a subspace can be characterised in terms of a set of key points based on an analysis of the collective profiles of the dimensions delineating the subspace, by which I mean some straightforward assessments of the statistical distribution of each dimension involved. This aspect of my subspaces will be examined in more detail in Chapter 4.3.2; first, though, I'll consider a couple basic measures for analysing word-vectors in context.

4.3.1 Two Measures for Probing a Subspace

In order to take a first pass at examining these robustly Euclidean features of my contextualised subspaces, I propose two geometric measures for exploring the conceptual geometry of a subspace, illustrated in Figure 4.1. The first is a distance metric, which defines a central point in a subspace and then considers the relationship of words to the semantic context of the subspace in terms of the distance of the corresponding word-vectors from this central point. The central point is defined as the mean point between the input word-vectors used to generate the subspace, or, for the purposes of Figure 4.1a, the central point of the eight word-vectors being analysed in this context. In this sub-

space featuring two hand picked co-occurrence dimensions selected from a base model built from a 5x5 word co-occurrence window traversal of Wikipedia, word-vectors relatively closely associated with the concept PREDATORY ANIMAL turn up near this central point.[3] So, for instance, cats (certainly in their taxonomical sense), more specifically lions, dogs, and, again more specifically, wolves all fall close to the central point, while sharks (certainly predators, and also animals, but perhaps less prototypically so), mice, humans, and lenders are more distant.

The second measure deployed here will be to analyse the norms of the word-vectors projected into the contextualised subspace, with my hypothesis being that word-vectors that are relatively far from the origin will be correspondingly relevant to the conceptual context from which the subspace has been generated. This prediction does not entirely play out in the subspace depicted in Figure 4.1b, where words like *human* and *lender* are about as far from the origin as *cat* and *shark*, and have higher norms than more prototypical denotations such as *lion* and *wolf*. As will be seen in subsequent results, beginning here and extending into the experiments described in the next chapter, in higher dimensional subspaces selected using the techniques outlined above, norm does prove to be a predictive measure of semantic relevance. Here again, the preponderance of co-occurrence statistics associated with a word over the course of a set of dimensions gives a higher dimensional subspace an advantage: if the selected dimensions are appropriately aligned, there will be a tendency for those word-vectors with some consistency of co-occurrence across all dimensions to extend towards the central fringe of the space, while those with inconsistent co-occurrence profiles will move towards the edges while remaining closer to the origin.

In the cases of both the distance from mean and norm measures, a threshold could, in principle, be established in order to determine a cut-off point for conceptual membership, either in terms of an absolute geometric measure – a radius from either the central point or the origin – or in terms of a set of nearest neighbours. This move would begin to move these subspaces towards Gärdenfors's (2000) notion of a region within a conceptual space, particularly in the case of the distance based metric illustrated in Figure 4.1a: here a clear sense of convexity as a criterion for a conceptual region exists, and likewise of betweeness as an indicator of conceptual inclusion. Importantly, though, these spaces as they stand lack the dimensional interpretability that characterises Gärdenfors's spaces, in that it is not possible to say that there is a dimension of size, or strength, or ferocity, or so forth along which a boundary for inclusion in the concept of PREDATORY ANIMAL

[3]Here it happens to be the case that choosing dimensions which actually nominate a concept likewise delineate a space where, at least in terms of the restricted vocabulary evoked in Figure 4.1, conceptual membership plays out in a geometrically predictable way, but I will not generally presume this to be the case.

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

Table 4-C: The top word-vectors in subspaces selected by input terms characteristic of WILD ANIMALS, for the JOINT and INDY dimension selection techniques, measured in terms of top norms within each subspace (*norm*), word-vectors closest to the mean point between the input word-vectors (*distance*), and also the smallest angle with this mean vector regardless of actual position in the subspace (*angle*).

can be identified.

Examples of the tendencies of both norms and relative distances are explored in Table 4-C and Table 4-D, where, as with the examples offered earlier in this chapter, input terms denoting things exemplary of the respective concepts WILD ANIMALS and PETS are used to generate subspaces, in this case using both the JOINT and INDY dimension selection techniques, once again using a base space built using a 5x5 word co-occurrence window. In these cases, the top 200 dimensions derived using each technique have been used to project subspaces, and then within those subspaces, the top ten word-vectors based on their norm and their distance from the mean point between the input word-vectors are reported. In addition to the two geometric measures described above, as a point of comparison, I also present results using an angular measure, where the word-vectors with the highest cosine similarity with the vector of the mean point between the input word-vectors are returned. This is offered as an approximation of what would be a typical approach in a standard static distributional model, to demonstrate why this measure doesn't work for the context sensitive spaces built using my methodology and also as a mechanism for further exploration of what's happening in these subspaces.

Notably, in the case of the norm measure, word-vectors that are exemplary of the conceptual category suggested by the intersection of the input terms seem to rise to the top of the subspace, so to speak: for both dimension selection techniques for the WILD ANIMAL type inputs, a list of wild animals, some rather exotic, are returned. A similar

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

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

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

The distance measure returns roughly similar results, including a number of denotations of appropriate animals. Here it is interesting to observe that other semantic types – in particular, adjectives in addition to nouns – begin to creep into the output: *wild*, *giant*, and *golden* are returned in the JOINT and INDY subspaces for the WILD ANIMAL input, and *giat* again comes up in response to the PETS input, along with, perplexingly, the verb *include*. It makes sense that the region near the mean point between the input vectors, where consistently high but perhaps not absolutely maximal PMI scores across these contextually characteristic dimensions are to be found, feature some of the descriptors and predicates associated with the concept being modelled, while the region at the

outer fringe of the space, where the words with the highest overall PMI values across the dimensions of the subspace, would be pointed denotations of instances of the concepts in question. The word-vectors corresponding to some of the more esoteric animals in particular are likely to have high co-occurrence frequencies with the same dimensions selected by the combination of the input terms relative to low independent frequencies precisely because of their rareness.

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

4.3.2 Replete Geometric Analysis

I will now propose a general method for a replete geometric analysis of a contextually projected subspace, based on the position of word-vectors in a space as well as the relationship between those word-vectors and points based on a more general analysis of the dimensions delineating the subspace which I will characterise as *generic points*. For the purposes of explicating this method, I will presume a subspace projected from an analysis of two input word-vectors A and B using one of the dimension selection techniques described earlier in this chapter, a presumption in line with the experiments to be described in Chapters **??** and 6. The premise is that these word vectors are to be analysed in terms of their semantic relationship; the precise nature of the relationship being

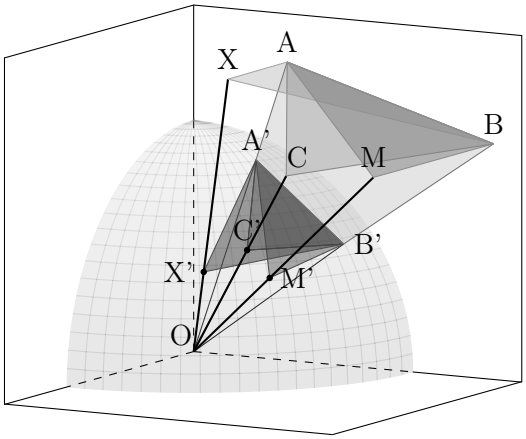


Figure 4.2: The geometric features of a subspace contextually projected based on an analysis of two input word-vectors.

analysed could be more or less anything, and in the next two chapters this method will be applied to the assessment of lexical similarity, relatedness, metaphor, and metonymy. The objective of this analytic method will be first to test the hypothesis that the geometry of contextually projected subspaces should be semantically informative, and second to compare the aspects of the geometry that are most informative for different semantic phenomena.

Figure 4.2 illustrates a generic three dimensional subspace, with point O as the origin. Points A and B are the two word-vectors that have been used to select the dimensions which define this subspace, and are likewise the word-vectors which will be analysed through the geometry of the subspace. In addition to these two points explicitly defined in terms of the values of projected word-vectors, two points are established based on an overall analysis of the dimensionality of the subspace: the *mean point* M and the *maximal point* X. M is defined as the vector of all the mean values for all the dimensions J delineating the subspace, so, if the dimensionality of J is d, M can be defined formally as follows:

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

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

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

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

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

Table 4-E: Dimensional profiles in terms of mean and maximum PMI values along dimensions, including mean values and standard deviation as well as the top five and bottom five dimensions for each statistic.

defined. The universal value chosen to define the dimensions of this vector is the mean value of the mean point M, so, formally, this point is the vector of that mean value repeated d times:

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

In the analysis of the semantic relationship between A and B in a given projection, these three vectors will be used as anchor points to establish the situation of A and B relative to the subspace overall: where C is an objectively central point in the subspace, M is in a sense central to a subspace relative to its particular dimensional constitution, and X is similarly indicative of the outermost possible extent of a particular subspace. The underlying intuition here is that, due to the frequentist components of the information theoretic co-occurrence statistics used to build the base space, different dimensions have different distributional profiles. To demonstrate this point, Table 4-E presents the mean values and standard deviations for the distribution of mean and maximum points from the top 20,000[4] most frequent co-occurrence dimensions, as well as the top five and bottom five values for each of these statistics for illustrative purposes.

The co-occurrence dimensions that tend to have lower mean and maximum values are clearly quite frequent words, and this is to be expected, given that the high frequency of

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

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

In addition to the situation of the points A, B, C, M, and X in a subspace, a normalised version of the subspace is considered, in which each vector is effectively measured at its intersection with a hypersphere of radius 1 emanating from the origin. These points are represented as A', B', C', M', and X' respectively in Figure 4.2. The purpose of considering these points is to take measure of the way in which the various vectors in a given subspace relate to the subspace as a whole, regardless of the extent of these vectors. So, for instance, the vectors A and B might have very different norms, but the distances A', B', C', M', and X' might still be very small—and, even then, the angle $\angle A'M'B'$ might be very large, suggesting that A and B both pass through the central region of the subspace but on different sides of the generic central point of the subspace. One of the objectives of this analytical method is to test whether this kind of information, which can be captured through a robust geometric description of a subspace, is semantically indicative.

So finally the various geometric features available for the analysis of a subspace are systematically outlined in Table 5-H. The points to be found in the space are broken down into three types, namely, the word-vectors themselves (points A and B), the generic points that emerge from an analysis of a subspace (points C, M, and X), and the normalised versions of all these points (A', B', C' M', and X'). The relationships between these points are construed across five categories as follows:

Distances Euclidean distances, such as the distance between the two word-vectors A

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

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

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

and B as well as the norms of the generic points, and, additionally, the mean distance of A and B from the origin;

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

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

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

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

These features have been selected as indicative of the overall comportment of the subspaces from which they are extracted, and, both independently and in conjunction, are expected to serve as indicators of the semantic phenomena characteristic of the word-vectors used to generate the subspace into which they are projected. So, for instance, I

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

As a more general hypothesis, I surmise that different sets of geometric features will collectively be predictive of different semantic phenomena. One of the primary objectives of the empirical work described in the next two chapters will be to establish a methodology for mapping features to phenomena and then using these correspondences as a mechanism for understanding the statistical characteristics that allow for the computational extraction of semantically and contextually useful information from large scale corpora. It will therefore ultimately be the comparison of the groupings of features corresponding to specific semantic phenomena that will provide the most significant outputs of the research reported here, and so the arrangement of features in terms of types and categories as outlined in Table 5-H is in this regard a schematic for the computational experimentation and corresponding evaluation and analysis at the core of this thesis.

4.4 A Mathematical Justification for Geometric Analysis

The application of geometry as a productive analytical tool for extrapolating semantic information from contextualised co-occurrence statistics has been, thus far, presented as a somewhat intuitive decision. There is a certain elegance to using quantifiable distances and angles as the analytical representation of choice, and this approach will, it will be seen, assist in the visualisation of what's happening statistically in the subspaces produced by my model. Notwithstanding these benefits, this section will offer a more mathematically thorough explanation of why a geometric approach is the right one for the types of statistics that are being used here, and in probabilistic models in general.

In order to understand the usefulness of geometry, it is worthwhile to consider again the information theoretical nature of the statistics being used here, and more generally in a plethora of distributional semantic models. Specifically, revisiting and restating Equation 4.1, the scalars of the base model are defined by considering a ratio of frequencies approximately equivalent to a ratio of probabilities:

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

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

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

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

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

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

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

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

Furthermore, when we consider the various geometric features of a contextualised subspace as the independent variables of a model designed to classify or quantify a semantic phenomenon, we are in fact looking for weighted linear combinations of operations on conditional probabilities that maximise the correlation between those statistics and a set of dependent variables generally based on human observations. In Chapter **??**, for instance, a linear regression will be used to try to learn to predict human ratings of relatedness and similarity based on geometric features of subspaces, and in Chapter 6 a logistic regression will be used to similarly classify binary judgements of metaphoricity. At this point, the geometry of the subspaces generated by my methodology becomes not only a convenient mechanism for humans to use to visualise the relationships between various statistical spaces, but actually also a handle for an algorithm to selectively learn rather complex combinations of probabilistic features. A machine learning approach to analysing the geometry of a contextualised subspace then becomes a mechanism for iterating through inferential expressions formulated as operations on conditional probabilities, and an effective model will extrapolate an interpretable treatment of these probabilities directly from the geometry of a subspace.

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

4.5 Comparing to Alternative Approaches

In order to evaluate the effectiveness of my methodology, it will naturally be necessary to compare the performance of the models I develop against other models. One way of doing this will, of course, be to compare to results other researchers have obtained experimenting with the data which will serve as the foundation for the results reported in the next two chapters. In the cases of results reported by other researchers, though, similar but variously different corpora have been used to train other models described in the literature. This is to be expected, and the results for large scale corpora should be fairly generalisable assuming a sensible choice of data (and the use of Wikipedia as all or a large portion of base data is quite common in the field), but nonetheless it will be useful to establish a baseline of results generated using models trained on the exact corpus to which I apply my methodology. And in the cases of metaphor and semantic type coercion in particular, which will be examined in Chapter 6, the datasets explored are relatively new and have not been approached by many researchers in the field, so any additional point of comparison will be valuable in evaluating my methodology.

Moreover, in most cases, other models have been designed in a task specific way: so, for instance, Schwartz et al. (2015) have developed a syntactic heuristic for identifying semantic similarity as compared to relatedness in particular, and Gutiérrez et al. (2016) describe a model that generates compositional adjective-noun representations geared towards metaphor detection. One of the key features of my models is that they are intended to be *general*: the geometries generated by my methodology are expected to be replete with semantic interpretability, allowing for the same potential for diverse and often surprising conceptualisation corresponding to the infinitely combinatory characteristic of natural language in use. For this reason, it is desirable to have a base case of a generic model that can be compared across the board to all the different tasks handled by my methodology.

With all this in mind, I propose two different points of comparison that, in addition to results extracted from existing literature, will be applicable to all subsequent experiments described here. The first involves factorising my base space using singular value decomposition (SVD), abstracting the space into a smaller set of abstract dimensions representing axes of maximum variance between PMI values. The second is an application of a well known and highly productive neural network model to the same underlying data that I've used. This will serve as a mechanism for comparing my results to what has proved to be another very effective methodology for the statistical modelling of semantics in general.

4.5.1 Static Interpretations of the Base Space

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

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

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

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

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

4.5.2 A Model Trained Using a Neural Network

In addition to the interpretations of the statistical base space described above, the neural network based models outlined by Mikolov et al. (2013a) under the rubric `word2vec` will be used as a point of comparison. These models have received a remarkable degree of attention in the NLP literature since their introduction a few years ago, so much so that the software was mentioned by name in 116 out of the 230 long papers published in the

2016 Proceedings of the Meeting for the Association for Computational Linguistics (Erk and Smith, 2016). The models have been taken, sometimes in modified form, as a source for representations of words *embedded* in vector spaces trained on large scale textual data, applied to tasks ranging from word relatedness and similarity ratings (Kiela et al., 2015) to analogy completion (Mikolov et al., 2013c), and have also been applied to multimodal tasks such as image labelling (Kottur et al., 2016).

The `word2vec` framework includes two different neural network architectures for generating word-vector representations based on traversals of large scale corpora. The *contextual bag of words* (CBOW) technique treats the terms in a co-occurrence window surrounding a target word w as input and attempts to learn a representative word-vector $\overrightarrow{w}$ that is predicted by processing the input word-vectors through a recursive neural network. The *skip-gram* technique, on the other hand, treats the representation $\overrightarrow{w}$ itself as input to a network which learns to predict word-vectors representing words on either side of the target word. In both cases, the model updates the scalars of the target word vectors in order to move them closer to the vectors representing each co-occurrence in which they're observed through backpropagation. In the case of the CBOW model, the terms co-occurring within a given window of the target word are combined into an average vector for the purpose of each training observation; with the skip-gram model, the selection of target output word-vectors is weighted based on their distance from the input word-vectors, and the model optimises the probability of two word vectors interpreted via the softmax function (see Mikolov et al., 2013b, for more details).

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

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

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

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

4.6 A Proof of Concept

In this section, I present a preliminary experiment performed using my contextually dynamic distributional semantic model. This experiment, conceived as a proof of concept, involves using multi-word phrases as input and evaluating my methodology's capacity for building subspaces where words associated with the conceptual category denoted by the input term can be reliably discovered. The experiment expands upon the notion of proto-conceptual spaces outlined in Section 4.3.1, examining whether the word-vectors that populate regions of subspaces are characterised by a certain categorical coherence. In the case of the data explored here, the experiment is specifically set up to feel out the contextual capacity of my methodology and compare it to a standard generic semantic space. The question asked is whether the shifts from subspace to subspace based on particular input yield productive alterations in the way that words both cluster and emerge from the melange of word-vectors that circulate around my base model.

The gist of this experiment is to take a word pair representing a compound noun – for instance, *body part* – and see if my methodology can use the word pair to contextually

generate a space where other words conceptually related to that compound noun can be found in a systematic way. This is conceived of as an entailment task, in that I will attempt to find phrases considered to be categorical constituents of the concept represented by the word pair, taking the WordNet lexical taxonomy as a ground truth. There is a scholastic back story here.

An early version of this experiment was reported in Agres et al. (2015). That first effort arose out of a question posed by a colleague regarding the feasibility of using a statical NLP technique for generating categorical labels that could be used to evaluate computational creativity in a domain specific way (for a psychological perspective on the difficulty of generating such terms in an objective way using human subjects, see van der Velde et al., 2015). So, for instance, given a creative domain such as MUSICAL CREATIVITY, could a distributional semantic model generate terms that are reliably relevant to the concept denoted by that phrase, rather than the potentially disparate properties independently associated with MUSIC and CREATIVITY? Intuitively there seems to be little reason to hope that the space halfway between these points in a general semantic space would somehow adequately represent the properties of the overall concept. The early work explored the dimensions contextually selected by analysing the co-occurrence features of word-vectors corresponding to inputs along the lines of the expository results presented anecdotally in Chapter 4, but without any rigorous evaluation.

Reviewer responses to a subsequent journal article (McGregor et al., 2015), designed as a more thorough introduction of the methodology, inspired a computationally oriented mode of evaluation. The experiment that has emerged involves attempting to recapitulate taxonomical conceptual relationships from the WordNet database (Fellbaum, 1998). Wordnet is a lexical taxonomy of *synsets*, basically semantic word senses, arranged into a hierarchy of entailment relationships, with each synset associate with a number of *lemmas*, word types indexed by that synset according to human annotators. There is precedent for the construction of *ad hoc* datasets from WordNet, with for instance Baroni et al. (2012), Riedl and Biemann (2013), and Melamud et al. (2014) all mining the extensive lexical taxonomy for gold standard entailment relationships. My experiment takes as input instances of synsets labelled by compound noun phrases and seeks to output as many of the lemmas listed associated with synsets that are hyponyms of the input synset. So, for instance, the synset body part has a hyponym EXTERNAL BODY PART, which has a hyponym EXTREMITY, which has a synset LIMB, which has a synset LEG associated with the lemma *leg*, and so *leg* would be considered a positive output for the input *body part*.[6]

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

4.6.1 Experimental Set-Up

12 of the top synset labels consisting of compound noun phrases are extracted from WordNet. These labels are extracted through a breadth first traversal of the tree of noun synsets, selecting the highest 12 synsets with multi-word labels with the constraint that none of the 12 can be parent nodes of any of the others: in this way, 12 distinct, non-overlapping conceptual categories are choosen. The experimental vocabulary is considered to be the intersection of the list of all WordNet noun lemmas associated with the vocabulary of my model (the 200,000 most frequent word types in Wikipedia), resulting in a total vocabulary of 32,155 words. The lemmas associated with all the hyponyms of each synset are extracted and grouped, and these words become the target words for my models' output. The 12 synset labels are itemised in Table 4-H.

With the target output established, the terms labelling a given synset are passed to my model as contextual input, with the corresponding word-vectors serving as the basis for dimensional selection using the JOINT, INDY, and ZIPPED techniques as outlined in Chapter 4. Here, the base space generated using a 5x5 word co-occurrence window is used, and 200 dimensional subspaces are returned; variations of these parameters will be tested in subsequent experiments. The subspaces returned by each of these techniques are explored to return the top terms using both of the procedures outlined in Chapter 4.3.1: the terms closes to the mean point between the input word-vectors in a subspace are returned, and the terms furthest from the origin – the terms with the largest norm – in a given subspace are returned. The top 50 terms found in a subspace each according to each measure are returned, as well as the top terms up to a limit n where n is the total number of lemmas associated with the target multi-word label. Accuracy scores for each of these sets of output are computed, so the total number of positive matches for hyponyms of the input synset out of the top 50 and top n terms returned.

As a point of comparison, results are likewise returned from two different `word2vec` models, one using the skip-gram methodology and one using the bag-of-words methodology, as described in Chapter 4.5.2. In line with the subspaces generated using my methodology, 200 dimensional models are used, and these models are built across 10 iterations of the corpus, using a 5x5 word co-occurrence window, applying a negative sampling rate of 10 and an initial learning rate of 0.025, as discussed in Chapter 4.5.2. Here the top terms in terms of proximity by cosine similarity to the mean point between the word-vectors associated with the input terms are returned, again taking the top 50 and top n for each input.

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

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

4.6.2 Results and Analysis

Results for the set-up described in the previous section can be found in Table 4-G, with both the average accuracy scores and the average ratio of model accuracy to baseline reported. Results for both the norm and distance from mean point methods are reported for subspaces derived using the JOINT, INDY, and ZIPPED dimension selection techniques, followed by results for the skip-gram and bag-of-words `word2vec` techniques. The first thing to note about these results is that all of the results are substantially above the baseline: the average ratios of model accuracy to the baseline (the likely accuracy achieved by randomly choosing words from the vocabulary for each input) are all above 2.5, and are above 3.2 for all of my methodologies. So it is clear that all these techniques are generating semantically significant relationships between word-vectors.

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

More surprising is the distinction between the strong performance of the norm measures and the less impressive performance of the mean point measure. In the case of accuracy among the top 50 terms returned by each model, my methodologies results using this Euclidean measure consistently fall short of the `word2vec` techniques. It would seem, then, that in the subspaces returned by my models, proximity to the input word-

vectors is not in itself an indicator of categorical inclusion in the conceptual space traced by the intersection of the correspond contextual input terms. Upon further consideration, there is a plausible explanation for this: revisiting the outputs for subspaces projected using denotations of animals as input, reported last chapter in Tables **??** and **??**, the norm measure produced specialised terms such as *chital* and *poodle*, while the distance measure generated relevant but not always categorical terms such as *wild*, *giant*, and *golden*. To give an example from the data used for this experiment, top-50 results from the JOINT distance measure returned for the input (*body, part*) include words like *portion*, *upper*, *shape*, and *whole*, while the results from PHYSICAL PROCESS include *method*, *complex*, and *affect*—so, terms that are conceptually relevant to the target domain but are not strictly part of the category BODY PART. We might characterise this trend in terms of a distinction between words which denote semantic *relatedness* versus *similarity*, a topic which will be addressed in depth in the next section.

Focusing on the accuracy of the results returned by the models up to the full length of each target set of lemmas, here results are weaker all around, which is not particularly surprising: as we move away from the regions where we expected to see the highest degree of conceptual consistency, mismatched terms begin to creep into the results. It is notable, though, that my methodologies outperform the neural network based models across the board, especially for the norm based measures but also in the case of this larger sample of the respective semantic spaces for the distance based measures. In fact, the stronger relative performance for the distance measure in these expanded regions of each type of subspace makes sense, since, as the norms measure moves closer to the origin in search of output and the distance measure likewise expands from the locus of its mean point, the results output by each measure will increasingly overlap (an overlaying of Figures 4.1a and **??** will illustrate this phenomenon). But the main point to take here is that, in the case of my methodologies, there is clearly a more persistent conceptual organisation to the space. As we expand from any point in the static type of semantic model generated by `word2vec`, we will undoubtedly begin to encounter the vagary and the messiness inherent in language and problematic for fixed lexical relationships. My methodologies, on the other hand, afford the *ad hoc* construction of semantic spaces which afford the situational corralling of the looseness and ambiguity inherent in a dynamic lexicon.

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

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

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

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

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

distinctions between similarity and relatedness.

Chapter 5

Relatedness and Similarity

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

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

A fundamental objective for a general semantic model is a mechanism for measuring

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

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

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

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

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

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

5.1 An Experiment on Relatedness

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

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

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

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

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

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

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

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

5.1.1 Relatedness: Methodology and Model

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

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

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

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

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

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

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

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

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

5.1.2 The Geometry of Relatedness

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

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

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

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

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

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

On a dimension-by-dimension level, similar PMI values, or at least similar ratios of values, between word-vectors relative to both the mean values for each dimension and

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

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

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

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

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

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

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

5.2 An Experiment on Similarity

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

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

In this section, I apply my own methodology to the SimLex data in order to investigate the extent to which context specific subspaces of word-vectors can accurately represent the similarity between words. As with the previous experiment exploring word relatedness, a primary objective here is to test the extent to which the geometric features of my subspaces both collectively and independently align with human ratings. In addition

to performing a linear regression mapping the full sets of geometric features generated for various combinations of parameters and likewise comparing the correlation between individual features and human similarity ratings, here I will also attempt to extract a set of features which optimally predict similarity while avoiding collinearity and without overfitting the resultant model. This approach will offer a mechanism for interpreting the dynamics at play between different features of contextualised statistical subspaces.

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

5.2.1 Similarity: Methodology and Model

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

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

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

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

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

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

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

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

5.2.2 The Geometry of Similarity

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.3 Comparing the Two Phenomena

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

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

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

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

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

The really interesting thing to note here, though, is that, outside of angular measures, the two different semantic relationships tend to be associated with different sets of geometric features. Relatedness is strongly associated with ratio type features, with the negative correlation with $\overline{A'C'} : \overline{B'C'}$ indicating that one related word tends to be significantly closer to the centre line than the other in INDY subspaces (this is also supported

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

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

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

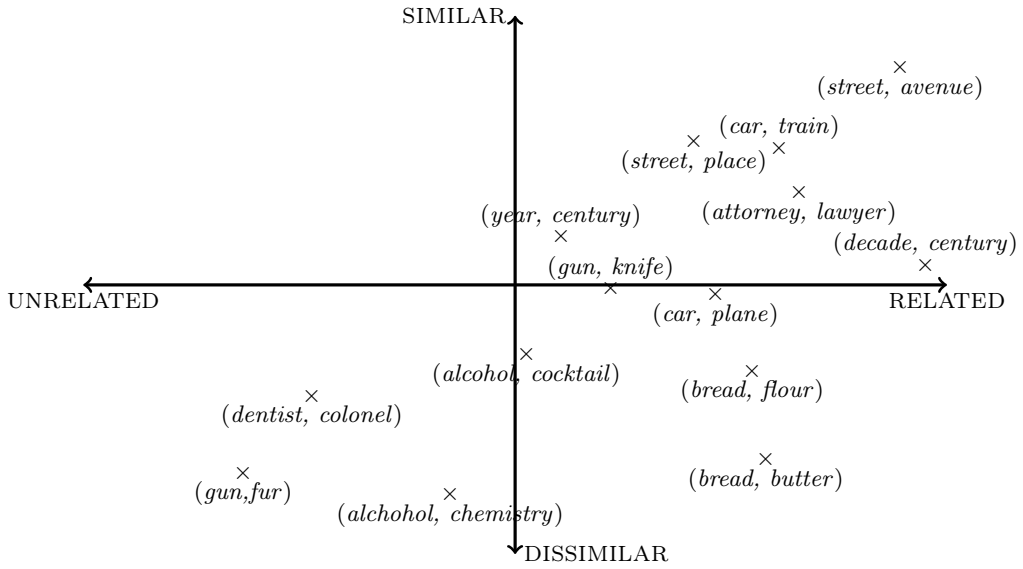


Figure 5.1: Noun pair scores along axes of relatedness and similarity as returned by a model built from features of 2x2 word co-occurrence window, 400 dimensional, INDY type subspaces.

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

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

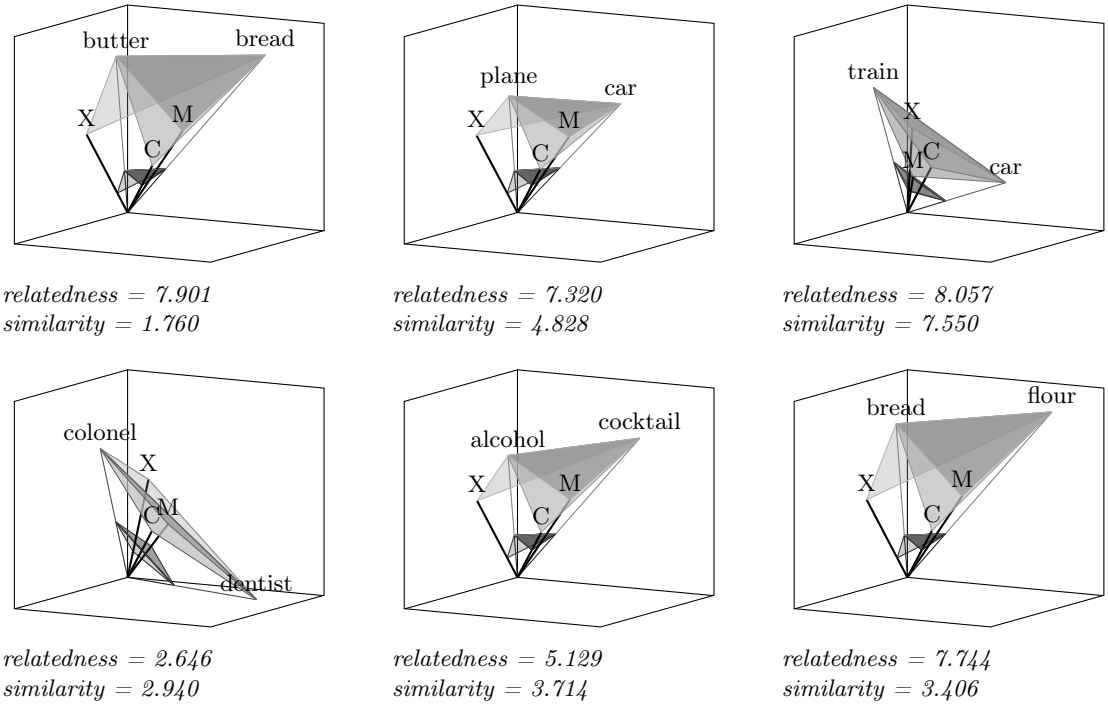
[7]Levy et al. (2015b) have similarly proposed that success in distributional semantic models capturing entailment relationships is in fact down to their ability to identify *prototypical hypernyms* that are simply more likely to be identified as categorically containing some other unseen word—but those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

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

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

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

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



relatedness = 7.901
similarity = 1.760

relatedness = 7.320
similarity = 4.828

relatedness = 8.057
similarity = 7.550

relatedness = 2.646
similarity = 2.940

relatedness = 5.129
similarity = 3.714

relatedness = 7.744
similarity = 3.406

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

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

5.4 Frames of Similarity

Tversky (1977), in his psychologically motivated reflections on the geometry of similarity, observes that relationships of similarity are fundamentally not symmetric: there tends to be a preference to consider the specific more similar to the general, and the peripheral more similar to the prototypical, than the other way around. So, to use Tversky's own example, *ellipse* is more similar to *circle* than *circle* is to *ellipse*; we might extend this conjecture to predict that *wolf* is more similar to *dog*, *radiologist* more similar to *doctor*, and *limping* more similar to *walking* than the converse propositions. Indeed, the frequentist axiom extrapolated through the geometric analysis of the previous section, stating that more common words denote things that are more likely to generally be a component of a similarity relationship, is broadly in line with this observation. Tversky makes the

point that the conventional conditions of geometric relationships – *minimality*, *symmetry*, and *the triangle inequality* – do not pertain in the case of similarity judgements, a point which if taken seriously serves to foil the project of a vector space model of word similarity.

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

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

These types of conceptual contexts are evident in the relatedness and similarity datasets which have been explored in this chapter. In the SimLex data, for instance, (*dentist*, *colonel*) is rated as one of the least similar word pairs at 0.40, while (*attorney*, *lawyer*) is, at 9.25, considered one of the most similar pairs. The difference seems reasonable enough in terms of a comparison between the two pairs, but the low rating of (*dentist*, *colonel*) leaves little room for either dentists or colonels to be even less similar to, say, gorillas, or electricity, or democracy, and so forth. What seems to be happening here

is that human evaluators are identifying an implicit conceptual frame in which each word pair is to be evaluated: in the case of attorneys, lawyers, dentists, and colonels, the frame is something like PROFESSION, and so the professional activities of colonels and dentists are judged to be more or less orthogonal, while attorneys and lawyers pursue very similar careers. The inclusion of some additional comparison, for instance (*dentist*, *grandparent*), would suggest a broadening of the conceptual frame to something like HUMAN, and a corresponding drawing together of words denoting professions in particular.

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

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

Faruqui et al. (2016) raise a number of issues with relatedness and similarity datasets, among them the uncertainty surrounding specific semantic phenomena and the lack of applicability of quantified word pair scores to practical NLP tasks. Those authors ultimately propose that quantitative evaluations of vector space models of word meaning

should avoid claims of generality, instead treating particular models as task specific implementations. There is something to be said for this approach, and even more to be said in support of the effort to apply statistical NLP techniques to activities in other fields where heterogeneous data and contextual complexity present potentially confounding factors to the relatively abstract and rigid representational structures of distributional semantic models. All the same, I maintain that word association tasks, particular a battery of tasks spanning a variety of semantic phenomena, can be a productive tool for exploring the capabilities of a methodology, and present the work that has been described in this chapter as a case in point.

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

The INDY dimensional selection technique in particular would lend itself to this type of programmatic extension of research into semantic relatedness, as it facilitates the open-ended concatenation of dimensions from an analysis of an arbitrarily large set of constituent word-vectors (the JOINT and INDY techniques, on the other hand, would presumably return increasingly uninteresting dimensions with universally non-zero values as the set of input words expands). A subspace built using the INDY technique based on an analysis of a set of words denoting, for instance, constituents of the concept PROFESSIONALS would acquire co-occurrence dimensions specifically salient to each of the input terms, and the construal of other word-vectors in the space along the collective profile

of dimensions would, I forecast, be indicative of their conceptual situation according to the various properties of being a professional. In such a space, we might predict that we would find, for instance, *surgeon* somewhere in the vicinity of the region between *barber* and *butcher*

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

Chapter 6

Metaphor and Coercion

In this chapter, I will extend the empirical work on exploring the application of my context sensitive distributional semantic models to two semantic phenomena which involve the application of words in situations where their meanings are in some sense conceptually altered: *metaphor* and *semantic type coercion*. The connotation of these terms will be explored throughout the course of this chapter, eventually arriving at a proposal for how to frame the idea of figurative language through a computational analysis. As an overview, the distinguishing characteristic of these phenomena as they are conventionally understood is that they involve cases where what might be thought of as the stable, encyclopedic understanding of some word sense – a *dictionary definition* of a word, so to speak – is in some way appropriated or subverted in order to, among other things, transfer information via the attributional conduits connecting figurative source to literal target.

My hypothesis is that, because figurative language always involves the contextual specification of word meaning, context sensitive geometries of lexical representations should provide an appropriate framework for identifying when this type of semantic phenomenon is in effect. Fraser (1993) demonstrates empirically that metaphor interpretation is, when a metaphor is presented to a subject out of context, an ambiguous exercise, and, to the extent that interpretations of de-contextualised metaphors can be predicted, the predicting factors are themselves culturally relative. Along similar lines, Bouveret and Sweetser (2009) propose that metaphor production involves the contextual alignment of overlapping semantic frames, and that this alignment likewise imports structure associated with one frame into the domain of another, evident in, for instance, the additional transposition of syntactic constraints from source to target. From a cognitive perspective, this coordinates a contextual theory of metaphor with the work on

conceptual frames from Barsalou (1992) discussed at the end of the previous chapter in the context of judgements of semantic similarity. From a modelling perspective, this suggests that a methodology for projecting semantic spaces where context specific perspectives can reveal *ad hoc* perspectives on semantic relationships should be a productive approach to identifying figurative language.

The idea that metaphor and metonymy are both instances of "a connection between two things where one term is substituted for another," (Gibbs Jr., 1993, p. 260) will quickly call to mind the premise of distributional semantics: if the motivation for building vector space models of word co-occurrence statistics is that related words have similar co-occurrence tendencies, then figurative language might be construed as a special case in which unrelated or at least conceptually divergent words are likewise found in similar sentential situations. The question, then, is whether statistical characteristics of the particular co-occurrences profiles selected by words with different meanings are predictive of figurativeness. A naive hypothesis might be that word combinations that are figurative should simply be further apart in a semantic space than word combination that are literal. If related words have similar co-occurrence profiles, then maybe unrelated words, for instance words with different conceptual entailments, should have less similar co-occurrence profiles. This conjecture, however, is belied first of all by the fact that, in the type of corpus containing a broad range of examples of language use necessary for building distributional semantic models, figurative language will already be built into the data (and, as Gibbs (1994) has pointed out, figurative language is going to built into any sample of language no matter how small or basic). A second problem is that, specifically to overcome the problems with modelling semantic relationships merely in terms of collocations, distributional semantics compares the co-occurrence profiles of words rather than their direct relationships, and it seems likely that word combinations prone to metaphoric interpretation might very well have at least overlapping profiles.

So the objective of the experiments reported in this chapter will be to explore the ways in which and the degrees to which a more fleshed out statistical description of contextually selected distributional semantic subspaces can reveal figurative language. As with the experiments on relatedness and similarity reported in the previous chapter, in addition to the relationship between target word-vectors in the subspaces they select, the statistical properties of the selected dimensions themselves will also be examined. And, again as with previous results, the instrument of analysis will be the geometric features of the subspaces in question, with, again, particular attention paid to the way in which the sets of features can collectively indicate figurative language. The two primary datasets explored represent binary decisions about metaphoricity and coercion respectively, and so my models will be applied to classification tasks here. In the case of metaphor, I test

whether a model learned based on classification data is generalisable to graduated human ratings of metaphoricity. With the coercion data, I will examine whether the addition of information about sentential context enhances the classification of word pairs. I will conclude the chapter with a reflection on some of the theoretical implications of the strongly positive results described here.

6.1 An Experiment on Metaphor

As pointed out by Shutova et al. (2012), statistical approaches to metaphor identification and interpretation have generally been formulated in the context of the *conceptual metaphor* theory of Lakoff and Johnson (2003). This model is founded on the principle that "we systematically use inference patterns from one conceptual domain to reason about another conceptual domain," (ibid, p. 246). Metaphors are then the mechanism for performing the mapping between these domains, and as such cut right to the core of cognitive processes. Statistical models of metaphor have accordingly treated metaphors as transformations of lexical representations, and vector space models of distributional semantics have naturally leant themselves to this type of approach. The construction of representations with the potential to interact with one another in semantically productive ways has in turn lent itself to the development of models that consider the compositional nature of metaphor, effectively treating the metaphor itself as a transformation of the underlying representations. So Utsumi (2011) constructs candidate metaphor-vectors by calculating the centroid of a number of vectors derived from an analysis of a noun-vector and a predicate-vector learned through latent semantic analysis, and then uses the spatial relationships between these composed vectors to analyse the metaphoricity of certain phrases. Hovy et al. (2013) similarly consider composition in their approach to metaphor classification, in this case by combining word-vector type representations with a model trained to identify metaphor based on dependency trees of sentences labelled for metaphoricity.

In the tradition of work on compositional distributional semantics explored by the likes of Mitchell and Lapata (2010), Baroni and Zamparelli (2010), and Coecke et al. (2011), among others, semantic types such as adjectives and verbs are modelled as tensors which perform transformations on nouns, which are modelled as vectors. In the normal run of things, compositional models therefore represent, for instance, noun phrases modified by adjectives as the product $A\overrightarrow{n}$, where $\overrightarrow{x}$ in a noun vector and A is a matrix representing an adjective learned from observations of attested instances of the adjective with other noun word-vectors. So the phrase *black dog* becomes a word-vector in the same space as the representation of just *dog*, and can be compared quantitatively and geometrically with

other phrases such as *white dog* or *big cat* and so forth. In the case of metaphor, these transformations are expected to map the word-vector representing metaphoric phrases into a region corresponding to the semantic domain of the original noun-vector modified by a metaphoric interpretation of the word associated with the tensor of a modifier or a predicate. So, for instance, in a model that effectively captures metaphoricity, the composition of the vector space representations corresponding to *brilliant light* would map to a region of space where comparisons between phrases like *dark illumination* and *red glow* are productive, while *brilliant child* might be expected to map into the proximity of *stupid boy* and *boring girl*.[1]

The data that I will use in this section to test my methodology was originally presented by Gutiérrez et al. (2016), along with an accompanying experiment on a novel model. It consists of 8,592 adjective-noun pairs, spanning 23 adjectives chosen for their membership in six different broad semantic categories that are prone to both literal and metaphoric use: so, for instance, *bitter*, *sour*, and *sweet* are considered constituents of the category TASTE. There are 3,473 different noun types used, with only 141 types, represented by 640 tokens, occurring in both literal and metaphoric phrases. Each pair has been rated as either literal or metaphoric by a pair of human annotators, with inter-annotator agreement measuring at Cohen's $\kappa = 0.80$; 4,593 of the pairs have been judged metaphorical. This dataset was conceived as something of an expansion of the similar but smaller corpus of adjective-noun phrases annotated with binary metaphoricity classifications presented by Tsvetkov et al. (2014) (and those authors tested their own data with an assortment of models, achieving highest f-scores by applying a random forest classifier to the features of an existing library of distributional semantic word-vectors).

In their own experimental treatment of the data, Gutiérrez et al. constructed a pair of compositional models in the mode of Baroni and Zamparelli (2010), learning adjective matrixes A to map from noun-vectors to noun-adjective phrase-vectors extracted from observations of co-occurrences of both nouns and phrases in a corpus. By creating separate tensor representations for literal and metaphoric instances of a given adjective, the authors can then compare the relationships between the vectors resulting from a noun-vector composed with literal and metaphoric senses of an adjective-vector to try to determine whether a given phrase would generally be classified as a metaphor or a literal expression by comparing the respective compared vectors to the phrase-vector as observed in the corpus. In a further attempt to generalise the method, and, notably, to apply the conceptual metaphor theory of Lakoff and Johnson (1980) to their computational model, the authors learn matrices performing linear transformations from lit-

[1] It should be noted that such a methodology at this point begins to assume dim shades of Gärdenfors's (2000) conceptual spaces, with different compositions inherently defining different regions of the space.

eral to metaphoric adjective-noun compositions and then compare the similarity between observed phrase-vectors and literal composed vectors versus transformed literal composed vectors to determine whether a given phrase is metaphoric or not.

The data described by Gutiérrez et al. will serve as the basis for testing my own context sensitive distributional semantic methodology's ability to classify phrases as literal or metaphoric, and the results of this experiment will be described in the following section. My hypothesis is that metaphor, and indeed all figurative language, is fundamentally entangled with the context mutually indicated by the representations of the words participating in the composition being analysed. In fact, I think that part of what is captured by the model described by Gutiérrez et al., and indeed a number of other researchers investigating statistical methods for metaphor classification, is precisely that there is a context inherent in the linear algebraic dynamics of composable lexical representations, and this is something which many researchers explicitly recognise. But I also think that the explicit projection of context specific semantic subspaces, the mainstay of my methodology, should provide an ideal testing ground to discover the way in which statistical geometry can directly broadcast the presence or absence and even potentially the degree of metaphor inherent in a given phrase. The following sections will test this hypothesis using a similar methodology to that applied to semantic relatedness and similarity in the previous chapter.

6.1.1 Methodology and Results

My own methodology is clearly less committed to maintaining distinct representations for different semantic types than the compositional models described above, instead modelling all words as untagged word-vectors based on their co-occurrences as observed across a large scale corpus. This feature of my research is in part theoretically motivated: in line with Langacker (1991), and *contra* the grammatic nativism or exceptionalism that has been a mainstay in theoretical linguistics, I would like to investigate the possiblity that "grammar is fully and appropriately describable using only symbolic units, each having both semantic and phonological import," (ibid, p. 290). In other words, the syntactic component of a natural language might be described in terms of the entanglements of the meaning-making structures – the lexical semantic representations – that arise in the course of language use, or maybe even as emergent properties of these entanglements.

With this in mind, I will approach the problem of metaphor classification with a similarly statistical and geometric methodology as was applied to relatedness and similarity in the previous chapter, outside of any prima fascie model of syntax or compositionality. For every pair of words in the data produced by Gutiérrez et al. (2016), I generate sub-

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.839	0.860	0.878	0.881	0.840	0.862	0.880	0.886
INDY	0.821	0.839	0.855	0.860	0.817	0.840	0.858	0.867
ZIPPED	0.839	0.864	0.876	0.878	0.833	0.854	0.873	0.880
ADJECTIVE	0.771	0.860	0.828	0.845	0.781	0.804	0.828	0.837
NOUN	0.819	0.861	0.843	0.847	0.806	0.821	0.838	0.843
SVD	0.685	0.703	0.703	0.697	0.677	0.694	0.687	0.684
SG	0.679	0.676	0.679	0.673	0.664	0.665	0.672	0.656
CBOW	0.669	0.681	0.677	0.672	0.669	0.673	0.677	0.671

Table 6-A: F-scores for metaphor identification based on a stratified ten-fold cross-validated logistic regression taking geometric features of various subspace types as input.

spaces of 20, 50, 200, and 400 dimensions using the JOINT, INDY, and ZIPPED techniques, projected from 2x2 and 5x5 word co-occurrence window base spaces. This data specifies a distince role for each word, one being a metaphoric source (the adjective) and the other being a target (the noun): so, for instance, a *bitter loss* is a loss, but presumably not one with an actual taste, and so the noun *loss* co-opts something of the quality of bitterness into its own conceptual domain. As such, it might be useful to generate subspaces based simply on an anlysis of the word-vectors corresponding to the adjective and the noun respectively. I do this by simply selecting the top d dimensions, in line with the dimensionality parameter for each model, for the term in question, and these spaces are labelled ADJECTIVE and NOUN in the results that follow.

In each subspace, I extrapolate the same 34 geometric features described in Table 5-H and applied in the previous chapter in the semantic relatedness and similarity experiments. Again because of the semantic asymmetry of the relationship between the input terms, an additional seven features are also available in these spaces: the adjective-vector norm divided by the noun-vector norm (A/B), likewise the lengths of the vectors between the adjective and the generic points divided by the lengths for the noun-generic-point vectors ($\overline{AC}/\overline{BC}$, $\overline{AM}/\overline{BM}$, and $\overline{AX}/\overline{BX}$), and the corresponding fractions of the normalised versions of these points ($\overline{A'C'}/\overline{B'C'}$, $\overline{A'M'}/\overline{B'M'}$, and $\overline{A'X'}/\overline{B'X'}$). These additional measures might offer a sense of wehther there are statistical tendencies that are specific to the semantic role being played by a word moving from literal to metaphorical relationships, and we might expect this to be particularly evident in the spaces selected by either the noun or the adjective on their own. As with the subspaces of relatedness and similarity, I normalise each feature across all word pairs to have means of 0 and standard deviations of 1.

In order to test the capacity of the geometric features of my subspaces to identify

metaphor, I perform a stratified ten-fold cross-validated logistic regression taking these features as independent variables and learning to predict the classifications assigned to the word pairs in the dataset. Balanced f-scores based on the precision and recall of my various dimensional selection techniques as well as static SVD factorisations of my base spaces and the `word2vec` models are reported in Table 6-A. The first thing to note is the strong performance across the board of the context sensitive methodology: the model based on my strongest performing subspace (JOINT, 5x5 window, 400 dimensions) substantially outperform the strongest versions of the static models (the SVD 5x5, 50 dimension model) with $p < .005$ based on a permutation test. The context sensitive models perform better, but only marginally better, in the 5x5 word window subspaces, suggesting that most of the useful information about the semantic properties that indicate a metaphoric projection are captured by the profile of terms co-occurring in close proximity to the target words. That this trend is reversed for the static spaces, with 2x2 word window spaces doing a bit better, further indicates that the peripheral information of wider ranging co-occurrences is specifically useful for a context sensitive analysis.

The JOINT technique gives the strongest results, suggesting that subspaces delineated in terms of co-occurrence dimensions mutually salient to both input terms offer the best platform for analysing metaphoricity. This makes sense: in the case of metaphor versus literalness, it is the co-occurrences that both words have in common that position their respective word-vectors in an indicative relationship relative to one another and the subspace overall. So for instance the co-occurrences salient to both *sweet* and *fruit* will have a particular conceptual profile that will not be evident in the dimensions jointly selected by *sweet* and *revenge*; this effect will be less evident for dimensions independently salient to each word. ZIPPED subspaces, where there will be at least some information about both words along every dimension, accordingly score almost as well as JOINT subspaces, with the INDY subspaces falling further behind.

Interestingly, the ADJECTIVE and NOUN spaces classify metaphor most accurately in 50 dimensional subspaces projected from the 2x2 word window base space. To the extent that part-of-speech can be a component of the analysis of these models, we can expect the smaller co-occurrence window to produce statistics that are more indicative of a particular grammatical class. The degradation of classification at higher dimensionalities for the smaller co-occurrence window setting is a little surprising, and it's worth noting that the INDY subspaces, which are basically blends of the ADJECTIVE and NOUN subspaces, don't exhibit the same tendency. In this case, it would seem the whole really is greater than the sum of the parts, with the dimensional selection of one word providing at least a degree of useful information about the other word not available in spaces salient to a single term. A similar pattern emerges for the static spaces: the SVD, SG, and CBOW

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.815	0.837	0.854	0.855	0.816	0.837	0.858	0.863
INDY	0.778	0.793	0.828	0.835	0.774	0.805	0.829	0.842
ZIPPED	0.810	0.838	0.847	0.854	0.799	0.828	0.844	0.853
ADJECTIVE	0.606	0.709	0.750	0.777	0.698	0.697	0.757	0.707
NOUN	0.806	0.808	0.828	0.833	0.796	0.812	0.824	0.829
SVD	0.679	0.691	0.695	0.690	0.665	0.674	0.678	0.676
SG	0.668	0.664	0.659	0.657	0.659	0.656	0.644	0.638
CBOW	0.657	0.665	0.665	0.661	0.656	0.660	0.666	0.660

Table 6-B: F-scores for metaphor identification with each of the conceptual categories identified by Gutiérrez et al. (2016) treated as a separate fold for cross-validation.

models all produce the most accurate classifications in 2x2 word window, 50 dimensional subspaces. One way to explain this is that more ambiguous information about word use begins to leek in at higher dimensionalities, servining to obscure the more standard indications available in either the most salient dimensions or the dimensions containing the most information about variance across the corpus.

There is another possibility to consider regarding the adjectives in this dataset in particular: as there are only 23 different adjective types, each adjective is observed multiple times in both metaphoric and non-metaphoric contexts. It is therefore possible that, in any given fold of the cross-validation of a classifier, the model might be learning how to guess whether a specific adjective is involved in a metaphor rather than something more general about the statistical geometry of metaphoricity. In order to avoid this trap, I reorganise the data into tranches based on the adjective in each pair, I use the eight conceptual categories outlined by Gutiérrez et al. (2016) in order to structure this new partioning.[2] I use each of these eight new sets of word pairs as a fold in a cross-validated logistic regression, such that the adjective in each phrase in each test set has not been observed in the training data.

Table 6-B presents the results from this reshuffled version of the experiment. The f-scores for metaphor classification returned by the context sensitive models are down slightly, but the difference is not significant at $p = .073$. The major change here is, as expected, in the ADJECTIVE subspaces: clearly when only information from the adjective in each word-pair is used to train a model, prior observations of a specific word type in the context of some other composition is a benefit. There is also a minor decrease in performance for the static models, which is interesting in that it indicates that, even

[2]Gutiérrez et al. (2016), identifying a similar problem, likewise develop a second model that learns metaphors as mappings between domains rather than just from noun-vectors to phrase, though their methodology requires them to use a reduced version of the data.

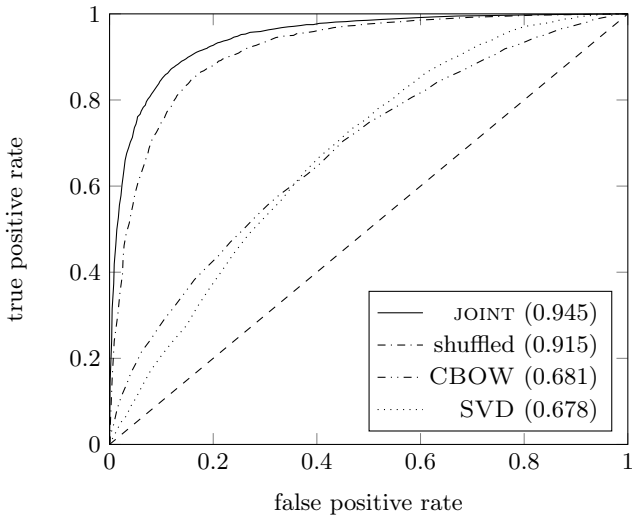


Figure 6.1: Receiver operating characteristic plots for a selection of models, with the area under the curve for each model type indicated in the legend.

when a single distance metric is used to classify metaphoricity, observations of a word in training help to subsequently test phrases involving that word. It is worth noting that of the 8,584 noun tokens spread across 3,473 noun types, 1,588 types, represented by 6,724 tokens, occur in more than one of the tranches delineating the conceptual categorisations of the adjectives, so it is possible that there is a small extent of learning to classify phrases based on previous observations of specific nouns.

In order to take a closer look at the way that different techniques model this data, and in line with the metaphor classification work of Tsvetkov et al. (2014), Figure 6.1 illustrates receiver operating characteristic curves for four versions of the approaches that have been described here: the JOINT technique with 400 dimensional, 5x5 word subspaces, the same technique applied to the version of the data shuffled to avoid training and testing on the same adjectives, and the CBOW and SVD models for the optimally performing 50 dimensional, 2x2 word window subspaces. True positive versus false positive rates are correlated at 99 increments in terms of the value of the output of a logistic regression model at which a phrase is determined to be metaphoric. The outcomes visualised here tell a similar story to Tables 6-A and 6-B, with the area under the curve statistics indicating a strong distinction between the context sensitive techniques and the static models. Perhaps the most interesting thing to note is the overall smoothness of the curves, which suggests a steady relationship between precision and recall at various classification thresholds.

With the trade off between true and false positives in mind, Table 6-C presents precision, recall, f-score, accuracy, and Cohen's kappa scores for the same models plotted

	precision	*recall*	*f-score*	*accuracy*	*kappa*
JOINT	0.879	0.894	0.886	0.877	0.753
shuffled	0.873	0.865	0.862	0.854	0.678
SVD	0.631	0.794	0.703	0.641	0.265
CBOW	0.638	0.721	0.677	0.632	0.253
Gutiérrez et al. (2016)	0.842	0.793	0.817	0.809	0.618
baseline	0.535	1.000	0.697	0.535	0.000

Table 6-C: Full classification statistics results for the models tested here as well as the results from the original literature and the majority class (metaphor) baseline.

in Figure 6.1. The trend to notice here is that context sensitive and static models tend to favour recall over precision (and the slight preference for precision in the JOINT 400 dimensional, 5x5 word subspaces for the shuffled version of the data reported here is an anomaly, as other approaches to that data exhibit the tendency towards higher recall). This evident enthusiasm for classifying phrases as metaphoric is a reflection of the data itself, which is slightly skewed towards metaphoric phrases, as described above and indicated in the performance of the majority class baseline, and this is reinforced by the relatively low accuracy scores for both context sensitive and static non-compositional distributional semantic models. It is noteworthy, then, that the model described by Gutiérrez et al. (2016) actually scores better for precision than recall, suggesting it actually tends to under-predict metaphoricity. This could perhaps be expected as a general distinction between statistical models based on unannotated data such as mine, which will arguably tend to favour a majority class, versus likewise statistical models operating on theoretically motivated mappings between representations, which have an apparent propensity for zeroing in with confidence on the properties of a compositional transformation that are indicative of metaphor—but at the expense of sometimes missing what might be considered outliers. In the same spirit, the jumpier nature of the receiver operating characteristic plots presented by Tsvetkov et al. (2014) is quite possibly an artefact of the decision points inherent in heuristically mapping model features from human made knowledge bases.

As a final point of comparison with other approaches to metaphor classification, I will return briefly to the unannotated character of my lexical representations. One of the most powerful features of the methodology described here is its ability to build a somewhat general model of a semantic phenomenon from a sufficiently comprehensive dataset, and the strong Cohen's kappa score of the best performing subspace selection technique, which begins to approach the aforementioned inter-annotator agreement level of $\kappa = 0.80$, is a testament to this. Following an analysis of the specific geometry of metaphor in the next section, Section 6.1.3 will assess the ability of my methodology to

JOINT		INDY		ZIPPED	
$\mu(A,B)$	0.787	C	0.767	$\mu(A,B)$	0.788
C	0.771	C/M	0.749	C	0.771
$\mu(A,B)/M$	0.764	$\angle AMB$	0.747	$\mu(A,B)/M$	0.769
$\angle COX$	0.762	C/X	0.746	X	0.767
X	0.762	$\mu(A,B)$	0.734	$\mu(A,B)/X$	0.759

ADJECTIVE		NOUN	
$\mu(A,B)/M$	0.745	$\mu(A,B)$	0.756
$\overline{AC} : \overline{BC}$	0.736	C	0.747
$\overline{AC}/\overline{BC}$	0.734	$\mu(A,B)/X$	0.728
$\mu(A,B)/X$	0.732	$\mu(A,B)/M$	0.721
$\angle ACB$	0.730	C/X	0.721

Table 6-D: Independent f-scores from the metaphor classification data for top five features of each subspace type for 5x5 word co-occurrence window, 400 dimension subspaces.

generalise even further from this data to a broader range of metaphors and to moreover move from classification to gradation based on observations of merely binary judgements of metaphoricity. For now, I simply note that it is remarkable that data about nothing more than the way that words tend to be collocated can, with the aid of a mechanism for contextualisation, reveal so much about the nature of the semantic relationship between the lexical components of an previously unseen phrase.

6.1.2 The Geometry of Metaphor

In this section, I will explore the geometric features which prove most productive in the classification of metaphor. As with relatedness and similarity in the previous chapter, I begin by examining the capacity of independent features to predict metaphor. Rather than a proper logistic regression involving multiple independent variables fed into a non-linear function, this analysis amounts to choosing a cut-off point in terms of the value of each feature separating literal and metaphoric phrases in the subspaces which an analysis of their corresponding word-vectors delineate. So the f-scores reported in Table 6-D can be understood as indicating the degree to which the values of a given geometric feature separate the dataset into distinct categories corresponding to human judgements of metaphoricity.

The scores themselves reflect the trend observed in Table 6-A and 6-B: the JOINT and ZIPPED subspaces produce features that are particulary good at classifying metaphor, with a decrease in performance in the INDY subspaces and then another step down in the single-word subspaces. None of the scores themselves come close to the levels of discrimination achieved by the models learned from full feature vectors, with the difference

between the performance of the best feature for the JOINT technique and that of the corresponding full featured space somehwat significant with $p = .006$. In terms of the actual features indicated by this analysis, two in particular figure prominently in one way or another, namely, the mean of the word-vector norms $\mu(A, B)$ and the norm of the central-vector C. In the first instance, the role of the relationship between word-vectors and the origin of the spaces that their salient co-occurrence dimensions delineate is once again reflective of the preliminary findings on conceptual geometry described in Chapter 4.6, where norm was seen to be an effective mechanism for defining a region of conceptual constituency. In the case of the distance of the central vector from the origin, the emergence of this feature, as well of the appearance of the norms M and X as components of various strongly predictive tendencies, indicate that here, as with similarity in the previous chapter, characteristics of dimensions outside of the situation of any particular word-vector along them might be in themselves indicative of metaphor: some words might simply be more likely to co-occur in the context of metaphoric language, and co-occurrence statistics should provide a handle for examining this tendency.

To further delve into the statistical geometry of metaphor, and in line with the results on relatedness and similarity described in the previous chapter, I once again search the state space of possible combinations of features to find the optimal feature vector for classifying metaphor in context sensitive subspaces. This is again treated as a beam search problem, though the search space expanded at each level of the search tree is here limited to the top 500 combinations of features given the larger size of the data being modelled. Table 6-E presents the optimal seven feature combinations discovered for the 5x5 word window, 400 dimensional JOINT subspaces based on both a standard ten-fold cross-validation and the version of the data shuffled in order to test on data not observed in each training phase. The f-scores achieved by these combinations of features, reported next to the respective labels at the top of the table, indicate a marginal decrease in the overall performance as compared to the full featured models of subspaces, but the results are still strong.

Angles between generic vectors, which were already evident as independently predictive features in Table 6-D, have a strong effect here, with the strong negative correlation of $\angle COX$ in the ten-fold cross-validation in particular suggesting that maximal values tend to be relatively similar across dimensions jointly selected by literal adjective-noun combinations, pulling the line of X closer to the centroid described by C. To put this differently, as pairs become more metaphoric, they tend to also become less consistent in the type of dimension that they co-select, as evidenced in the increasing variance in the maximum values of these dimensions. Perhaps the most interesting thing to observe here, though, is the strong correlation between ratios of word-vector to generic vector distances

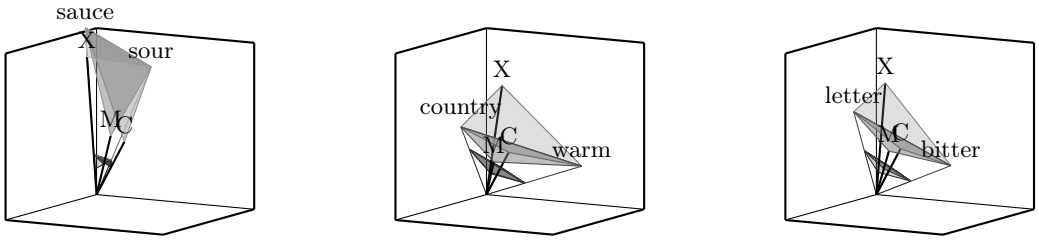
	10-fold ($f = 0.869$)	shuffled ($f = 0.830$)
	DISTANCES	
word-vectors	-	-
generic vectors	$M = -1.448$	-
	ANGLES	
word-vectors	$\angle ACB = -0.775$	-
normalised	-	-
generic vectors	$\angle COX = -1.618$	$-0.271 = \angle COM$
	$\angle COM = 0.974$	$0.045 = \angle MOX$
	MEANS	
word-vectors	$\mu(\overline{AM}, \overline{BM}) = -1.124$	$-1.007 = \mu(\overline{AC}, \overline{BC})$
normalised	-	-
	RATIOS	
word-vectors	-	$0.492 = \overline{AM} : \overline{BM}$
		$-0.620 = \overline{AX} : \overline{BX}$
normalised	-	$-0.168 = \overline{A'C'} : \overline{B'C'}$
	FRACTIONS	
word-vectors	$\overline{AC}/\overline{BC} = 0.325$	-
generic vectors	$M/X = 1.305$	$0.252 = A/B$

Table 6-E: The seven most predictive features for metaphor classification, compared between ten-fold and sight-unseen cross-validation of logistic regression on statistics extrapolated from 5x5 word window, 400 dimensional JOINT subspaces.

in the case of the version of the data shuffled to test on unseen adjectives, but not in the case of the stratified cross-validation. The positive correlation with the balance of the distances from the word-vectors to the mean vector M means that subspaces where the word-vectors have a relatively even relationship to the weighted centre are, in fact, more metaphoric (and their relationship to the maximum vector is comparatively less balanced, with this vector in turn being less central to the space per the observations regarding $\angle COX$). But more generally, it is noteworthy that the balance between word vectors and generic vectors is informative about metaphoricity specifically in models tested on unseen adjectives: this balance is in effect a projection into space of quotients of joint probabilities of observing words and co-occurrence terms divided by the typical or maximal probabilities of being observed with the co-occurrence terms, and from it we can infer that these quotients are generally predictive of metaphor in context, even without word-specific training data.

Figure 6.2 presents visualisations by way of three dimensional projections of word-vectors and generic vectors from 400 dimensional JOINT subspaces selected from the 5x5 word window base space.[3] In the example of the uncontroversially literal phrase *sweet*

[3]These projections have been rendered using the same regression technique as applied to the images for related word pairs in the previous section, but the coordinates of X have been divided by 1.5 instead



literal: sweet watermelon *neutral: warm country* *metaphoric: bitter letter*

Figure 6.2: Three dimensional projections of word-vectors and generic vectors in subspaces for pairs at the extents and in the middle of the literal-metaphorical spectrum, taken from 5x5 word window, 400 dimensional subspaces selected using the JOINT technique.

watermelon, the word-vectors are characteristically far from the origin and close to one another, corresponding to the predictivity of $\mu(A, B)$ in particular. At the other extent of the spectrum, the highly metaphoric phrase *bitter letter* is characterised by a dropping of the word-vectors and a widening of the angle between them; the generic vectors, meanwhile, are now further from the origin relative to the word-vectors that select the subspace and, at the same time, draw closer to one another in particular at the normalised layer of the subspace. But most interestingly, at a relatively neutral point, occupied by the intriguingly ambiguous phrase *warm country*, which the logistic regression trained on these subspaces assigned a score of close to 0.5, there is actually evidence of an intermediary widening out of the overall array of points even as the word-vectors remain fairly far from the origin.

The exciting thing about this last observation is that it suggests that, rather than existing on a linear or even monotonic scale, metaphor may itself actually be a multi-dimensional phenomenon, with a characteristic particular to highly ambiguous word combinations that is to some extent separate from the statistical features of straightforward literalness and clear cut metaphoricity. The broad arrangement of word-vectors in space engendered by the contextualisation of the phrase *warm country*, in contrast to the relatively tight relationship of the generic vectors, can be interpreted as revealing an uncertainty regarding the semantic properties being transferred in this small composition, corresponding to a drifting of the word-vectors and a contracting of the generic vectors across the jointly selected co-occurrence profile. Here, once again, the statistical geometry of a subspace can be productively mapped to a theoretical statement about the nature of a semantic phenomenon as characterised by a selectively contextual and quantitative representation of observations about the way that words are used, by and large outside of any strong preconditions symbolically encoded in the computational framework.

of 2.

6.1.3 Generalising the Model

One of the interesting things about feature-based classification is that there is typically an inherent commitment to degree of class membership, even when the training data used to build a model is simply binary. This is true of any model which uses, for instance, a logistic regression technique for determining class, as there is a cut-off point along the spectrum of model output and a corresponding proximity to that point for any given sample, and it is especially obvious when the features of the model are actually geometrical measures. In this section, I will apply the models learned from the the Gutiérrez et al. (2016) data to another dataset designed to assess metaphor as a matter of degree rather than simply as a binary situation, and a dataset that additionally deals with a different type of metaphor in terms of composition. The question explored here is whether the geometric features of context specific distributional semantic analysis of word-vectors will provide binary classification models with adequate information for projecting metaphoricity along a continuous scale.

The data used for this experiment was originally reported by Jankowiak et al. (2015), and was used to train a model based on an earlier version of methodology as described by Agres et al. (2016). This data consists of 228 predicate-object word pairs selected to cover three degrees of metaphor, consisting of literal pairs such as *announce willingness*, conventionally metaphoric pairs such as *cut pollution*, and novel metaphors such as *smell excuses*. 102 human participants provided metaphoricity scores on a seven point Likert scale, and the average scores were compiled into the dataset that will used to test models learned from the geometric features output by my context sensitive methodology.[4] Specifically, I will experiment with two different classification model techniques. In the first instance, I will take the output of the logistic regression described above, trained on the Gutiérrez et al. (2016) data, as assigning probabilities to the metaphoricity of an input word pair, and I will in turn measure the degree to which these probabilities correlate with the degree of metaphoricity collectively assigned by human raters. In the second instance, I'll use the binary metaphor classification data to train a support vector machine.[5] Applying a radial basis function kernel, I analyse the correlation between distance from the discriminatory hyperplane and the human ratings. In both cases, and in line with results reported in the previous chapter, Spearman's correlations are the unit of analysis.

Table 6-F presents results for both modelling techniques, focussing on features extrapolated from 5x5 word window, 400 dimensional subspaces using the JOINT approach and

[4]Studies were also conducted to gather ratings for *familiarity* and *meaningfulness*, but those ratings will not be modelled in this thesis.

[5]This is implemented using the python scikit-learn `SVC` module.

features	1	3	5	7	9	full
			logistic regression			
JOINT	0.368	0.355	0.033	0.085	0.279	-0.033
ADJECTIVE	-0.377	0.355	0.044	0.513	0.511	0.335
			support vector machine			
JOINT	0.352	0.359	0.042	0.045	0.243	0.158
ADJECTIVE	-0.170	0.247	-0.021	0.407	0.418	0.236

Table 6-F: Spearman's correlation with human verb-noun metaphoricity scales judgements based on logistic regression and support vector machine models trained on adjective-noun classification data, taking feature vectors of various lengths as independent variables.

through an analysis on the adjective in each word-pair from the input data. Feature vectors of various lengths, picking the optimal geometric features for each dimensional selection technique, are used to feed input to each model. In terms of the models trained on features from JOINT subspaces, there is a clear trend towards strong performance with one or three features, weaker performance with five or seven features, stronger performance again with nine features, and then a drop-off again in the full featured space. The relatively low performance with the full set of features is not particularly surprising: there is clearly an encroaching incidence of generalisation error here as the models become flooded with data about various and certainly collinear statistical features of contextual geometry. At the shallow end of the feature selection parameters, on the other hand, the single measure $\mu(A, B)$ (per Table 6-D) once again points to the efficacy of word-vector norm as a predictive characteristic of contextualised co-occurrence subspaces.

The really remarkable outcome here, though, is the very strong performance of the models learned from the top seven and nine features extracted from subspaces selected by PMI values of the adjective word-vectors alone. This is particularly interesting given that the data being tested actually consists of a different type of grammatical relationship, namely, predicate-object pairs. It would seem, then, that the co-occurrence dimensions most salient to either verbs or adjectives generate a geometry in which their relationship to potential arguments can play out in similar ways in terms of the metaphoricity inherent in the semantic context: the interaction between the selecting vector, the noun-vector, and the generic vectors translates from one type of composition to another in an isomorphic way. This explantion, including the claim that the mapping of predictive features from one type of metaphor to the other is to a large extent isomorphic, is supported by the particularly strong performance of the logistic regression at seven and nine dimensions, where the logistic function takes a polynomial with coefficients learned in the training phase as direct input. The more complex non-linearity afforded by the support vector machine appears to actually somewhat confound the mapping from verb-noun to

adjective-noun phrases—though the difference between the correlations at nine dimensions is not statistically significant at $p = .104$ based on a Fisher r-to-z transformation.

The one area where a support vector machine provides a clear improvement in performance is in the full dimensional models extrapolated from JOINT subspaces. In this case, it would seem that the radial basis function classification actually does a better job of avoiding the overfitting in a higher dimensional feature space. But, putting questions of model choice aside, there is clear evidence here for the generality of the contextual geometry of metaphor, and also a strong case for the appropriateness of machine learning techniques for providing an appropriate mechanism for the computational manipulation of co-occurrence information to build a more nuanced model of degree of metaphor based on relatively rudimentary classification data. Crucially, it is the context sensitivity of my methodology that facilitates the exploration of a multi-dimensional feature space in which the non-linear nuances of this particular semantic phenomenon can be discovered; a model providing a singular static relationship between lexical representations could not offer the context specific underpinning for generating a geometry replete with interpretable statistical features. Finally, there are signs here to invite further research, and indeed some grounds for hoping that a context sensitive approach might have the scope for handling more sophisticated tasks such as metaphor interpretation and generation.

6.2 An Experiment on Coercion

In this section, I will apply my methodology to the classification of a phenomenon closely related to metaphor, namely, *semantic type coercion*, by which the semantic type of a word is shifted through its interaction with another word: in the cases examined here, verbs that select for a particular semantic type will be seen to coerce nouns from one conceptual category to another by taking those nouns as arguments. So, for instance, in phrases like *denied wrongdoing* or *heard footsteps*, the nouns in play are standing in for a conceptually relevant but different type of noun, and the literal versions of these phrases would go something like *denied committing wrongdoing* or *heard the sound of footsteps*, where the verbs select arguments of types along the lines of ACTIVITY and PERCEPTION respectively. This phenomenon is often referred to as *logical metonymy*, identifying it as a subspecies of the more general figurative phenomenon metonymy by which a thing is denoted by a conceptually related lexical representation.

Coercion is one of the semantic phenomena targeted by Pustejovsky's (1995) theory of a *generative lexicon*, by which nouns are semantically modelled as having a *qualia structure* which maps out the way that a thing relates to itself, the world, and the agents

interacting with it in that world on four different levels of abstraction, with the general objective of arriving at "a model of meaning in language that captures the means by which words can assume a potentially infinite number of senses in context, while limiting the number of senses actually stored in the lexicon," (ibid, p. 104). In terms of coercion, qualia provide the basis for a process of *projection* by which a variety of semantic types can be extracted from a complex type (or a *dot object* in Pustejovsky's lingo) in order to fulfil the typing requirements of a predicate in open ended ways. The model that emerges here – one built on dynamically interactive lexical semantic representations contingent on some sort of general conceptual context – begins to look like the general linguistic stance that has motivated my own methodology.

This theoretical commitment suggests a schematic by which a symbol manipulating system might begin to get a handle on productive and context sensitive lexical representations of things in the world. To this end, Jezek and Hanks (2010) have described an ontology based on a computational analysis of co-occurrence patterns designed to facilitate the modelling of what is ultimately a sliding scale of statistically enhanced semantic representations, or "shimmering lexical sets," (ibid, p. 19), as the authors put it. Applying a similar notion that coercion is probabilistic rather than discreet, Lapata and Lascarides (2003) use co-occurrence statistics to try to predict the verbs which, in the role of for instance participles, successfully resolve instances of coercion. And, under the rubric of *logical metonymy*, Shutova et al. (2013) expand upon the work of Lapata and Lascarides by extracting verb senses from WordNet to build a class based model, to some extent recapitulating the categorical distinctions that characterise many theoretical approaches to coercion. The motivation behind this last system is the apt observation that, in the case of coercion, "humans are capable of interpreting these phrases using their world knowledge and contextual information," (Shutova et al., 2013, 11:2).

Returning to the theoretical issues regarding grammaticality raised earlier in this chapter, the analysis of coercion within the framework of the generative lexicon points to something more like a graduated typology, sliding from specific instances of processes, things, and the like to more general conceptual categories and finally to entire classes of words. As Langacker (1991) has pointed out, there is a lurking ambiguity in grammatical class distinctions, with various conceptual schema existing in any natural language for moving between classes: so, to borrow an example from Langacker, phonological and symbolic dynamics facilitate a conceptually coherent progression from *sharp* to *sharpen* to *sharpener*, and the rules that are extrapolated as an explanatory framework for such transitions are just a way of systematising the cognitive networks that underpin this linguistic phenotype.[6] And as Copestake and Briscoe (1995) point out in their probabilistic

[6]Wittgenstein's (1967) quip regarding "grammatical fictions," (ibid, ¶307) also seems pertinent.

account of coercion, selectional preferences are at least to a certain extent conditioned by factors involving word frequency, suggesting that there could be grounds for a distributional mechanism for modelling semantic shifting.

With this in mind, my hypothesis is that, as with metaphor in the previous section, a syntactically neutral statistical model with a context generating capacity should be able to capture the way in which, in the case of argument type coercion, a predicate specifies some conceptual contingency of the coerced object in order to accommodate its selectional preference. The purpose of this set of experiments (an early version of which is reported in McGregort et al., 2017) is to test this broad hypothesis, and to explore the particular statistical features of co-occurrence which afford appropriate contextualisations. This will serve, to a certain extent, to address a question raised by Pustejovsky and Jezek (2008), who illustrate some of the difficulties inherent in extracting typological structure from a distributional analysis of a large-scale corpus. The point made there is that "generative mechanisms in the semantics, such as coercion, modulate meanings in context and allow words to behave distributionally in unexpected ways with respect to their selectional properties," (ibid, p. 209). Those authors show how a model involving a dynamic between a statistical approach such as a distributional semantic model and a theoretical structure such as the generative lexicon can accommodate some of this unexpectedness. My goal in the following experiments is to explore the extent to which a context sensitive approach to distributional semantics can, without the structure of a symbolic formalism or pre-formulated grammatical or typological annotations, address the pertinent theoretical issues raised by the kind of analysis offered by Pustejovsky and Jezek.

6.2.1 Methodology and Results

The data which will be used to test my methodology in this section was originally presented by Pustejovsky et al. (2010) as a task for the ongoing International Workshop on Semantic Evaluation series of computational semantic modelling challenges. The data consists of 2,071 sentences (originally split into a test set of 1,039 training instances and 1,039 testing instances)[7] each containing a marked verb and object, with the object classified as either coercive or not. The verbs cover various conjugations of five different verb stems, each identified as selecting for a different semantic type as an argument: the verbs (and the semantic type selected) are *arrive* (LOCATION), *cancel* (EVENT), *deny* (PROPOSITION), *finish* (EVENT), and *hear* (SOUND). The objective, then, is to train a model to indicate that the phrase *finish the party* is not coercive, in as much as we accept

[7]The data is available under task seven at `http://semeval2.fbk.eu/semeval2.php?location=data`.

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.604	0.619	0.630	0.657	0.634	0.672	0.673	0.691
INDY	0.666	0.677	0.703	0.693	0.652	0.660	0.707	0.679
ZIPPED	0.568	0.624	0.610	0.647	0.596	0.625	0.658	0.663
VERB	0.664	0.675	0.698	0.704	0.631	0.652	0.699	0.700
NOUN	0.601	0.628	0.643	0.633	0.518	0.565	0.603	0.641
SVD	0.511	0.523	0.539	0.412	0.521	0.409	0.483	0.563
CBOW	0.498	0.508	0.531	0.493	0.496	0.544	0.535	0.496
SG	0.518	0.565	0.575	0.529	0.534	0.523	0.583	0.557

Table 6-G: F-scores for coercion identification based on a ten-fold cross-validated logistic regression taking geometric features of various subspace types as input.

that *party* denotes a member of the conceptual category EVENT, whereas *finish the food* is because what is actually being finished is the event of eating food, not the food itself. For the purposes of the original presentation the data is split into a training set and a testing set of roughly equal size, but questions of the most meaningful partitioning of the data will be discussed below.

Two amendments are made to the data as presented. First, of the 2,071 verb-object pairs, 78 contain multi-word objects not compatible with the vocabulary used for my model, reducing the total number of word pairs to 1,992, 591 of which are considered coercive. Second, of these remaining computable word pairs, 903 are duplicates (they are presented in unique sentences, but for the first phase of analysis here only verb-noun pairs will be consider; sentential context will be addressed below). This leaves a total of 1,029 word pairs, 399 of which are deemed coercive. As with the metaphor data in the previous section, I train a logistic regression model to discriminate between regular argument selection and coercion. I once again take the two words being analysed as input to generate a number of different context specific distributional semantic subspaces, treating the 34 geometric features outlined in Table 5-H plus the seven additional fractional features specific to asymmetric input terms described above in Section 6.1.1 as the independent variables of the regression analysis.

Table 6-G presents the f-scores derived from the precision and recall results of a ten-fold cross-validation of these logistic regression models. Most obviously, these numbers are considerably lower than the comparable results for metaphor outlined in Table 6-A, but this is to some extent mitigated by the relative scarcity of instances of coercion in the data: a minority class baseline always classifying word pairs as coercive would, based on the above data statistics, give $f = 0.558$. The top score of $f = 0.707$ for the context sensitive models, achieved by the 5x5 word window, 200 dimensional INDY dimension

selection technique, is substantially better than the baseline with the probability of this difference happening by chance at $p = .028$, and the difference with the skip-gram model with the same parameter are likewise notable, if not outright statistically significant, at $p = .075$. Of the three dimensional selection techniques that use both words as input, the INDY method achieves the overall highest scores (as opposed to the JOINT technique for metaphor), but it must be noted that these top results come at 200 dimensional subspaces selected from both 2x2 and 5x5 word window spaces, suggesting that there is a degradation in the usefulness of information included on dimensions past a certain point of saliency for a given input word. The progression of results as dimensionality increases is evident elsewhere here as well, with the single word input dimensional selection techniques as well as with the static SVD and `word2vec` models. The SVD models in particular perform erratically on this task, hinting that the angular relationships in a centred space of word-vectors which has proved effective on previous tasks provides only marginal information about the selectional relationships between predicates and objects.

In line with the metaphor results is the overall poor performance of the static models, which generally do somewhat worse than the baseline and substantially worse than the context sensitive models. Of particular note is the decline of the SVD models and the comparative ascent of the `word2vec` skip-gram methodology: the sentential context predicting mechanism of the skip-gram approach seems to better capture the typological relationships between predicates and arguments than a principal component analysis of the dimensional variance in a base space of co-occurrence statistics. But in fact, the results here are across the board less regular in their relationship to parameters of dimensionality and co-occurrence window size, with a more even distribution of relatively high and low scores for both 2x2 and 5x5 word co-occurrence window models, and comparatively strong outcomes occasionally popping up for 20 or 50 dimensional spaces. The seemingly erratic output of the model gives an overall impression of an unanchoring between the statistics of co-occurrence and the semantic phenomenon being explored here. Perhaps in the case of coercion, or at least in terms of the data sampled here, many predicate-object combinations are, regardless of the influence of the verb on the noun's conceptual situation, too conventional for type shifts to be detected in a meaningful way in terms of co-occurrence profiles.

Another telling feature of these results is the quite strong performance of the subspaces selected by an analysis of the verbs alone. In fact, this is likely to be an artefact of the data itself: only five different verb stems are present, and some are arguably marked by their own semantic peculiarities, with, for instance, *finish* coercing 152 out of the 252 arguments it takes in the data, where the rate for *deny* is only 29 out of 183 instances. In order to find out if the models being tested here are actually just learning, in one

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.338	0.397	0.362	0.381	0.345	0.428	0.404	0.386
INDY	0.454	0.386	0.436	0.459	0.369	0.350	0.411	0.410
ZIPPED	0.256	0.297	0.363	0.358	0.324	0.352	0.377	0.357
VERB	0.233	0.334	0.361	0.448	0.307	0.401	0.352	0.336
NOUN	0.306	0.398	0.406	0.401	0.243	0.293	0.317	0.340
SVD	0.295	0.252	0.276	0.126	0.217	0.173	0.301	0.288
CBOW	0.368	0.329	0.248	0.162	0.302	0.316	0.245	0.177
SG	0.349	0.333	0.281	0.194	0.366	0.351	0.316	0.229

Table 6-H: F-scores for coercion identification taking each verb stem type as a separate fold of a cross-validation.

way or another, specific rules about particular inputs, I rearrange the data into five folds corresponding to the five verb types present, training a model on each combination of four different verbs and then testing the model on the classifications of word-pairs involving the fifth. F-scores are reported in Table 6-H.

There is indeed a notable drop-off in scores across the board here, with the hypothesis that there is no difference between the top INDY 400 dimensional, 2x2 word window score here and the top score from the unshuffled version of the data fairly unlikely at $p = .030$. On the other hand, the progression of scores as dimensionality increases remains jagged, with the static models particularly notable in their poor performance at higher dimensionalities. So it would seem that a great deal of what is being learned here may be specific to the verbs and the types of the arguments they take, a hypothesis supported by the relatively weak showing for the verb-only dimension selection technique. On the other hand, the verb-only, noun-only, and INDY techniques, unlike the various other methods, do now evince a steady increase in performance as dimensinoality increases, suggesting that with this rearrangement of the data these approaches are now at least discovering much of what can be classified about coercion based on co-occurrence statistics. In fact, it should be remarked that each INDY subspace is composed of the first half of the dimensions selected by the verb-only technique, combined with the first half of the noun-only subspaces, so the correspondence between these approaches isn't surprising. It is noteworthy that here subspaces built from a conjunction of dimensions associated with the two words in play are most indicative of the categorical shifting of a noun's type, rather than the subspaces formed by dimensions which are each in themselves representative of something of a conjunction in the salient co-occurrences of both words, as was the case for metaphor classification.

Figure 6.3 presents a receiver operating characteristic plot comparing between the

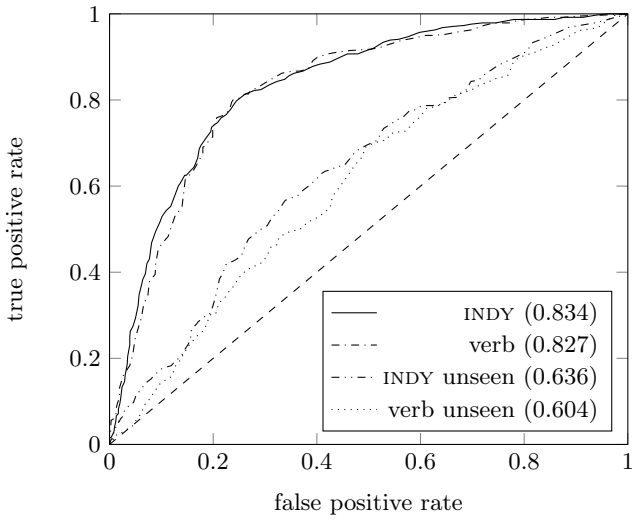


Figure 6.3: Receiver operating characteristic plots for a selection of models for coercion classification, with the area under the curve for each model type indicated in the legend.

verb-only and INDY techniques for both the regular and rearranged versions of the data, with areas under the curve indicated in the legend. As expected, the verb-only and INDY techniques are comparable for the unaltered version of the data, but the model learned from verb-only subspaces falls off once each verb stem is treated as its own fold of the data. Of particular note is the way that the rearranged verb-only curve flattens out in the middle: the relative drop-off in true positives in the mid-range of cut-off points for classifying coercion tells us that there is a lull in the precision of the model here, with mistakes being made on the interpretation of subspaces projected by unfamiliar verbs (keeping in mind that, in the unaltered version of the data, the subspace projected by two different instances of the same verb morpheme would be identical, and so it is only the variance in the relative situation of the noun-vector in these subspaces that needs to be analysed to evaluate coercion). The relative jumpiness of the curves as compared to the smooth trajectories observed for the metaphor data in Figure 6.1 can be attributed to the scale of the data, with the massiveness of the metaphor dataset providing a steadier progression as the criteria for positive classification are relaxed. On the whole, though, the story here is a similar one of a fairly balanced advance of recall and a correspondingly steady decline in precision as the model becomes increasingly permissive in its classification of coercion.

JOINT		INDY		ZIPPED	
$\mu(\overline{A'X'}, \overline{B'X'})$	0.526	$\mu(\overline{A'X'}, \overline{B'X'})$	0.547	$\mu(\overline{A'C'}, \overline{B'C'})$	0.392
$\mu(\overline{A'C'}, \overline{B'X'}$	0.496	$\mu(\overline{A'C'}, \overline{B'C'})$	0.544	$\mu(A, B)/C$	0.349
$\mu(\overline{A'M'}, \overline{B'M'}$	0.453	$\mu(A, B)/C$	0.522	$\mu(\overline{A'X'}, \overline{B'X'})$	0.321
$\mu(A, B)/C$	0.442	$\angle AOB$	0.517	$\mu(\overline{A'M'}, \overline{B'M'})$	0.237
$\angle AOB$	0.429	$\mu(\overline{A'M'}, \overline{B'M'})$	0.504	$\angle AOB$	0.209

VERB		NOUN	
$\overline{AC}/\overline{BC}$	0.580	$A : B$	0.528
$A : B$	0.412	A/B	0.486
A/B	0.387	$\mu(A, B)/C$	0.486
$\mu(\overline{A'M'}, \overline{B'M'})$	0.384	$\angle AMB$	0.427
$\mu(\overline{A'X'}, \overline{B'X'})$	0.374	$\angle ACB$	0.423

Table 6-I: Independent f-scores from the coercion classification data for top five features of each subspace type for 2x2 word co-occurrence window, 400 dimension subspaces, validated on unobserved verbs.

6.2.2 The Geometry of Coercion

Following the procedure which has proved productive for the analysis of semantic phenomena in preceding experiments, I will now study the statistical geometry associated with the contextual classification of coercion, beginning with an analysis of individual features and moving on the a consideration of optimal combinations of features. In Table 6-I, I once again report the top five performing (in terms of f-score) features for each of the context sensitive dimension selection techniques described in the previous section. These features have been tested on the more prohibitive version of the coercion data rearranged to avoid training and testing on the same verb stem types, and with this in mind the improvement in scores here as compared to Table 6-H is remarkable. With the exception of the ZIPPED subspaces, all other techniques exhibit substantial improvements on the models learned from the full set of statistical features, with the difference in the verb-only spaces especially notable and a substantial improvement with the chance of the results being randomly attained at $p = .070$.

Also of note is the character of the features that are most predictive for each dimensional selection technique. For all three methods involving both words as input for subspace selection, the mean values of distances at the normalised level of subspaces feature prominently (and it should be noted that the angle $\angle AOB$, which also features here, is perfectly correlated with the distance between the normalised word-vectors A' and B'). This indicates that the angles formed between the word-vectors and the generic vectors are especially associated with coercion, and this as opposed to metaphor, where the distance of various vectors from the origin as well as the ratios of these distances are

particular predictive. That the averages of the angles with the generic vectors seem generally more significant than the angles between the word-vectors themselves is evidence that it is the absolute and combined situation of the word-vectors in the context of their subspaces, rather than their relationship to one another, that can be interpreted in terms of a typological semantic relationship such as coercion.

All of this is in opposition to the top features for the subspaces selected based on a single input term, where fractions and ratios are prominent. Of particular note is the significance of the relationship between the word-vectors, and this makes sense: given that the relevance of one word-vector in a subspace selected by the other is only incidental and in no way built into the space itself, differences in the relative lengths of the word-vectors and their relative distances to generic points are particularly indicative of degrees of inclusion in the profile characteristic to the dimension-selecting term. The implication is that, with a co-occurrence profile chosen by one word, it is simply the prominence of the other word with respect to this profile that is indicative of the typological relationship between the verb's selectional constraints and the noun's categorical expectation. Furthermore, the resurgence of verb-only spaces here in terms of a singular feature, namely, the fraction of the verb-centre-vector distance $\overline{AC}$ to the comparable distance $\overline{BC}$ tells us that the comparative situation of the word-vectors to the absolute centre of a subspace varies between coercive and non-coercive cases of argument selection. It's worth noting here that, in verb-selected subspaces projected from the 2x2 word window base space, co-occurrences dimensions will correspond to terms that tend to be observed in close proximity to the verbs themselves, so we can expect these dimensions to be characterised by arguments of the verbs and modifiers of those arguments: it isn't hard to imagine how, in terms of modifiers in particular, the typical characteristics of the arguments normally selected by a verb would serve as a kind of template for testing the typological fit of a new candidate argument, with relative proximity to the centre, along with the extent of the noun vector along these characteristic co-occurrence dimensions, being good metrics for determining the fit.

These independent feature results are suggestive of the types of statistics that are associated with coercion, but not of the direction of these correlations, let alone the dynamics between different statistics. To examine the geometry of coercion more in depth, I once again perform a beam search to discover the top seven features associated with both the INDY and verb-only subspace selection techniques in 400 dimensional subspaces projected from 2x2 word window base spaces, training models on the rearranged version of the data and applying a vector inflation factor in order to avoid collinearity between input features. Results are reported in Table 6-J. Remarkably, a very different picture emerges than what was observed above regarding independent features, with neither the mean

	INDY ($f = 0.681$)	VERB ($f = 0.688$)
	DISTANCES	
word-vectors	-	-
generic vectors	-	$-0.833 = X$
	ANGLES	
word-vectors	$\angle AMB = -0.564$	-
	$\angle ACB = -0.103$	
normalised	-	$0.290 = \angle A'M'B'$
generic	-	$1.241 = \angle COM$
		$-0.214 = \angle COX$
	MEANS	
word-vectors	$\mu(A, B) = 1.656$	-
normalised	-	$0.452 = \mu(\overline{A'X'}, \overline{B'X'})$
	RATIOS	
word-vectors	$\overline{AM} : \overline{BM} = 0.450$	-
normalised -	-	
	FRACTIONS	
word-vectors	-	$2.315 = \overline{AM}/\overline{BM}$
normalised	$\overline{A'M'}/\overline{B'M'} = -0.259$	-
	$\overline{A'X'}/\overline{B'X'} = 0.203$	-
generic vectors	$C/M = -1.257$	$-2.398 = C/M$

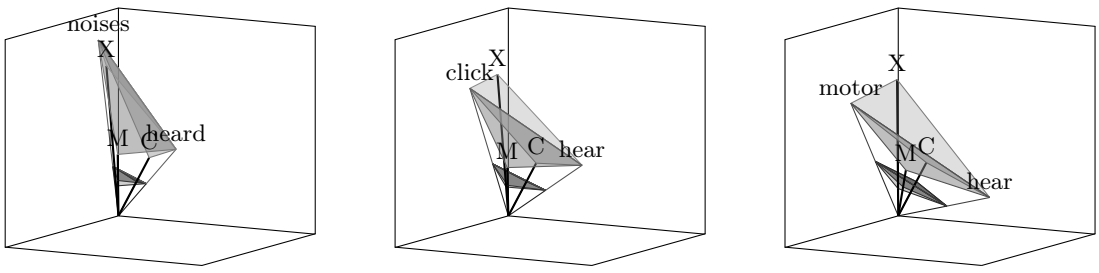
Table 6-J: Comparison of the seven most effective features for coercion classification in 2x2 word, 400 dimensional subspaces for INDY versus VERB based dimension selection.

distances between the norms of the INDY subspaces nor the angles and ratios individually observed in the verb-only subspaces making an appearance. In fact, one of the most notable characteristics of the respective feature vectors is, on the one hand, the spread of the features across several different categories of co-occurrence statistic, but then also the balance between non-normalised features in one category for one technique versus the normalised components of the same category for the other technique.

So, for instance, the angles $\angle AMB$ and $\angle ACB$ both correlate negatively with coercion (meaning the angles are wider for more coercive word pairs) in the INDY type subspaces, implying that the word-vectors are more likely to be found on opposite sides of these two central points in the space in the case of coercive pairings (but not necessarily on opposite sides of the lines extending from the origin through these points–they could, for instance, be above and below a point relative to the origin). The positive correlation with the angle $\angle A'M'B'$ in the verb-only subspaces, on the other hand, indicates that the word-vectors tend to be on opposite sides of the lines extending through the mean vector. This is interesting, since we can safely assume that the verb-vector will occupy a relatively central position in a subspace defined entirely by dimensions with which the verb has a high expectation of co-occurrence: it would seem that the noun-vector

essentially pivots towards the co-occurrence dimensions that are most strongly associated with the verb, meaning that the words that are especially characteristic of the immediate syntagmatic situation of the verb tend to have a stronger association with nouns of a type not paradigmatically selected by the verb. In the case of fractions of lengths between vectors, on the other hand, the very strong positive correlation between coercion and $\overline{AM}/\overline{BM}$ in verb-only spaces suggests that as the noun-vector (associated with B) retracts towards the origin relative to the verb-vector, coercion is more likely. This makes sense in this type of subspace, since nouns of types that categorically satisfy a verb's selectional constraints will tend to have higher PMI values along the dimensions selected by that verb. The negative value for the normalised version of the same fraction in the case of the INDY subspaces, however, means that the respective angles between the word-vectors and the mean vector tend to be more balance in cases of coercion.

The one point of consistent comparison across the two techniques is the fraction of the length of the central vector C divided by the length of the mean vector M. As discussed in Chapter 5.3 in the context of the comparison between relatedness and similarity, the negative correlation here for both the INDY technique and the verb-only technique indicates an increase in the likelihood of classifying a relationship as coercive as variance across the mean values of features delineating a subspace increases. If high variance were only associated with coercion through the INDY technique, it could be argued that coercion simply correlates with subspaces patched together from two different independently selected co-occurrence profiles with a tendency towards have very different mean values: for instance, one word might select co-occurrence terms that occur more frequently and therefore have lower mean values than the other. Given that this effect is even stronger for the verb-only subspaces, however, where the co-occurrence profile of just one term is in play, the prominence of this feature actually indicates that, as with similarity, there are some words (and, in particular, verbs) which just tend to be more coercive than others. Moreover, these words tend to have a particular statistical characteristic by which the terms with which they co-occur tend to have more varied mean values. In the end, this can be reduced to an observation that might not seem particularly surprising, even if it also wasn't immediately obvious prior to this geometric analysis: words that tend to be involved in coercive relationships also tend to co-occur with a range of other words that are more varied in terms of their own essential characteristics such as frequency. This interpretation is further supported by the negative correlation between the length of the maximum vector X and coercion in verb-only subspaces. This generally indicates that there are some basic differences between the types of verbs, and the corresponding dimensional profiles, that tend to coerce their arguments, and specifically implies that coercive verbs tend to be observed in close proximity to at least some higher frequency words.



non-coercive: heard noises *neutral: hear click* *coercive: hear motor*

Figure 6.4: Three dimensional projections of word-vectors and generic vectors in INDY subspaces for pairs at the extents and in the middle of the spectrum of coercion.

Figure 6.4 illustrates word-vectors and generic vectors projected into exemplary instances of subspaces at three different phases of coercion. Similarly to the literal stage of metaphor illustrated in Figure 6.2, the least coercive instance is characterised by a relatively tight space, with the noun-vector particularly prominent and acute angles at the vertexes of the generic points. The next two stages of coercion go through something of the reverse of the process observed with metaphor, however, beginning with a contraction and a bit of an widening in the somewhat neutral case of *hear click*, and then opening up into a more disparate arrangement across the subspace for the unambiguously coercive *hear motor*. The angles at the vertices of the generic point in particular, as well as the balance between the distances between the word-vectors and these points, follow the trend indicated by the coefficient weightings reported for INDY subspaces in Table 6-J. The shift in the overall balance of the word-vectors is interesting to note, as well: while the ratio $A : B$ wasn't indicated in the feature-wise analysis of the space, the prominence of the vector for *noise* in the projection on the left suggests that there could be some nouns which are less likely to be susceptible to coercion, and indeed it is tricky to imagine the context in which the fairly abstract word *noise* could be adapted to a conceptual category other than NOISE.

A more general implication of this geometric analysis is the idea that coercion is a graduated rather than a binary phenomenon. This runs counter to what has been the conventional approach to this semantic phenomenon in the field, which considers coercion to be effectively an activation of a rule based process of constraint satisfaction: Pustejovsky (1993) surveys the typical approach to coercion, and indeed to lexical ambiguity in general, as a process of contextualised *selectional restriction* over a set of discreet word senses (though see Lapata and Lascarides, 2003; Shutova et al., 2013, for computational applications of a probabilistic, corpus based model). It seems apparent, though, that some instances of predicate type coercion are less obvious than others. *Hear click* illustrates this point nicely, as we are are forced to pause while we consider whether *click*

categorically denotes NOISE or something more like *process* or even *event*. Furthermore, to borrow an instance from *BriscoeEA1995*, among others, phrases such as *enjoy a book* are clearly coercive and moreover open to interpretation as to the exact mechanism of type shifting: is the book being read or written?

The prodigious use of terms such as *book* in the literature suggests that the axis of concretion and abstraction may be involved in these determinations. To the extent that concrete terms might be understood as relatively low level nodes in a conceptual taxonomy working its way up to the more abstract types, the distance of a word like *motor* from a paradigmatic denotation such as ENTITY could therefore facilitate its coercion into the class NOISE. Rather than modelling this semantic phenomenon as a consequence of a symbolic system of typed representations, however, I have sought to explore the extent to which co-occurrence features can prefigure subsequent high-level decisions about coercion. To the degree that the output of my methodology can be considered a positive result, the finding might be summarised like this: there is evidence here for a lexical semantic model grounded in statistical, interactive, contextually adjustable representations, with categorical commitments regarding the conceptual indices of semantic units emerging from the dynamics of the representation in the course of language use.

6.2.3 Adding Sentential Context

To revisit Pustejovsky and Jezek's (2008) hypothesis regarding the mechanisms of semantic type coercion, the explication of type shifts arises fundamentally in a conceptual and, accordingly, linguistic context. While I have, as discussed in Chapter 3.2, endeavoured to treat context as a more generally cognitive rather than strictly textual phenomenon, one of the appealing things about the data provided by Pustejovsky et al. (2010) is that the word pairs classified in terms of coercion are embedded in sentences, and these sentences do naturally offer the basis for some sort of conceptual handle on the interaction between predicate and object. In this section, I will perform a final experiment on semantic type coercion in which the content of each sentence is provided to my models as an additional input for projecting co-occurrence subspaces in which the relationship between word-vectors and generic vectors can be analysed.

To begin with, I take each sentence and, using the Stanford Parser (Toutanova and Manning, 2000),[8] extract a part-of-speech tag for each word in each sentence in the data. (Note that, now that I'm using the full sentences which provide unique data for every word-pair, I can use full data set of 1,992 sentences.) For each sentence, I group

[8] As provided by the `nltk` package for python.

window		2x2				5x5			
dimensions		20	50	200	400	20	50	200	400
JOINT	nouns	0.157	0.174	0.244	0.283	0.193	0.244	0.257	0.271
	verbs	0.121	0.155	0.190	0.237	0.117	0.163	0.215	0.229
	adjectives	0.083	0.113	0.179	0.187	0.119	0.131	0.183	0.207
	adverbs	0.042	0.091	0.155	0.154	0.101	0.128	0.171	0.174
INDY	nouns	0.092	0.133	0.147	0.157	0.158	0.170	0.148	0.168
	verbs	0.117	0.126	0.173	0.165	0.147	0.209	0.174	0.201
	adjectives	0.123	0.114	0.162	0.172	0.173	0.161	0.151	0.184
	adverbs	0.115	0.137	0.139	0.120	0.167	0.146	0.121	0.111

Table 6-K: F-scores for coercion detection in full featured subspaces based on JOINT and INDY analyses of parts of speech found in each sentence containing a verb-noun pair.

together all the words other than the target verb and object that satisfy one of four grammatical class descriptions: nouns, verbs, adjectives, and adverbs. I then run four different experiments, one treating the words for each of these grammatical classes as the input for dimension selection and then analyse the situation of the word-vectors in the projected subspaces in terms of the now familiar catalogue of geometric statistical features. I treat the output for each grammatical class as the data for a logistic regression, first applying standard mean normalisation and then, setting values for sentences where no words of a particular grammatical class are available to zero, train a model to learn to classify each word pair as coercive or not coercive. Results for each grammatical class, with various dimensional and co-occurrence window parameters, using the JOINT and INDY dimension selection techniques, are reported in Table 6-K.

The scores here are, clearly, low, with the difference between the top score of $f = 0.283$ for nouns in the 2x2 word window, 400 dimensional JOINT subspaces and the minority class baseline of $f = 0.458$ approaching significance at $p = .020$. Beyond that, it is notable that the JOINT subspaces perform substantially better that the INDY subspaces (at $p = .056$ for the 2x2 word window, 400 dimensional noun subspaces). This suggests that a coherent subspace representing the intersection of co-occurrence profiles across the instances of a given grammatical class found in a sentence, rather a subspace merged out of the divergent co-occurrences associated with each word independently, presents a more cohesive basis for the analysis of the conceptual relationship between a predicate and an object. More to the point, though, it seems that sentential context, as far as this analytical technique is concerned, provides only marginal information for contextualising the relationship between a verb and the potentially shifted type of one of its arguments. Turning back to the data, the sentences analysed were extracted from the BNC corpus, and so in many instances have a particularly colloquial or conversational character: it is not clear how much useful additional co-occurrence context could be found in sentences

	precision	*recall*	*f-score*	*accuracy*
INDY	0.708	0.603	0.651	0.805
VERB	0.726	0.610	0.663	0.813
Roberts and Harabagiu (2010)	-	-	-	0.961
Roberts and Harabagiu (2011)	-	-	-	0.812
EBL	0.630	0.498	0.556	0.764
EBL*	0.833	0.690	0.755	0.871
MINORITY CLASS	0.291	1.000	0.458	0.297
MAJORITY CLASS	0.000	0.000	0.000	0.703

Table 6-L: Coercion scores for the best performing context sensitive methods trained and tested on the original data configuration and compared to various other results.

such as, "We will be *denying* the *charges* strongly in court," or, "I mean the *train* could have been *cancelled* or something," or simpy, "He *finished* his *drink*." In the end, then, sentential context could easily be seen as activating other non-linguistic cognitive context that is not built into the type of statistical model explored here.

Finally, Table 6-L presents the outcomes of models tested on the data as originally presented, split into a training and test set each containing instances of all five verb stems. In addition to the top performing INDY and verb-only subspaces, I present results for two different *example based learning* techniques. In the first instance, denoted EBL, a rule is learned for each verb stem type in the training data, based simply on whether that verb is usually observed to take a noun of the specified type as an argument or to coerce the type of its argument, and then that rule is applied across all instances in the test data. In practice this means that forms of *finish* are predicted to be coercive, and all other verbs are predicted to take arguments of the expected type. The enhanced EBL* technique learns a rule for each verb-noun combination, resorting to the EBL rule when it encounters an unobserved pairing in the training data. The very strong results here simply indicate that a large number of combinations are attested in both the training and test data, a peculiarity which I sought to overcome through the rearrangement of the data described above.

I also survey the two results previously reported in the literature here. Roberts and Harabagiu (2011) present a probabilistic model that assigns a likelihood to classes associated with both predicate arguments and nouns based on observations across a large scale corpus and then uses the training set of the SemEval data to learn to identify a threshold of coercion. So, where my models learn something about the way that predicates and objects interact in different co-occurrence contexts, this model seeks to learn something about coercion based on dependency-enhanced distributions. The similar accuracy scores suggest that in the end the two approaches are using different mechanisms to arrive at a

similar sense of coercion outside of sentential context, though it would be interesting to compare precision and recall scores or even line-by-line model output in order to better understand the specific ways that each techniqe makes its determinations. The superior results of Roberts and Harabagiu (2010) are based on a model which specifically extracts predicate argument expectations and noun classes from WordNet and combines these features with a number of other features, combining data driven approaches with knowledge bases. The end result can be thought of as an enhancement on the EBL* technique described above.

Based on the results reported by Roberts and Harabagiu (2011), we might reasonably speculate that building word-vectors based on dependency relationships – for instance, treating the distance between words in a parse tree rather than absolute distance in a string as the boundary condition for co-occurrence window size, as Padó and Lapata (2007) have discussed – might significantly enhance a model's ability to classify coercion. But this would come at the expense of building a model that doesn't have some degree of syntactic commitment already built into it, and it is likewise easy to imagine how such an approach would open itself up to accusations of tautology: if coercion as a binary case is a grammatical abstraction, then such a model would be to some extent recapitulating the premise of coercion data structuring. Of greater interest would be, in line with the degrees or axes of metaphoricity briefly explored in Section **??**, establishing data indicating novel or exceptional cases of coercion, or alternatively of coercion scored along a continuous scale. This might require a reworking of the standard dogma of semantic type shifting, but then one of the objectives of my methodology is to provide a computational framework for exploring and possibly shifting the lines of theories about language and meaning.

6.3 Interpretation and Composition in Context

One of the tricky things about figurative language is its ephemerality: if we stare at it for long enough through a theoretical lens, it seems to vanish, as is evident in the deflationary case made by Sperber and Wilson (2012). But on the other hand, if we ask someone in street whether the phase *buy a story* is more metaphoric than *buy a book*, we can reasonably expect the answer will almost always be "yes", and it would be a mistake to dismiss the evidence that in a colloquial sense some compositions are clearly metaphoric, and others are clearly not. This raises a challenging point with regard to the comparison between metaphor and coercion, the two instances of figurative language explored in this chapter: is metaphor perhaps to some extent a more overt case of coercion, or maybe a specific case that is in some way or another a little more subtle? Part of the problem here

is that the distinctions between these phenomena begin to exceed the capacity for what can reliably be quantified about language in a clinical setting, with evaluative criteria that will depend on the opinion of an expert which comes pre-packaged with inevitable biases.

The experiments presented in this chapter have focused on the classification of non-literal language: the simple task of determining whether the way that a set of words are used pertains to some encyclopaedic sense of their lexical semantic role, without regard to the explication of any sort of interpretation of how semantics are subverted or what the metaphor communicates. But Shutova (2010) has made the case that, in a cognitively plausible sense, metaphor classification should be seen precisely as metaphor interpretation, and this theoretical stance has been backed up by a data-driven computational model that involves classification of metaphor by way of a round-trip paraphrasing technique. This in turn invites a consideration of the entanglement between the interpretation and composition of figurative language: metaphor and metonymy are things that, in some particular conceptual context, are done by one lexical entity to another, as the apt term *coercion* would itself suggest. If composition happens in some cognitive context, then interpretation presumably involves the identification or simulation of that context, as the relevance theoretical account of metaphor surveyed in Chapter 2.3. So this is in line with Carston's (2010) description of how metaphor involves the generation of an *ad hoc* concept pertaining to the semantics of the shifted lexeme in the specific context of a particular linguistic encounter.

In fact, it is tempting to go so far as to say that figurative language is identified precisely as those instances of language where recourse to a conceptual context is necessary to interpret a lexical composition, and furthermore that the degree of figurativeness correlates with the extent of context construction involved in an interpretation. If this postulate holds water, it means that figurative language is arguably about the perception of contextuality in the semantic complex that fulgurates between words and ideas as much as it is about the alignment of properties between conceptual domains. Under this theoretical regime, my context sensitive methodology should offer a good framework for identifying figurative language, because models built using this methodology can be indicative of the extent to which contextualisation draws generic or encyclopaedic representations into alignment with one another. This hypothesis is supported by the evidence from the metaphor experiments in Section **??**, where strong results for both metaphor classification and the translation of metaphor classification models to the task of rating metaphoricity on a continuous scale indicate that the relationship between lexical semantic concepts in context sensitive spaces does tell us something about the degree to which *ad hoc* concepts are being extrapolated in the course of composing the input terms.

The lower numbers associated with the coercion results described in Section **??**, and particularly to the detection of predicate argument coercion involving unattested verbs, are to some extent a natural product of the nature of the data, given that the data offered by Pustejovsky et al. (2010) presents coercion as a minority case. It's also worth considering, however, the way in which the theoretical framing of coercion seeks to impose a fairly rigid structure on the conceptual hierarchy inherent in lexical typing. The presumption is broadly, in the specific case of predicate argument shifting examined here, that a given verb takes a singular class as its argument, and that noun sense are likewise defined in terms of distinct classes, and so coercion is in principle understood as a discretely binary phenomenon. But there are clearly issues of hierarchy and level of abstraction at stake in making class distinctions with regard to the denotations of nouns, and in fact there are some instances of coercion – "cancel a train", for instance, or "hear a click" – where the conceptual shift, while perhaps strictly present according to a particular categorical lexicon, does not involve a significant conceptual reconstruction by way of contextualisation.

This, then, raises a valid question: is the role of figurative language exclusively, or even for that matter primarily, to port attributes from one conceptual domain to another? Or is what metaphor does, as Davidson (1978) has famously suggested, really about something more fundamentally phenomenological than just the efficient transmission of propositions? So, where, for instance, Sweetser (1990) sees polysemy as an intermediate stage bridging the progress from literal to metaphoric usage, my methodology leaves itself open to the possibility that all usage is, in fact, first and foremost pragmatic, and only secondarily lexicalised. By this interpretation, words have semantic affordances in terms of their potential to convey cognitive content intersubjectively, and they are picked up and used in much the same way that a cognitive agent might adapt an object designed or just perceived as being for one purpose as an implement in another activity—using a shoe as a hammer, for example, or a chair to fend off a lion. The cognitive foregrounding of this nascent theory can be found in the ecological psychology of Bateson (1972) and Gibson (1979), and the linguistic correlary seems to be in line with what psycholinguists inspired by biosemiotics such as Rączaszek-Leonardi and Nomikou (2015) are saying about the way that language is primarily about affording cognitive value to interlocutors, including but hardly limited to truth values.

This theoretical speculation is a potential extrapolation of my methodology rather than a precondition for it, and is offered primarily as an example of how this statistical approach might become a component of productive line of philosophical enquiry. The point, though, is that with a geometric methodology, relationships between lexical semantic representations can be recast as Gibsonian affordances: there is a mechanism for the

direct perception of opportunities for meaning making in the actual layout of the statistical environment. Meaning is, then, something which is directly perceived and acted upon, in the sense that a geometric mode of representation allows us, in an abstract sense, to imagine how an agent might participate in symbolically grounded semantic activity without resorting to the computation of the conceptual transpositions involved in the analysis of figurative language. By situating semantic representations in situationally induced spaces, the probabilisitic contingencies between words in particular contexts can in principle be mapped directly to the perception of an opportunity for meaning making without the need to resort to an interposing layer of conceptual computation. It would be a mistake to go any further than this in terms of arguing for the cognitive plausibility of a model that is fundamentally statistical and computational, but, inasmuch as it it a desirable thing to consider the possibility that the prolific use of metaphor in the course of ordinary linguistic communication is an artefact of a more general cognitive propensity to grab whatever is available in an environment and make use of it, the methodology described here seems like an acceptable starting point for further investigation.

Chapter 7

The Geometry of Conceptualisation: Analogies

In this chapter, as a final empirical investigation into the potentialities of context specific distributional semantic techniques, I will investigate the capacity of my methodologies to model analogy. For the purposes of the computational and geometric modelling of semantics, analogy can be seen as a kind of meta-phenomenon: an analogical equation involving two sets of lexical representations indicates that there is some underlying intensionality that conceptual binds the denotations of the representations. So, for instance, the metaphor "that surgeon is a butcher" can be extended through a mapping between the conceptual domains of SURGERY and BUTCHERY to arrive at semantic formulae such as *surgeon:scalpel :: butcher:cleaver* or hospital:patient :: abattoir:carcass. Furthermore, if these relationships can be mapped geometrically in a semantic space, then we should have on our hands a productive mechanism for configuring a general semantic model—and one which may overcome some of the issues of interpretation and composition raised in the last chapter. If we can connect a general region of butchery to a region of surgery in a semantic space, for instance, then we might be able to extrapolate such metaphoric turns of speech as "the surgeon hacked at me with her scalpel" from a model without committing to the claim that the model (or, for that matter, and agent) has actually interpreted the metaphor in an online way.

The idea that there is a geometric component to analogy is at least hinted at by Tversky (1977), who, as discussed in Chapter 5.4, raises the issue of inequalities and asymmetries in relationships of synonymy. Gentner (1983) extends Tversky's insights to a model explicitly targeting analogy through the application of isomorphic *structure mappings* that identify congruities between conceptual domains based on composite symbolic

representations. From a computational perspective, Veale and Keane (1992) describe a system that functions through a series of *spatial operators* which facilitate mappings between conceptual domains by way of a schema of collocations, containments, and orientations, though these operations do not involve the instantiation of Euclidean measures. Subsequent symbolic computational models of metaphor in particular have seized on the mechanism of modelling mappings between conceptual structures that are, to a greater or lesser extent, based on the identification of congruities and a corresponding geometrical logic of sorts, and a small sample of work in the field has been surveyed in Chapter 2.3.

The empirical work described here will, naturally, focus on a statistical rather than symbolic approach to modelling analogy by way of spatial mappings between domains—and, in this case, domains, in the spirit of Gärdenfors (2000), are represented roughly as regions in a Euclidean space. It is important to note, though, that one of the primary components of the productive symbolic approaches to analogy mentioned above goes away once we move into distributional semantic spaces: where the features of symbolic representations are generally constructed to be interpreted as actual attributes of the denotations being modelled, the dimensions of distributional semantic spaces are simply indices to information about co-occurrences observed in a digital corpus (this has already been discussed in Chapter 3.3 in the context of Rimell's (2014) work studying the relationship between co-occurrence overlap and entailment, and again in Chapter 3.4 by way of Derrac and Schockaert's (2015) model treating directions in factorised distributional spaces as conceptual themes). So there is a trade-off between access to a continuous Euclidean space of lexical semantic representations with geometric measures facilitated by the statistical nature of the representation building process and the loss of interpretable features in a symbolic conceptual scheme. My hypothesis here, in line with experiments described in the previous two chapters, is that a process of contextualisation can generate spaces where collections of co-occurrence dimensions representing conceptually oriented profiles of language use will provide an appropriate ground for modelling analogy in terms of rigorous Euclidean relationships. And in the case of analogy in particular, as will be seen in the following section, there is already a body of work offering compelling evidence that distributional semantic statistics can map conceptual relationships onto the geometry of word co-occurrence.

7.1 Analogies as Parallel Vectors

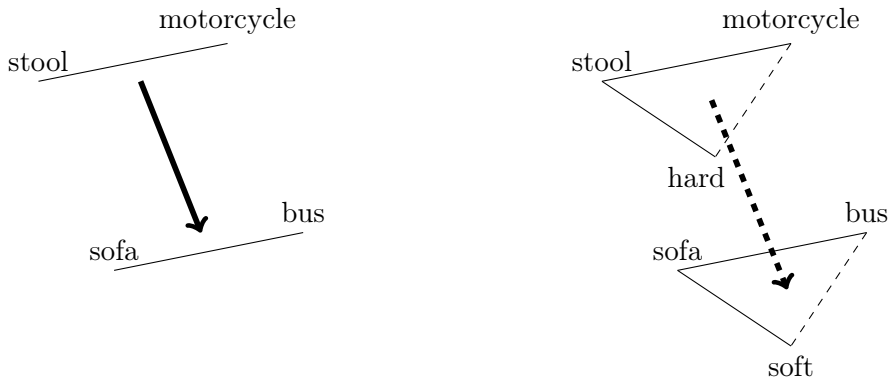
The `word2vec` distributional semantic modelling techniques, which have served as a point of comparison and discussion throughout this thesis, was originally presented with a test set of 19,544 four-word analogies, constructed by the model artchitects and devised to

cover a range of relationships which the designers categorised as broadly *semantic* or *syntactic* (Mikolov et al., 2013a,1).[1] So, for instance, the data presents relationships such as, on the one hand, *Bangkok:Thailand :: Paris:France* or *boy:girl :: man:woman*, and, on the other hand, *calm:calmly :: lucky:luckily* or *aware:unaware :: efficient:inefficient.* The task involves feeding a semantic model the first three terms and then measuring the rate at which it is able to accurately predict the fourth term.

The neural network architecture of the `word2vec` approach produces remarkably strong results on this task through the application of a simple geometric device. Within the normalised space of word-vectors generated over the course of iterative traversals of a large-scale digital corpus, given an unfinished analogy of form $A : B :: C : X$, the model simply finds the vector $\overrightarrow{x}$ most closely fulfilling the equation $\overrightarrow{b} + \overrightarrow{c} - \overrightarrow{a} \approx \overrightarrow{x}$, where $\overrightarrow{a}$, $\overrightarrow{b}$, *and* $\overrightarrow{c}$ are the word-vectors corresponding to the three known elements of the analogy, and returns the vocabulary word associated with $\overrightarrow{x}$. The original literature reports an accuracy rate of 0.61 for the CBoW model, which is all the more impressive when we consider how many ways there are to choose the wrong solution to an analogy from a vocabulary of one million words. (It should be noted that similarly strong results have been reported for the hybrid frequentist-neural model of Pennington et al., 2014, .)

But the really remarkable thing about these results is that the models build these spaces in a completely unsupervised manner with respect to the actual task of analogy solution. This means that the arrangement of word-vectors plays out in a tidy conceptual geometry, interpretable through simple linear algebraic operations, simply by virtue of the way that words tend to come up in proximity to one another in the course of colloquial written language use (the original results were obtained from models trained on the Google News Corpus, and, as will be seen below, the same models trained on Wikipedia achieve comparable scores). Much has been made of this: Levy and Goldberg (2014a) postulate about the procedural equivalence of iterative and statistical models mitigated by parameterisation issues, while AroraEA2016 have attempted to explain mathematically how the application of a random walk type algorithm to statistical models results in a recapitulation of the strong neural network results. At the time of writing, there is a generally accepted intuition afoot in the field that, along the lines of the distributional hypothesis itself, it makes sense that the gradual nudging of word-vectors by a neural network based on observations of co-occurrences should push words into situations where orientations and distances in space broadly map to conceptual relationships between representations; there is not, however, a well-formed mathematical explanation of why these techniques are so effective at projecting semantic relationships into space. At any rate,

[1] The analogy data is included in the package that can be downloaded at `https://code.google.com/archive/p/word2vec/source/default/source`.



(a) This makes sense. (b) Does this make sense?

Figure 7.1: The analogical components of overlapping conceptual frames do not necessarily map neatly into a singular space.

the import of all this is that, in `word2vec` type distributional semantic spaces, at least a certain type of analogy plays out along the lines illustrated in Figure 7.1a, as a close approximation of a parallelogram at the surface of a normalised hypersphere in a space delineated by abstract dimensions acting as handles for the backpropagating action of a neural network.

So, to put it in plain language, the line between two points representing a conceptual relationship in one region of the space should be parallel to and in the same direction as two points in another region representing an analogous conceptual relationship under a different overall conceptual scheme. There is, however, an objection to be raised here. Returning once again to Tversky's (1977) observations about the asymmetry of similarity, there are problems once we begin to extend the geometry of analogy to more complex conceptual structures, as illustrated in Figure 7.1b: for any given mapping between anything other than the most trivial conceptual domains, there will be some intensional component of the denotation that does not conform to the presumed isomorphism of the analogy in a distributional semantic space. So, while it may be possible to map relatively atomic elements analogically in a space of fixed semantic points, it seems there will always be a breakdown when it comes to trying to discover isomorphisms between entire domains.

Chen et al. (2017) have noted geometrical anomalies specifically in the context of distributional semantics, demonstrating that the presumption of conceptual parallelism is considerably more consistent in some domains than others, and further using clinical studies on human respondents to feel out some of the disconnects in analogical chains inherent in these types of semantic models. What becomes clear is that static distributional models are very effective at providing a semantically productive geometry some of

the time, but they lack the adaptability that is fundamental to environmentally situated cognition and so do not make the open-ended kind of connections between words and concepts that are characteristic of semantics. What is necessary is an element of flexibility, and I propose that my methodology for contextualising semantic spaces offers the proper kind of framework for providing the required openness to

7.2 Contextualising Analogical Geometry

In this section, I will explore the distributional contexts of analogical semantic relationships. The premise of this investigation arises from the disconnect illustrated in Figure 7.1: it seems impossible to imagine how a static semantic space could consistently represent analogies as well-formed geometric entities. Rather, I maintain that analogy is always to at least a certain extent context specific. Following on this, my hypothesis is that there should always be some contextualisation which permits the satisfactory mapping of an analogy in a semantic space. In the case of a simple four word analogy, which will be the focus of the work presented here, this means that an analogy is modelled as a parallelogram, with each of the word-vectors denoting the components of the analogy as a vertex, to a degree of precision that precludes any other word-vector as being mistaken as a component of the lexical-conceptual complex.

The empirical work described here will proceed initially with a strategy of reverse engineering of sorts, seeking to validate the possibility of discovering the kind of spaces that we would like to find for mapping analogies through contextualising operations on distributional semantic spaces delineated by literal co-occurrence dimensions. This will lead on to an examination of some of the ways that context sensitive approaches might be applied to an analogy completion task, though, not surprisingly, it turns out to be considerably easier to find a space where an analogy works out as expected than to discover an analogy in a sizeable state space of possible subspaces. In the end, this last empirical component to my research, which expands upon work originally presented in McGregor et al. (2016) will hopefully serve as an incipient to further research in terms of the potentialities and capacities of context sensitive distributional semantics.

7.2.1 Projecting Probability into Space

Before we engage with an exploration of the analogical potential of context specific subspaces, a brief review of the mathematics of distributional semantic spaces with literal co-occurrence dimensions will serve to reinforce the connection between the geometry of

analogy and the probabilistic grounding of my methodology. Returning to the definition of a co-occurrence statistic outlined in Chapter 4.4, recall that the pointwise mutual information between a word w and a co-occurrence term c is the unexpectedness associated with an observation of c in proximity to w, which can be expressed in terms of joint and compound probabilities (and the equation is approximate because we're ignoring the skewing factor of 1 and the smoothing constant described in Chapter 4.1:

$$PMI(w, c) \approx \log\left(\frac{p(w, c)}{p(w) \times p(c)}\right) \tag{7.1}$$

The basic assumption of the geometric approach to analogy, meanwhile, is that the components of an analogy map into a parallelogram sitting in some askance situation in a high dimensional space, a state of affairs which can be expressed using linear algebraic terms for a suppositional analogy $A : B :: C : D$ and the corresponding word-vectors:

$$\overrightarrow{a} - \overrightarrow{b} \approx \overrightarrow{c} - \overrightarrow{d} \tag{7.2}$$

For any arbitrary dimension i, this can then be reduced to a difference between logs:

$$\log\left(\frac{p(a, i)}{p(a) \times p(i)}\right) - \log\left(\frac{p(b, i)}{p(b) \times p(i)}\right) \approx \log\left(\frac{p(c, i)}{p(c) \times p(i)}\right) - \log\left(\frac{p(d, i)}{d(a) \times p(i)}\right) \tag{7.3}$$

This expression can be significantly reduced by merging the arguments of the logarithms on either side of the equation into ratios and then dropping the logs:

$$\frac{p(a, i) \times p(b)}{p(b, i) \times p(a)} \approx \frac{p(c, i) \times p(d)}{p(d, i) \times p(c)} \tag{7.4}$$

Or, converting the ratio of joint and independent probabilities to conditional probabilities and with a little more algebra:

$$p(i|a, i|d) \approx p(i|b, i|c) \tag{7.5}$$

To again state this plainly, our analogy optimisation function seeks to find those dimensions where the combined probability of observing a given co-occurrence term in the context of A and also (though not necessarily simultaneously) D is as close as possible

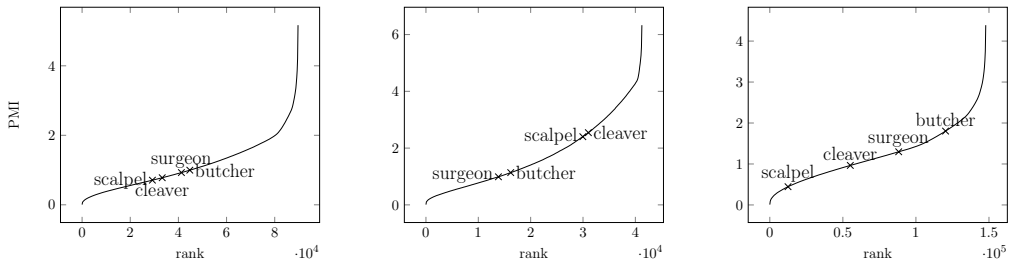
to observing the same term in the contexts of B and C. So for instance we are interested in discovering a dimension that is as likely to occur in a context of *surgeon* and *cleaver* as it is to occur in a context of *butcher* and *scalpel*. If we can discover a profile of such dimensions, then we can productively map this particular analogy onto a contextualised distributional semantic space.

This property of semantic spaces defined in information theoretical terms is an artefact of the conversion of products and ratios into sums and differences through the mechanism of logarithmic functions. To coin a term, logarithms *geometrise* a space of probabilistic statistics, allowing us to perform operations on shapes in Euclidean space that correspond to hypotheses about joint and conditional observations of events, in this case co-occurrence events in a large scale corpus. It must be emphasised, however, that the interpretability of probabilities in geometric terms only holds in spaces where dimensions still map to literal co-occurrence statistics, and so this property is a feature of my methodology but not of semantic spaces that have been factorised or learned through the abstract operations of a neural network. The next objective, then, is to search for the appropriate techniques for specifying a context in order to map out a given analogy.

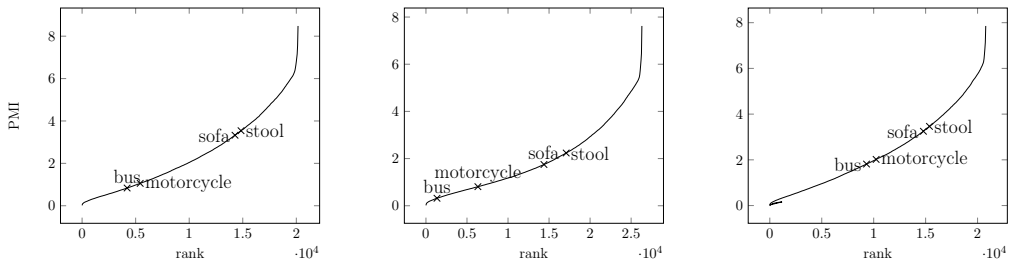
7.2.2 Finding Contexts for Analogies

I next investigate whether or not co-occurrence dimensions satisfying the conditions laid out above can be discovered in co-occurrence subspaces contextualised using the methods developed and explored throughout this thesis. In particular we are interested in discovering the dimensions which most closely satisfy the equation $(\overrightarrow{a} - \overrightarrow{b}) - (\overrightarrow{c} - \overrightarrow{d}) = 0$ for the word-vectors corresponding to the components of the analogy $A : B :: C : D$. This relationship can be examined on a dimenion-by-dimension basis, beginning by extracting dimensions that are known to have non-zero values for some or all of the word-vectors involved in the analogy. Figure 7.2 presents a histographic analysis of just such an analysis for two different analogies: the aforementioned *surgeon:scalpel :: butcher:cleaver*, representing the frequently discussed conceptual mapping from SURGERY to BUTCHERY, and, from Figure 7.1, *stool:sofa :: motorcycle:bus*, indicating a mapping from FURNITURE to VEHICLES. In both cases the best three dimensions for satisfying the balance of values indicating parallel relationships between the legs of the analogy are selected from the top 400 dimensions projected from 5x5 word co-occurrence window based spaces taking the first three of the four components of each analogy as input to the ZIPPED methodology.

What stands out here is the way that the analogical word-vectors tend to cluster into pairs. This makes sense, since the formula described above indicates instances where the relationship between two of the word vectors is very similar to the relationship between



(a) SURGERY ⟺ BUTCHERY

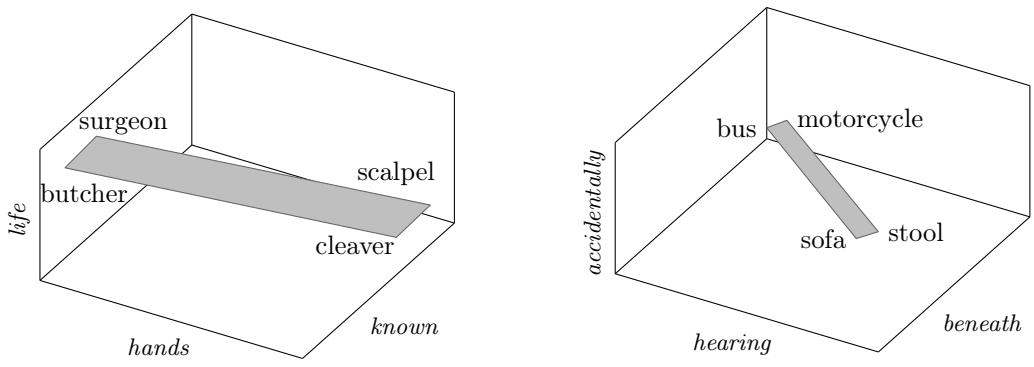


(b) FURNITURE ⟺ VEHICLES

Figure 7.2: Histograms of the top three co-occurrence dimensions satisfying the expected arithmetic of analogy.

the other two: this requirement is nicely satisfied with pairing word-vectors up with one another. These dimensional values are pushed into two-dimensional projections of three-dimensional spaces in Figure 7.3, and here the well-defined parallelograms expected from this method of dimension selection become apparent. In fact, more than just parallelograms, the shapes that begin to emerge are specifically rectangular in nature. If we imagine extending the process of selecting dimensions where target word-vectors are clustered into pairs into higher dimensional subspaces, we can see that the vertices of the shapes that would evolve would tend towards right angles, and so this indicates an additional geometric feature of the relationships between lexical semantic representations that we might associate with analogy.

Another interesting thing to note about the configurations in Figure 7.3 is the oblong nature of the shapes. In fact, it seems as though the word-vectors are orientating themselves in terms of types—though not necessarily in alignment with the conceptual categories delineating the analogical mappings. So, where *bus* and *motorcycle* might be seen as occupying a VEHICLE extent of the subspace as opposed to the FURNITURE region inhabited by *sofa* and *stool*, *surgeon* and *butcher* seem to be establishing a region of PROFESSIONS while *scalpel* and *cleaver* could be construed in a domain of IMPLEMENTS.



(a) SURGERY ⟺ BUTCHERY (b) FURNITURE ⟺ VEHICLES

Figure 7.3: The geometry of two analogies projected into subspaces defined by the three most analogically accurate dimensions.

These distinctions are, naturally, a peculiarity of the dimensions themselves, with *hands* in particular specifying a high co-occurrence with IMPLEMENTS, while the preposition *beneath*, with its spatial intimations, remits high values for *sofa* and *stool* (denotations of things that other things can be beneath); the prevalence of these same terms in co-occurrences with *hearing* is less obvious but nonetheless indicative.

One of the things to take away from this small-scale qualitative analysis is that, at the end of the day, any four-point analogy can be cut along at least two different conceptual axes, corresponding to the intensions that semantically bind the representations along each edge of the rectangle. We might easily speculate that the shapes found in well-formed analogical subspaces will be elongated along the axes that correspond to what humans might tend to classify as the conceptually salient distinction inherent in the analogy, but there are presumably also a variety of ways to make an orthogonal distinction, and we can reasonably expect these secondary characteristics of the analogical relationship to emerge in higher dimensional spaces. In fact, a reasonable hypothesis, raised in McGregor et al. (2016), is that contextualised analogical constructs should, as dimensionality increases, begin to assume a more sqaure shape and a more central position in a subspace.

So I think it makes sense to expand this approach to analogy modelling to cover more analogies, and to examine the way that higher dimensionalities provide a basis for geometric analysis. In order to efficiently and systematically test the viability of context sensitive subspaces for analogy solution, I randomly select a subset of the data, described above, designed by Mikolov et al. (2013a) and project 400 dimensional subspaces from both 2x2 and 5x5 word window base spaces using the JOINT, INDY, and ZIPPED techniques, taking, with an eye towards an analogy solving methodology, only the first three of the four words in each analogy as input. Following this step, and as with the examples mentioned

dimensions		5	10	20	50	100	200
2x2	JOINT	0.911	0.972	0.989	0.986	0.970	0.916
	INDY	0.722	0.908	0.976	0.985	0.967	0.873
	ZIPPED	0.921	0.975	0.991	0.987	0.970	0.919
5x5	JOINT	0.941	0.987	0.996	0.997	0.995	0.957
	INDY	0.697	0.908	0.973	0.984	0.962	0.895
	ZIPPED	0.934	0.987	0.999	0.998	0.997	0.968

Table 7-A: Accuracy rates for solving analogies when choosing subsets of optimal dimensions from 400 dimensional subspaces picked taking the first three elements of each analogy as input.

above, this experiment becomes an instance of what we might call space-fitting: finding the subspace derived from three of the four analogical terms that is expected to most appropriately fulfil our geometric expectations, the testing, based on full information about the analogy, the degree to which the space does in fact fit the shape. In each of the six resulting subspaces (two base spaces by three dimension selection techniques), I rank dimensions in order of their proximity to satisfying the equivalence relationship between legs of the analogy. I think explore analogical accuracy as a function of various dimensional threshold levels, considering an analogy to be accurately solved if the label of the word-vector that most closely satisfies $\overrightarrow{x} = \overrightarrow{b} + \overrightarrow{c} - \overrightarrow{a}$ corresponds to D in $A : B :: C : D$.

Results for this experiment are reported in Table 7-A, with accuracy scores given for subspaces of 5, 10, 20, 50, 100, and 200 dimensions. In the case of both the JOINT and ZIPPED techniques, the chances of finding a satisfactory subspace are strong across the board. This means that, on the one hand, it should be possible to pick as few as the right five dimensions out of a set of 400 and still find a subspace where more than 90% of the analogies in this sampled dataset are accurately modelled, and, on the other hand, there is a way to cut the set of 400 dimensions picked without knowledge of the fourth component of an analogy in half and get find likewise reliably productive geometries. The INDY technique doesn't do quite as well here, particularly at the lower dimensionalities where there is presumably less of a chance of finding many non-zero values for all the components of an analogy along co-occurrence dimensions that might have achieved high scores for a single input term independently in part by way of being infrequent and perhaps specialised. And of course, there are quite a few ways to pick either five or 200 out of a set of 400, so we do not yet have an analogy solving or generating engine on our hands.

This last point leads to a further question: what if there is some way, given the vast combinatory spaces of dimensional subsampling available here, to solve more or less *any*

dimensions		5	10	20	50	100	200
2x2	JOINT	0.654	0.814	0.896	0.930	0.881	0.466
	INDY	0.115	0.234	0.341	0.369	0.267	0.045
	ZIPPED	0.616	0.806	0.892	0.929	0.887	0.489
5x5	JOINT	0.657	0.828	0.901	0.921	0.835	0.402
	INDY	0.129	0.253	0.338	0.384	0.277	0.051
	ZIPPED	0.589	0.790	0.888	0.915	0.876	0.418

Table 7-B: Accuracy rates for solving randomly completed analogies when choosing subsets of optimal dimensions from 400 dimensional subspaces picked taking the first three elements of each analogy as input.

version of an analogy? If the contextualisation process is so open ended that we can geometrically constrct more or less any conceivable semantic relationship, then the first step of the contextualisation process, in which only part of the analogy is used to generate a subspace from which subsequent fully informed selection are made, doesn't really get us anything at all in the way of using three points of an analogy to find the appropriate context for discovering the fourth point. With this in mind, I rearrange the 1,000 analogy sample of the data used to generate the results in Table 7-A such that the fourth component of each analogy is randomly selected from all possible fourth components across the list. Table 7-B reports results for selecting lower dimensional subspaces expected to solve these random analogies, applying the same procedure as described above for identifying optimal dimensions and then testing at various dimensionalities.

On the one hand, these results are impressively – even surprisingly – good. It turns out, for instance, that there is some set of 50 dimensions to be selected from the 400 dimensional subspace projected by applying the JOINT technique to the inputs

XXX

that solves the unlikely analogy

XXX

On the other hand, though, these scores are substantially lower than those reported for established analogies in Table 7-A. This is particularly the case for higher dimensionalities, where the options for discovering a multitude of dimensions facilitating the mapping of a randomly generated analogy evidently become confounded, and the difference is greatest of all for relatively large sets of dimensions chosen from the INDY subspaces. So it would seem that the overlap between independently selected co-occurrence dimensions is actually indicative of some degree of conceptual coherence after all, evidenced by the relative likelihood of solving an attested analogy versus a random one. (It's also interest-

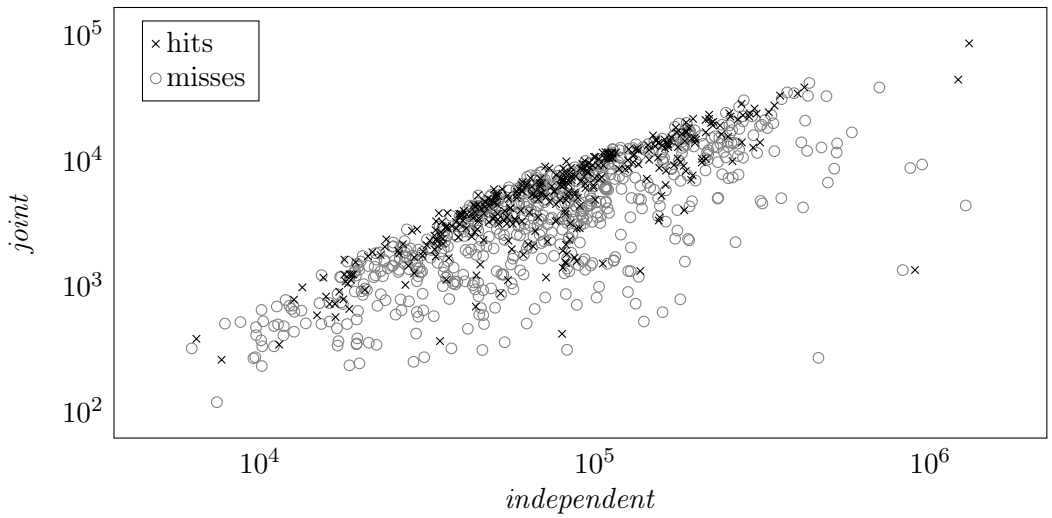


Figure 7.4: A scatter plot of hits and misses for analogy solution as functions of the number of jointly non-zero dimensions versus independently non-zero dimensions, with both axes scaled logarithmically.

ing to note that subspaces derived from smaller co-occurrence windows are more apt to yield what we might call forgiving analogical options for the randomly generated data, indicating once more that there is more conceptual association in the syntagmatics available in a wider co-occurrence window, as discussed in the context of relatedness versus similarity in Chapter 5.1.1.)

7.2.3 Searching for Solutions to Analogies

Having established that there are in principle analogically productive subspaces to be discovered based on taking part of an analogy as input to a context sensitive distributional semantic model, I now explore the capacity of my methodology for completing partial analogies.

It's not particularly surprising that there is a fairly strong positive correlation between the total number of jointly non-zero dimensions and the number of independent non-zero dimensions: word-vectors with more non-zero dimensions are more likely to have an overlap with other word-vectors. But the sharpness of definition of the upper boundary of this relationship is notable, with a dense clustering of both positive and negative results with a relatively high overlap compared to relatively low forming a distinct ridge along the upper left boundary of the distribution. The implication here is a long tail of increasingly disconnected word sets drifting off to the lower right hand of the plot: this could mean that there are mismatches between the co-occurrence profiles of the input

terms as well as the relative frequencies of the terms.

7.3 A Note on the Data

It must be mentioned that the data that has been analysed in this chapter is of a very specific character. The analogies put together by the team at Google are populated by a high percentage of proper names, in particular place names and also currencies, demonyms, and the like. This belies a particular view of language and indeed cognition which is at odds with the premise motivating the model described in this thesis, as outlined at the beginning of Chapter 3. Proper names are, as Russell (1905) has pointed out, particular kinds of words with peculiar denotational properties in that they refer to specific and unique entities or correspondingly specific classes of entities. This is not to say that they do not admit ambiguity – *Paris* is the name of, among other things, a classical character, and *Berlin* the name of a 1980s new wave band – but there tends to be a certain clarity of intent when these types of words are used. These types of analogies are exemplary of cases where language coalesces into a relatively stable conceptual representation, and, notwithstanding cases of polysemy, it's arguably not particularly surprising that these relationships emerge as commensurable directions in a likewise stable representational space.

Furthermore, it is telling that the designers of the dataset have chosen to refer to the variety of analogy typified by *Denmark:Danish :: China:Chinese* as *syntactic*, whereas the relationships denoted by *grandfather:grandmother :: grandson:granddaughter* is considered *semantic*.

References

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

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

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

Agres, K. R., McGregor, S., Rataj, K., Purver, M., and Wiggins, G. A. (2016). Modeling metaphor perception with distributional semantics vector space models. In *Workshop on Computational Creativity, Concept Invention, and General Intelligence*.

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

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

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

Barnden, J. A. and Lee, M. G. (1999). An implemented context system that combines belief reasoning, metaphor-based reasoning and uncertainty handling. In *Modeling and Using Context: Second International and Interdisciplinary Conference*, pages 28–41.

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

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

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

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for distributional semantics. *Computational Linguistics*, 36(4).

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193.

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

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

Barsalou, L. W. (1993). Flexibility, structure, and linguistic vagary in concepts: manifestations of a compositional system of perceptual symbols. In Collins, A., Gathercole, S., and Conway, M., editors, *Theories of memory*, pages 29–101. Lawrence Erlbaum Associates, London.

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

Barwise, J. and Perry, J. (1983). *Situations and Attitudes*. MIT Press, Cambridge, MA.

Basharin, G. P., Langville, A. N., and Naumov, V. A. (2004). The life and work of a.a. markov. *Linear Algebra and its Applications*, 386:3–26. Special Issue on the Conference on the Numerical Solution of Markov Chains 2003.

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

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

Birke, J. and Sarkar, A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL-06*, pages 329–336.

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

Black, M. (1955). Metaphor. In *Proceedings of the Aristotelian Society*, volume 55, pages 273–294.

Black, M. (1977). More about metaphor. In Ortony, A., editor, *Metaphor and Thought*, pages 19–41. Cambridge University Press, 2nd edition.

Boden, M. A. (1990). *The Creative Mind: Myths and Mechanisms*. Weidenfeld and Nicolson, London.

Boden, M. A. (2006). *Mind as Machine: A History of Cognitive Science*. Clarendon, Oxford.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 1st edition.

Bouveret, M. and Sweetser, E. (2009). Multi-frame semantics, metaphoric extensions and grammar. *Annual Meeting of the Berkeley Linguistics Society*, 35(1):49–59.

Brentano, F. (1974/1995). *Psychology from an Empirical Standpoint*. Routledge, London. Translated by Antos C. Rancurello and D. B. Terrell and Linda L. McAlister.

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

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

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

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

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

Carnap, R. and Bar-Hillel, Y. (1952). An outline of a theory of semantic information. Technical Report 247, Research Laboratory of Electronics, MIT.

Carston, R. (2010). Metaphor: Ad hoc concepts, literal meaning and mental images. 110(3):297–323.

Carston, R. (2012). Metaphor and the literal/nonliteral distinction. In Allan, K. and Jaszczolt, K. M., editors, *The Cambridge Handbook of Pragmatics*, pages 469–492. Cambridge University Press.

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.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford University Press.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. The MIT Press, Cambridge, MA.

Chen, D., Peterson, J. C., and Griffiths, T. L. (2017). Evaluating vector-space models of analogy. In *39th Annual Conference of the Cognitive Science Society*.

Chomsky, N. (1959). A review of b. f. skinner's verbal behavior. *Language*, 35(1):26–58.

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

Church, A. (1940). A formulation of the simple theory of types. *Journal of Symbolic Logic*, 5(2):56–68.

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

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive*

Sciences, 10(8):370–374.

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

Coecke, B., Sadrzadeh, M., and Clark, S. (2011). Mathematical foundations for a compositional distributed model of meaning. *Linguistic Analysis*, 36(1-4):345–384.

Coeckelbergh, M. (2016). Can machines create art? *Philosophy & Technology*.

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

Colton, S. (2008). Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium on Creative Intelligent Systems*.

Colton, S., Cook, M., Hepworth, R., and Pease, A. (2014). On acid drops and teardrops: Observer issues in computational creativity. In Kibble, R., editor, *Proceedings of the 50th Anniversary Convention of the AISB*.

Copestake, A. and Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67.

Croft, W. and Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge University Press.

Davidson, D. (1974). On the very idea of a conceptual scheme. In *Proceedings and Addresses of the American Philosophical Association*, volume 47, pages 5–20.

Davidson, D. (1978). What metaphors mean. In *Inquiries into Truth and Interpretation*. Clarendon Press, Oxford, 2nd edition.

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

Deacon, T. W. (2011). *Incomplete Nature: How Mind Emerged from Matter*. W. W. Norton & Company, New York, NY.

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.

Dennett, D. C. (1991). *Consciousness Explained*. The Penguin Press, London.

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

Descartes, R. (1641/1911). *The Philosophical Works of Descartes*. Cambridge University Press. Translated by Elizabeth S. Haldane.

dos Santos, C. and Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, Dublin, Ireland.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. CSLI Publications.

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

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

Dunn, J. (2013). Evaluating the premises and results of four metaphor identification systems. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I*, pages 471–486. Springer Berlin Heidelberg.

Eco, U. (1976). *A Theory of Semiotics*. Indiana University Press, Bloomington.

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

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

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

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

Everett, D. L. (2005). Cultural constraints on grammar and cognition in pirahã: Another look at the design features of human language. *Current Anthropology*, 46(4):621–646.

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

Fass, D. (1991). Met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Fauconnier, G. and Turner, M. (1998). Conceptual integration networks. *Cognitive Science*, 22(2):133–187.

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

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

Firth, J. R. (1959). A synopsis of linguistic theory, 1930-55. In Palmer, F. R., editor, *Selected Papers of J. R. Firth 1952-59*. Indiana University Press.

Floridi, L. (2011). *The Philosophy of Information*. Oxford University Press.

Fodor, J. (2001). *The Mind Doesn't Work that Way: The Scope and Limits of Computational Psychology*. MIT Press.

Fodor, J. A. (2008). *LOT 2: The Language of Thought Revised*. Oxford University Press.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Fox, C. and Lappin, S. (2005). *Foundations of Intensional Semantics*. Blackwell

Publishing, Oxford.

Fraser, B. (1993). *Interpretation of novel metaphors*, pages 307–341. Cambridge University Press, 2 edition.

Fredkin, E. (2003). The digital perspective. *International Journal of Theoretical Physics*, 42(2):145–145.

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

Gallie, W. B. (1956). Essentially contested concepts. *Proceedings of the Aristotelian Society*, 56:167–198.

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.

Gargett, A. and Barnden, J. (2013). Gen-meta: Generating metaphors using a combination of ai reasoning and corpus-based modeling of formulaic expressions. In *Proceedings of TAAI 2013*.

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.

Gelder, T. V. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, 92(7):345–381.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Gibbs, Jr., R. W. (1994). *The Poetics of Mind*. Cambridge University Press.

Gibbs Jr., R. W. (1993). *Process and products in making sense of tropes*, page 252âĂŞ276. Cambridge University Press, 2 edition.

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

Grefenstette, E. and Sadrzadeh, M. (2011). Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

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

Halawi, G., Dror, G., Gabrilovich, E., and Koren, Y. (2012). Large-scale learning

of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1406–1414.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

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

Haugeland, J. (1993). Mind embodied and embedded. In Houng, Y. H. and Ho, J., editors, *Mind and Cognition: 1993 International Symposium*, pages 233–267. Academica Sinica.

Hegel, G. W. F. (1816/1989). *Science of Logic*. Humanities Press, Atlantic Highlands, NJ. Translated by A. V. Miller.

Heidegger, M. (1926/1962). *Being and Time*. Basil Blackwell, Oxford. translated by John Macquarrie and Edward Robinson.

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

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

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

Hobbes, T. (1651). *Leviathan*. Andrew Cooke.

Hoffmeyer, J. (1997). *Signs of Meaning in the Universe*. Indiana University Press.

Hovy, D., Srivastava, S., Kumar, S., Sachan, J. M., Goyal, K., Li, H., Sanders, W., and Hovy, E. (2013). Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57.

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

Hume, D. (1738/2000). A treatise of human nature. Oxford University Press.

Husserl, E. (1900/2001). Logical investigations, volume 1. Number v. 1 in (International library of philosophy and scientific method.). Routledge. Translated by J. N. Findlay.

Hutto, D. D. (2001). Consciousness and conceptual schema. In Pylkkänen, P. and Vadén, T., editors, *Dimensions of Conscious Experience*, pages 15–43. John Benjamins.

Indurkhya, B. (1997). Metaphor as change of representation: an artificial intelligence

perspective. *Journal of Experimental & Theoretical Artificial Intelligence*, 9(1):1–36.

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

Jankowiak, K., Naskręcki, R., and Rataj, K. (2015). Event-related potentials of bilingual figurative language processing. In *Poster presented at the 19th Conference of the European Society for Cognitive Psychology*, Paphos, Cyprus.

Jezek, E. and Hanks, P. (2010). What lexical sets tell us about conceptual categories. *Lexis*, 4:7–22.

Johnson, M. (1990). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press.

Jordanous, A. K. (2012). *Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application*. PhD thesis, University of Sussex.

Jr, R. W. G. and Tendahl, M. (2006). Cognitive effort and effects in metaphor comprehension: Relevance theory and psycholinguistics. *Mind and Language*, 21(3):379–403.

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

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

Kant, I. (1787/1996). *Critique of Pure Reason*. Hackett Publishing Company, Indianapolis, IN. Translated by Werner S. Pluhar.

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

Kartsaklis, D. and Sadrzadeh, M. (2013). Prior disambiguation of word tensors for constructing sentence vectors. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1590–1601.

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.

Kauffman, S. A. (1995). *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. Oxford University Press.

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

Kiela, D. and Clark, S. (2014). A systematic study of semantic vector space model

parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) @ EACL 2014*, pages 21–30, Gothenburg.

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

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, 7(2):257–266.

Koch, C. (2004). *The Quest for Consciousness: A Neurobiological Approach*. Roberts and Company.

Koestler, A. (1964). *The Act of Creation*. Hutchinson, London.

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

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

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. University of Chicago Press.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

Lakoff, G. and Johnson, M. (2003). *Metaphors We Live By*. University of Chicago Press, 2nd edition.

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.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Langacker, R. (1991). *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Mouton de Gruyter, Berlin.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical Prerequisites*. Stanford University Press, Stanford, CA.

Lapata, M. and Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):261–315.

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

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

Lee, M. G. and Barnden, J. A. (2001). Reasoning about mixed metaphors within an implemented artificial intelligence system. *Metaphor and Symbol*, 16(1-2):29–42.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64:354–61.

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

Levy, O. and Goldberg, Y. (2014a). Linguistic regularities in sparse and explicit word representations. In *Eighteenth Conference on Computational Natural Language Learning*.

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

Levy, O., Goldberg, Y., and Dagan, I. (2015a). Improving distributional similarity with lessons learned from word embeddings. *Transaction of the Association for Computational Lingusitics*, 3:211–225.

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

Locke, J. (1689/1997). An essay concerning human understanding. Penguin, London.

Luong, T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 104–113.

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

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, 49:199–227.

Malandrakis, N., Potamianos, A., Elias, I., and Narayanan, S. S. (2013). Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech and Language Processing*, 21(11):2379–2392.

Margolis, E. and Laurence, S. (2007). The ontology of concepts–abstract objects or mental representations? *Noûs*, 41(4):561–593.

Maturana, H. and Varela, F. (1987). *The Tree of Knowledge*. Shambhala, Boston, MA.

Translated by Robert Paolucci.

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

McGregor, S., Purver, M., and Wiggins, G. (2016). Words, concepts, and the geometry of analogy. In *Proceedings of the Workshop on Semantic Spaces at the Intersection of NLP, Physics and Cognitive Science (SLPCS)*, pages 39–48.

McGregor, S., Wiggins, G., and Purver, M. (2014). Computational creativity: A philosophical approach, and an approach to philosophy. In *Proceedings of the Fifth International Conference on Computational Creativity*.

McGregort, S., Jezek, E., Purver, M., and Wiggins, G. (2017). A geometric method for detecting semantic coercion. In *Proceedings of 12th International Workshop on Computational Semantics*.

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

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

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

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

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

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

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

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439.

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.

Narayanan, S. (1999). Moving right along: A computational model of metaphoric reasoning about events. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence*, pages 121–127.

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

Ortony, A. (1975). Why metaphors are necessary and not just nice. *Educational Theory*, 25(1):45–53.

Ortony, A., editor (1993). *Metaphor and Thought*. Cambridge University Press, 2nd edition.

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

Pantel, P. (2005). Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 125–132, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pattee, H. H. (2001). The physics of symbols: Bridging the epistemic cut. *Biosystems*, pages 5–21.

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

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

Pierce, J. R. (1980). *An Introduction to Information Theory*. Dover, New York, 2nd edition.

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language*. William Morrow.

Plato (1892). *The Republic*. Oxford University Press.

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

Pustejovsky, J. (1993). Type coercion and lexical selection. In Pustejovsky, J., editor, *Semantics and the Lexicon*, pages 73–94. Kluwer Academic Publishers.

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

Pustejovsky, J. and Jezek, E. (2008). Semantic coercion in language: Beyone distributional analysis. *Rivista di Linguistica*, 20(1):181–214.

Pustejovsky, J., Rumshisky, A., Plotnick, A., Jezek, E., Batiukova, O., and Quochi, V. (2010). Semeval-2010 task 7: Argument selection and coercion. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 27–32.

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

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). A word at a

time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, pages 337–346.

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

Reimer, M. (2001). Davidson on metaphor. *Midwest Studies in Philosophy*, 25:142–155.

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

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

Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17(1):67–99.

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

Rączaszek-Leonardi, J. and Nomikou, I. (2015). Beyon mechanistic interaction: value-based constraints on meaning in language. *Frontiers in Psychology*, 6(1579).

Roberts, K. and Harabagiu, S. M. (2010). Utdmet: Combining wordnet and corpus data for argument coercion detection. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 252–255.

Roberts, K. and Harabagiu, S. M. (2011). Unsupervised learning of selectional restrictions and detection of argument coercions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 980–990.

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

Rowlands, M. (2010). *The New Science of the Mind*. The MIT Press, Cambridge, MA.

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

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

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

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.

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.

Sapir, E. (1970). *The Status of Linguistics as a Science*, pages 65–77. University of California Press.

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

Schütze, H. (1992b). Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE conference on Supercomputing*, pages 787–796.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

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

Searle, J. R. (1979). Metaphor. In Ortony, A., editor, *Metaphor and Thought*. Cambridge University Press.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Shanahan, M. (2010). *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. Oxford University Press.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

Shutova, E. (2010). Models of metaphor in nlp. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 688–697.

Shutova, E. (2013). Metaphor identification as interpretation. In *Proceedings of *SEM 2013*.

Shutova, E. (2015). Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Shutova, E., Kaplan, J., Teufel, S., and Korhonen, A. (2013). A computational model of logical metonymy. *ACM Transactions on Speech and Language Processing*, 10(3):11:1–11:28.

Shutova, E., Teufel, S., and Korhonen, A. (2012). Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Skinner, B. F. (1957). *Verbal Behavior*. Copley Publishing Group, Acton, MA.

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

Sowa, J. F. (2006). *Semantic Networks*. John Wiley & Sons, Ltd.

Sperber, D. and Wilson, D. (1995). *Relevance: Communication and Cognition*. Blackwell, 2nd edition.

Sperber, D. and Wilson, D. (2012). A deflationary account of metaphors. In Wilson, D. and Sperber, D., editors, *Meaning and Relevance*, pages 97–122. Cambridge University Press.

Sweetser, E. (1990). *From Etymology to Pragmatics: Metaphor and Cultural Aspects of Semantic Structure*. Cambridge University Press.

Thomas, M. S. C. and Mareschal, D. (1999). Metaphor as categorisation: A connectionist implementation. In *Proceedings of the AISB '99 Symposium on Metaphor,*

Artificial Intelligence, and Cognition, University of Edinburgh.

Thompson, E. (2007). *Mind in Life*. Harvard University Press, Cambridge, MA.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215.

Toutanova, K. and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 63–70.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 248–258. The Association for Computer Linguistics.

Turney, P. D. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, pages 491–502, London, UK, UK. Springer-Verlag.

Turney, P. D., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690.

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

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

Utsumi, A. (2011). Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive Science*, 35(2):251–296.

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

van Genabith, J. (2001). Metaphors, logic and type theory. *Metaphor and Symbol*, 16(1-2):23–57.

Veale, T. (2012). From conceptual mash-ups to bad-ass blends: A robust computational model of conceptual blending. In *Proceedings of the Third International Conference on Computational Creativity*, pages 1–8.

Veale, T. (2016). Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 34–41, San Diego, California. Association for Computational Linguistics.

Veale, T. and Hao, Y. (2007). Comprehending and generating apt metaphors: A web-driven, case-based approach to figurative language. *AAAI*, pages 1471–1476.

Veale, T. and Keane, M. T. (1992). Conceptual scaffolding: A spatially foounded meaning representation for metaphor comprehension. *Computational Intelligence*, 8(3):494–519.

Veale, T., Valitutti, A., and Li, G. (2015). Twitter: The best of bot worlds for automated wit. In Streitz, N. and Markopoulos, P., editors, *Distributed, Ambient, and Pervasive Interactions: Third International Conference, DAPI*, pages 689–699. Springer International Publishing.

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

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

Whitehead, A. N. and Russell, B. (1927). *Principia Mathematica*. Cambridge University Press.

Whorf, B. L. (2012). *Science and Linguistics (1940)*, pages 265–280. MIT 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*, pages 136–143.

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

Wiggins, G. A. (2006). Searching for computational creativity. *New Generation Computing*, 24:209–222.

Wiggins, G. A. (2012). The mind's chorus: Creativity before consciousness. *Cognitive Computing*, (4):306–319.

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.

Wittgenstein, L. (1953/1967). *Philosophical Investigations*. Basil Blackwell, Oxford, 3rd edition. trans. G. E. M. Anscombe.

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

Znidarsic, M., Cardoso, A., Gervás, P., Martins, P., Hervás, R., Alves, A. O., Oliveira, H. G., Xiao, P., Linkola, S., Toivonen, H., Kranjc, J., and Lavrac, N. (2016). Computational creativity infrastructure for online software composition: A conceptual blending use case. In *Proceedings of the Seventh International Conference on Computational Creativity, UPMC, Paris, France, June 27 - July 1, 2016.*, pages 371–379.