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:

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

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.

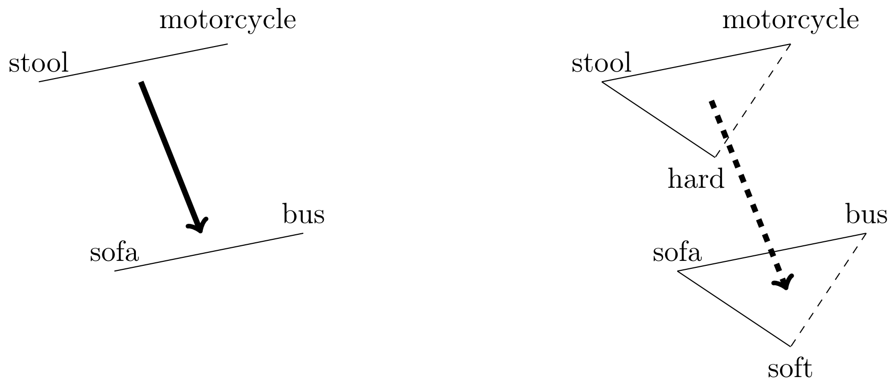
2.1 Analogies as Parallel Vectors

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 2.1: 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 2.1a, 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 2.1b: 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

2.2 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 2.1: 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.

2.2.1 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:

$$PMI(w,c) \approx \log\left(\frac{p(w,c)}{p(w) \times p(c)}\right) \tag{2.1}$$

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:

$$\overrightarrow{a} - \overrightarrow{b} \approx \overrightarrow{c} - \overrightarrow{d} \tag{2.2}$$

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

$$\log\left(\frac{p(a,i)}{p(a) \times p(i)}\right) - \log\left(\frac{p(b,i)}{p(b) \times p(i)}\right) \approx \log\left(\frac{p(c,i)}{p(c) \times p(i)}\right) - \log\left(\frac{p(d,i)}{d(a) \times p(i)}\right) \tag{2.3}$$

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:

$$\frac{p(a,i) \times p(b)}{p(b,i) \times p(a)} \approx \frac{p(c,i) \times p(d)}{p(d,i) \times p(c)} \tag{2.4}$$

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

$$p(i|a, i|d) \approx p(i|b, i|c) \tag{2.5}$$

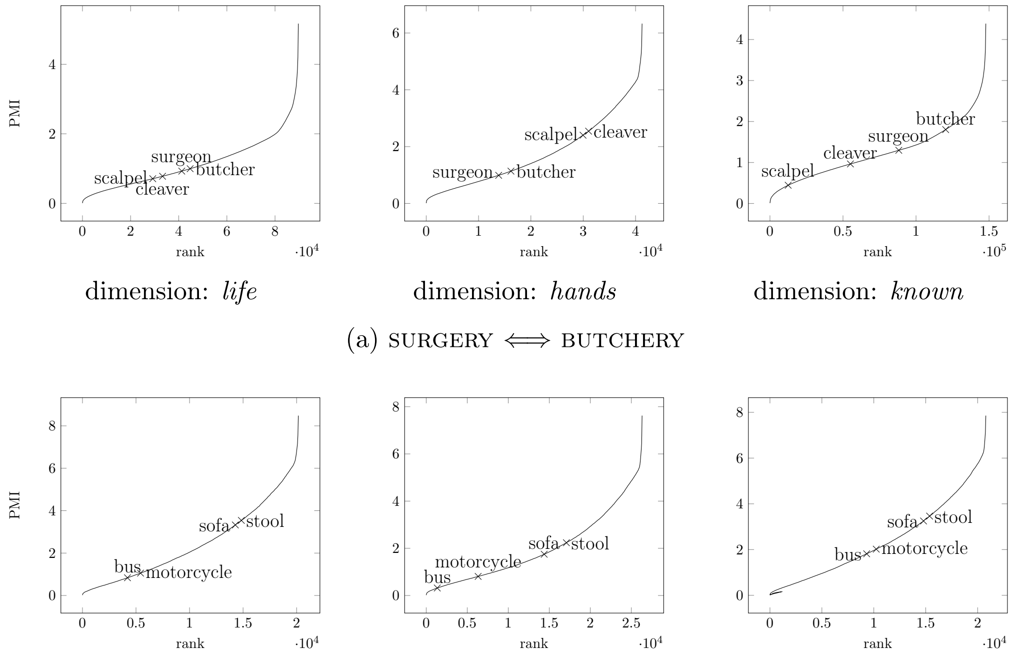
To again state this plainly, our analogy optimisation function seeks to find those dimensions where the combined probability of observing a given co-occurrence term in the context of A and also (though not necessarily simultaneously) D is as close as possible to observing the same term in the contexts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

2.2.2 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 $(\vec{a} - \vec{b}) - (\vec{c} - \vec{d}) = 0$ 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 2.2 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 2.1, *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



dimension: *life* dimension: *hands* dimension: *known*

(a) SURGERY $\Longleftrightarrow$ BUTCHERY

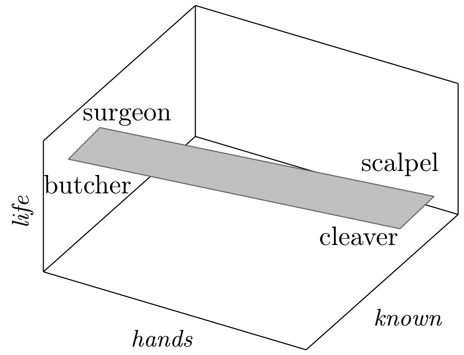
dimension: *beneath* dimension: *hearing* dimension: *accidentally*

(b) FURNITURE $\Longleftrightarrow$ VEHICLES

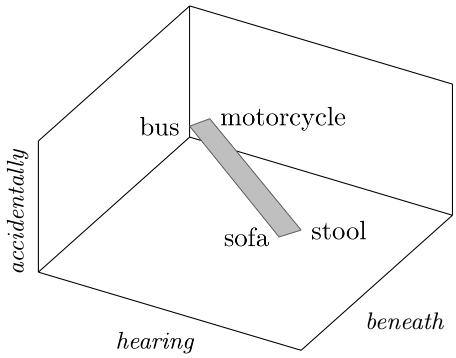
Figure 2.2: Histograms of the top three co-occurrence dimensions satisfying the expected arithmetic of analogy.

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 2.3, 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 2.3 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 2.3: 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 996 analogies from 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

	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 2-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.

examples mentioned 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 2-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 2-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 996 analogy sample of the data used to generate the results in Table 2-A such that the fourth component of each analogy is randomly selected from all possible fourth components across the list. Table 2-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 (*Athens, Greece, Berlin*) that solves the unlikely analogy *Athens:Greece :: Berline:Impossible*. On the other hand, though, these scores are substantially lower than those reported for established analogies in Table 2-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 interesting 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.)

	dimensions	5	10	20	50	100	200	400
2x2	JOINT	0.020	0.052	0.082	0.164	0.221	0.271	0.274
	INDY	0.010	0.079	0.163	0.315	0.406	0.450	0.469
	ZIPPED	0.016	0.055	0.128	0.234	0.280	0.294	0.282
	CBoW	-	-	0.180	0.435	-	0.588	0.612
	SG	-	-	0.191	0.394	-	0.630	0.674
	SVD	-	-	0.035	0.033	-	0.025	0.026
5x5	JOINT	0.016	0.048	0.094	0.199	0.291	0.326	0.327
	INDY	0.017	0.077	0.182	0.340	0.438	0.511	0.509
	ZIPPED	0.016	0.058	0.150	0.307	0.357	0.378	0.367
	CBoW	-	-	0.200	0.448	-	0.607	0.632
	SG	-	-	0.175	0.397	-	0.622	0.672
	SVD	-	-	0.043	0.045	-	0.047	0.051

Table 2-C: Accuracy rates for analogy solution by various techniques with various parameters, taking the first three words in an analogy as input and then providing the fourth word as output.

2.2.3 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. The procedure applied here is simply to take the first three terms for each analogy as input and then use an analysis of the corresponding word vectors to project subspaces following the JOINT, INDY, and ZIPPED methods. These spaces then become the basis for a geometric solution to the analogy, taking the fourth point to be the word-vector that most closely completes the parallelogram indicated by the three input vectors. Unlike with the experiments on relatedness, similarity, metaphor, and coercion described in the previous two chapters, this experiment is completely unsupervised; the hypothesis tested is that contextualisation alone should provide a basis for the geometric modelling of the conceptual alignments inherent in analogy.

Results for my methodologies as well as the `word2vec` modelling techniques and static SVD models, with various parametric specifications, are reported in Table 2-C. With the exception of the most low dimensional subspaces, the INDY technique performs the best of the context sensitive methodologies, eventually offering the expected solution to more than half of the analogies in the dataset for higher dimensional subspaces (and the difference here is approaching significance with $p = .028$ based on a permutation test). This is an interesting result in the context of the lower INDY scores in Tables 2-A and 2-B above: taken altogether, this suggests that this technique returns sparser subspaces consisting of more specialised co-occurrence dimensions salient to only one of the inputs, which make it harder to consistently find dimensions where the analogical relationships play out in the desired way, given full knowledge of the conceptual relationships being mapped, but by the same token less likely to find dimensions that complete an arbitrary analogy, as well. So on balance, co-occurrence dimensions that are more likely to be have a few strong values for a small set of relevant words would seem to correspond to the kind of contextualised conceptual specificity that lends itself to the productive generation of analogical completions.

The context sensitive approaches exhibit steadily improving performance up to 200 dimensions, but then seem to level off after this point, suggestive once again that, in the case of count-based distributional models, a sufficient but still, relative to the dimensionality of the base spaces, small number of highly relevant co-occurrence dimensions are best suited for extrapolating conceptual geometry. Also of note is the improvement across the board moving from lower to higher sized co-occurrence windows. So it would seem that in the case of the semantic alignments at play in analogy, the syntagmatic information available across wider portions of sentences gives more of a semantic handle to the resulting geometric relationships of word-vectors. Finally, and not dissimilarly to results for previous experiments, there is not too much to pick between the JOINT and ZIPPED techniques, though the latter does, as it moves towards the specificity afforded by the INDY method, offer marginally better outcomes.

In terms of the static semantic models, the SVD approach completely collapses on this test. It would seem that the move of dimension-wise normalisation of the abstracted matrix of optimally informational vectors, which proved so effective for word relatedness modelling and at least adequate on other tests, knocks the lexical representations of the model out of the neat geometric relationships that are sought here. The `word2vec` techniques, on the other hand, are outstanding, with any marginal differences in performance between what is reported here and results reported in the original literature as discussed above presumably attributable to variations in the underlying corpora involved in training the models (and the probability of the difference between the top 2x2 window, 400 dimensional skip-gram model and the 5x5 window, 200 dimensional INDY subspaces being attributable to variations in the data is $p = .018$). The skip-gram technique, which Mikolov et al. (2013c) conjectured should be particularly good at handling what they termed as semantic (meaning more overtly intensional) analogy, does consistently better than the CBoW technique at higher dimensionalities, though, interestingly, not at lower dimensionalities, indicating that the co-occurrence predicting approach gains in semantic potency as more informational leeway is allowed into the model.

In order to tease out the operations of various methodologies a bit more, I separate the data into 12 different conceptual categories, in line with the divisions provided by

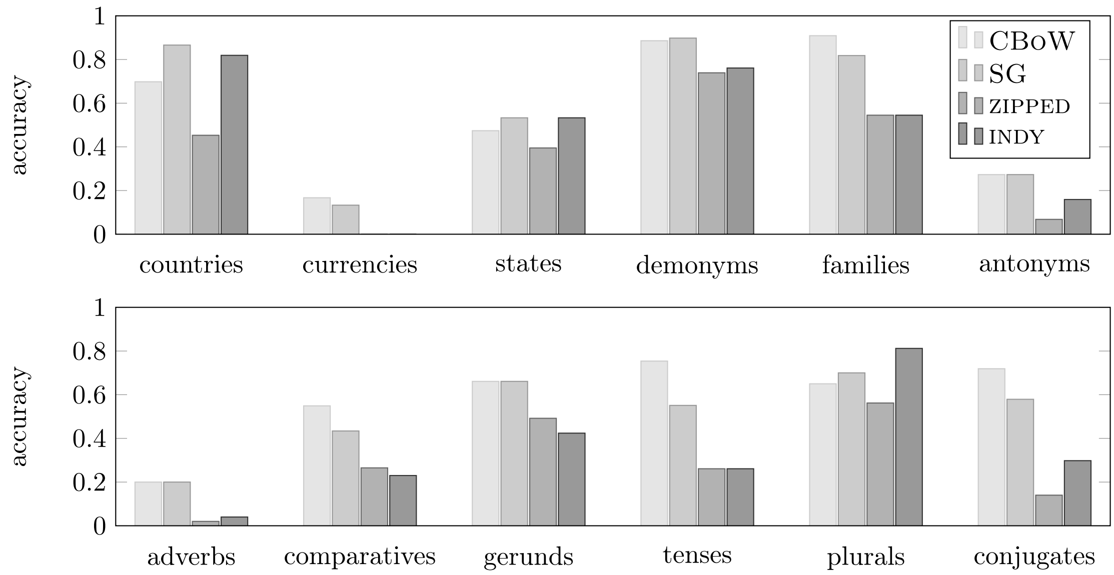


Figure 2.4: Accuracy rates for various modelling techniques on semantically delineated subsets of the data, for 5x5 word window, 200 dimensional spaces.

the original data, though I will not dwell at this point on the theoretical issues raised by the distinctions initially made between semantics and syntax. The categories split out as follows, including the numbers of instances of each category in the sampled data:

232 countries and capitals (*Germany:Berlin :: Tehran:Iran*);
30 countries and currencies (*Mexico:peso :: India:rupee*);
152 American states and cities (*Cincinnati:Ohio :: Memphis:Tennessee*);
22 countries and demonyms (*Denmark:Danish :: China:Chinese*);
50 familial relationships (*father:mother :: son:daughter*);
44 antonyms (*logical:illogical :: informative:uninformative*);
113 adjectives and adverbs (*swift:swiftly :: furious:furously*);
59 adjectives and their comparative and superlative forms (*good:better :: tough:tougher*);
88 verbs and gerunds (*go:going :: write:writing*);
69 present participles and past tense verbs (*screaming:screamed :: saying:said*);
80 singular and plural nouns (*woman:women :: pineapple:pineapples*);
57 singular and plural verb conjugations (*think:thinks :: say:says*).

Figure 2.4 illustrates the accuracy scores for these subsets of the data for four different semantic modelling techniques, the two `word2vec` methods and my ZIPPED and INDY methods, all with parameters set to 5x5 word windows and 200 dimensions for the sake of comparison, presented as a bar chart for the sake of visual comparison within and across categories. The context sensitive methodologies, represented by the darker bars to the right in each categorical cluster, perform somewhat comparably with the neural network models in the cases of conceptual domains such as demonyms and plurals, with the INDY method in particular outperforming all other methods for mapping the expected relationships between singular and plural nouns (though the difference is only marginally significant, with $p = .031$ on a permutation test between the INDY results, with accurracy of 0.812, and the skip-gram results of 0.700). The INDY also does a good job of mapping pairs of countries and capital cities, again with no statistical significance between the INDY score of 0.819 and the skip-gram score of 0.866 (p = .150), though the ZIPPED technique does fall down here, with an accuracy of just 0.453. The overall impression here is that the contextual methodologies seem to do particularly well in instances where there is an expectation of close co-occurrence relationships along one axis of the analogical construct: so, singular and plural forms of nouns are likely to have similar co-occurrence profiles, particularly under the broader 5x5 word window tabulating parameter, and likewise with the names of countries and the gentilics of their inhabitants.

Where context sensitive projections do less well is on the analogies involving shifts that relate to the more abstract conceptual domains of grammatical classes, such as the comparison of adjectival and adverbial forms or the mappings between antonyms. And it is interesting to note that these approaches fail to complete even a single instance of mapping from country name to national currency. One theory here is that the dimensions of currency word-vectors are too sparse and obscure to develop into a coherent conceptual region within subspaces constrained by co-occurrences primarily relating to country names, a hypothesis supported anecdotally by instances of failed analogies such as *Denmark:krone :: Angola:Angolan* and *Japan:yen :: Argentina:Argentine*—indeed, in most cases, inputs involving one currency and two country names seem to almost always map to the demonym rather than to the currency of the unpaired country. A sense emerges that these subspaces match the expected patterns in the dataset best when there is recourse, by way of co-occurrence tendencies, to distinct regions in the contextualised conceptual space of lower dimensional projections. In fact, on the whole, all the methods examined here seem to do better on tests involving conceptual domains which we might qualitatively describe as consisting of more distinctly defined regions, with the looser semantic of comparisons between adjectives and adverbs or pairs of antonyms tending to foil the models. For instance, while there is a clear semantic relationship to be drawn between words such as *happy* and *unhappy* or, orthogonally, *unhappy* and *unimpressed* when they are extracted from their natural sentential habitat, there is in fact quite a bit of semantic distinction in how these words are actually used, and presumably a corresponding disparity in co-occurrence profile. This insight will be revisited below in Section 2.3.

Finally, in order to analyse the performance of the INDY technique in terms of statisti-

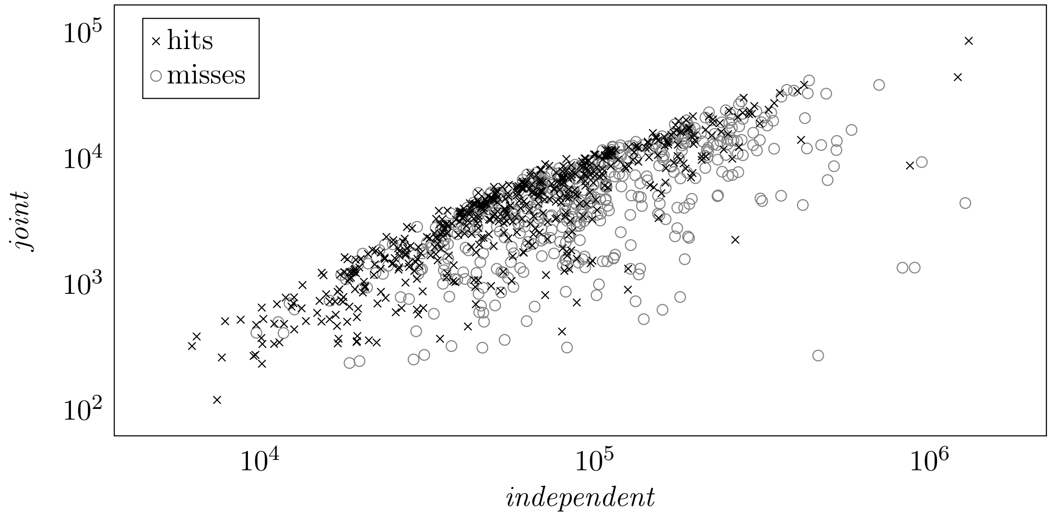


Figure 2.5: A scatter plot of hits and misses for analogy solution as functions of the number of jointly non-zero dimensions versus independently non-zero dimensions, with both axes scaled logarithmically.

cal features of co-occurrences rather than suppositional conceptual categories, Figure 2.5 presents hits and misses in the 200 dimensional, 5x5 word window subspaces as functions of the number of non-zero valued co-occurrence dimensions jointly shared by all three input word-vectors versus the sum of independent non-zero valued dimensions for each word-vector. The trend that emerges here is a distinct clustering of hits for subspaces picked by words that have an overall lower number of positive co-occurrence dimensions, and then secondarily for subspaces invoked by words with a relatively high overlapping co-occurrence profile compared to their independent distributions of co-occurrences. In the first instance, this would seem to indicate that more specialised relationships, denoted by words that either simply come up less or else come up in more particular sentential contexts, are more prone to picking the subspaces where a fourth component of an analogy will be successfully identified. In the second instance, partial analogies involving words that have a relatively significant degree of overlap in terms of the way that they tend to be used. In both cases, it makes sense that the INDY technique, which will hone in on specialised co-occurrences, should provide a good bases for contextualising partial analogies.

So it looks like what's happening is that, as the overlap between the words involved in an analogy decreases, the expected location of the fourth point is pushed into problematic regions of the space on a dimension-by-dimension basis. This can be visualised by once again considering the dimensional analysis of analogy illustrated in Figure 2.2. In the case where the dimension selecting word is the first component of the analogy (A in $A : B :: C : D$), the expected location of the fourth point is actually pushed into the negative region of the dimension, where there is no information to be found. In cases where B or C are the selecting words, D would be expected to be paired with this selector, and so to have a relatively high value along the same dimension. Overall, in these cases where, again, there is minimal co-occurrence information shared between the different input components, the fourth point is pushed into an increasingly unlikely region of a subspace, and the likelihood of a semantically plausible output decreases. This *post hoc* analysis suggests that more nuanced approaches to the dimensional selection process, for instance something like a partial application of the ZIPPED technique, where there is a guarantee of some information for all input words, might be prescribed in cases where, based on frequentist features of the input, difficulty in completing an analogy is expected.

More generally, 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 one boundary of this relationship is notable, even given the logarithmic scaling of the plot, 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 perimeter 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. One quite plausible hypothesis, which is certainly open to future empirical exploration, is that the nature of the distribution illustrated here is peculiar to analogy, and that, if random trios of words were selected, the lower right region of the plot would be more filled in as we discovered more instance of words with a wide range of co-occurrence features but relatively little overlap between one another corresponding to conceptual disparity.

2.3 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*. Both of these examples exhibit conceptual relationships that to a certain extent map to morphological features of the representations involved, and so exemplify some of the characteristics of conceptual schemas described in Langacker's (1991) cognitive grammar. It seems like the designers of this dataset resorted to some instinctive assumptions about the categorical nature of concepts in their selections, ultimately landing on classifications which are pointedly non-ambiguous. This is well motivated, and the data has provided a very useful tool for exploring the geometry of contextualised distributional semantic subspaces, but there is also an element of conceptual absolutism at play and a corresponding drift from the looseness and ambiguity that overwhelmingly characterise natural language as it is actual used by humans. We might reasonably conjecture that these types of conceptual relationships are particularly conducive to the geometry of static semantic spaces such as those generated by `word2vec`.

? has made some very interesting observations regarding the way that analogical geometry actually plays out in `word2vec` models. For one thing, it turns out that the parallelograms expected to emerge in analogical mappings are in fact, in the case of the data tested here, elongated to the point of being virtually lines, to such an extent that they are effectively lines and the models would usually return one of the input terms if it weren't for a system constraint prohibiting this. This supports the observation made regarding the conceptual comparmentalisation noted in Figure 2.3, but it also suggests that the types of analogy being analysed here involve pairs of words that, in the static spaces generated by the neural networks of the skip-gram and CBoW methods, are clustered into very distinct semantic regions of static spaces. A related finding was that in a number of cases, the expected completions to analogies were so proximate to one of the input words that a simple nearest neighbour approach based on a single input would return the specified output. This is particularly the case in analogies involving demonyms and capital cities, where the propensity for words related along these conceptual axes to be observed in similar sentential contexts is presumably so high as to cause the models

input	JOINT	INDY	SG	CBoW
stool, motorcycle, sofa	car	bike	motorbike	motorbike
surgeon, butcher, scalpel	knife	butchers	knife	butchers
scalpel, cleaver, surgeon	advisor	physician	beckwith	physician
key, piano, string	violin	cello	cello	violin
bird, fly, fish	catch	freshwater	anchovies	bait
fire, hot, ice	cold	topped	bubbling	hockey
cold, ice, hot	chocolate	top	hockey	hockey
theater, world, actor	commentator	won	medallist	medallist
candle, life, wind	disturbance	winds	lives	lives
discontent, winter, glorious	summer	olympics	autumn	summer

Table 2-D: Results for a few different models, with 5x5 word windows and 200 dimensional spaces, for completing a small set of analogies spanning various conceptual domains.

to push their vectors into virtual collocation.

In order to explore the performance of distributional semantic models as they move outside the boundaries of analogies between well define conceptual domains, I experiment with a small, curated set of analogies that are designed to transgress in various ways the type of conceptual boundaries inherent in the data explored in the previous section, outlined in Table **??**

The poor performance of the INDY technique on these examples suggests that this method is also possibly better suited for handling stable lexical representations; the strong results in Tables 2-C may, in fact, be an artefact of this property of these subspaces, and this invites further research into subspace selection techniques.

It is worth recalling, as mentioned in Chapter 4.3.1, that the dimensions that compose a subspace cannot be interpreted in a one-by-one way; rather, they collectively constitute a context in which semantic relationships emerge. Along the same lines, I will suggest that the conceptual geometry which affords analogy, and the attendant potential for, for instance, metaphor making, is not always open to instant interpretation in terms of the informational transfer at play in the linguistic construction. This supposition is in line with the relevance theoretical approach to metaphor discussed originally in Chapter 2.3 and again in the context of my own methodology in Chapter **??**.

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.