Linguistics

M.Phil Thesis

Distinguishing Intersective and Non-Intersective Adjectives in Compositional Distributional Semantics

Author
Matthias Lalisse

Supervisor
Dr. Ash Asudeh

Trinity 2015

Contents

1 Introduction **2**

2 Semantics in Vector Spaces **3**
2.1 Distributional meaning 5

3 Adjective semantics **5**
3.1 Intersectives and non-intersectives: Basic properties 6
3.2 Tests for intersectivity 8
3.3 Problems with non-intersective adjective denotations 11

4 A probabilistic model of word meaning **13**
4.1 Basic concepts 13
4.2 Lexical state vectors 14
4.3 Computing probabilities from state vectors 15
4.4 A mixed state picture of adjective meanings 17
4.5 Density operators: Mixed states 18
4.6 Entropy of statistical mixtures 20
4.7 Entropy of composite systems 21

5 Experimental implementation **23**
5.1 Corpus 23
5.2 Word vector models 23
5.3 Joint left-right context model 24
5.4 Adjective dataset 25
5.5 Basis truncation 26
5.6 Linking hypothesis 28

6 Results **30**

7 Conclusion **33**
7.1 Future work 34

8 Appendix: Adjective dataset with semantic tags ($n = 300$) **37**

1 Introduction

The proper formal definition of adjective denotations has been since the inception of formal semantics, due to the non-uniformity of interpretations for this part of speech. It makes strong intuitive sense that attributive adjectives denote properties of individuals. After all, what is an adjective used for if not to indicate some property that is not already contributed by the noun it appears with? Indeed, a common characterization of adjectives holds that

> they typically denote properties–most centrally in the domains of size, shape, colour, worth, and age... The core semantic function of adjectives seems to be to provide terms for individual properties. (Huddleston and Pullum, 2002, p. 528–9)

However, this analysis, carried out in full generality, turns out to be untenable for a large class of adjectives called *non-intersective*–to wit, those for which no interpretation can be given a single property/set corresponding to the adjective.

The purpose of this thesis is to study the properties of intersective and non-intersective adjectives, which have a rich characterization in set-theoretic terms, in the setting of distributional semantics. Distributional semantic representations have become prominent in recent years, especially in applications. However, the wide availability of distributional data on the internet and other sources, coupled with the computational power now available for analyzing it, means that a whole new field of models, data, and hypotheses have become available for investigation. The main hypothesis on which distributional semantics rests is that the patterns of distributions of words carry information about their meaning. If this is so, and if the notion of meaning intended by this distributional model is at all relevant to semantic theory, then we should observe some correspondence between the two sets of concepts. And indeed, some recent studies have found precisely these correspondences (Boleda et al., 2012, Geffet and Dagan, 2005).

In formal semantics, natural language meanings are generally given as statements of logic interpreted in a model. In the distributional conception, words with similar distributional patterns in corpora are also similar in meaning. The kind of distributional context that is considered relevant is a parameter of the model, but in general, *context* is taken to refer to the words co-occurring with a given word being modelled, within a certain window of n words. On the basis of such representations, it is minimally possible to detect synonymy between words based on this relation of similarity of contexts. However, much work has been aimed at expanding the range of semantic phenomena that can be sensibly talked about and analyzed in the distributional setting.

In this study, I present a theoretical framework for interpreting the types of models in use in distributional semantics, and spell out a theory-driven hypothesis

about the distributional characteristics of intersective and non-intersective adjectives that is tested against data obtained form a corpus of English.

It should be remarked that the model adopted here is strictly based on untransformed co-occurrence statistics from the generating corpus, since the aim is to test hypotheses related to the characteristics of the data, not to optimize performance on some semantic task. One problem with the full implementation of the distributional hypothesis for research purposes is that the computational resources required are prohibitive. The measurement statistics compiled are potentially immense, which has led practically-oriented researchers to carry out truncations for statistical transformations of the underlying distributional spaces in order to make the problem of finding, ranking, and making use of semantic relationships computationally feasible. However necessary for the purposes of rendering these problems tractable in a practical setting, transforming co-occurrence statistics by various means renders these models only loosely interpretable as tests of distributional properties, since they are not linked to genuine statistics about the characteristics of the language being modelled.

The metric proposed here for detecting distributional analogues of the formal properties of intersectives and non-intersectives is based on the information-theoretic notion of *entropy*, which is a measure of the uncertainty associated with a probability distribution. Specifically, I use the von Neumann entropy (Kartsaklis, 2014, Neumann, 1955), which is a generalization of the classical Shannon entropy measure (Shannon, 1948) that applies to vector spaces. Intuitively, the more uncertain the outcome of a random experiment, the higher the entropy associated with the experiment. Briefly, the hypothesis is that intersective adjectives consistently denote the same properties across all uses, whereas non-intersectives vary depending on the meaning of their noun argument. If property denotations predict distributional characteristics, it is expected that nouns modified by an intersective adjective will have lower uncertainty (= lower entropy) about their location within the semantic space. Conversely for non-intersective adjectives. Section 2 reviews the essentials of distributional semantics, and lays out relevant literature, including work that has aimed at linking formal and distributional semantic models. Section 3 explains the distinction between intersective and non-intersective adjectives, defines operational tests for discriminating them. Section 4 crystallizes the probabilistic model of word meaning implied in distributional semantics work, and links this to an explicit method for computing representations of lexical states corresponding to modelled words. It is shown how these lexical states are related to the co-occurrence statistics. A minimal mathematical framework is developed for the purposes of proving the framework's internal consistency and relevance to distributional hypothesis-testing. The expected correspondences between distributional uncertainty and (non-)intersectivity are explained.

Density matrices are the key mathematical objects employed for the characterization of distributional uncertainty around adjectives. Density operators are used to represent probability distributions over states of a system when there is uncertainty about what state the system is in. Within distributional semantics, density matrices have been proposed for a variety of tasks related to disambiguating word meanings. Kartsaklis (2014) proposes a density operator-based word representations that capture various sense of a word in a compact form, to enable disambiguation. Blacoe et al. (2013) use density matrices to represent probability distributions over dependency relations, and define a similarity relation between density matrices that exploits the power of these objects to represent ambiguity and sense-selecting behavior.

In addition to representing ambiguity of meaning, density operators come with a measure of entropy or uncertainty that allows the level of uncertainty about the meanings of words or groups of words to be quantified. Hence, these representations are a good candidate for measuring the level of uncertainty associated with the meanings of different classes of adjective. In 4, I explain the von Neumann entropy, a generalization of the classical Shannon entropy measure defined, which quantifies the amount of uncertainty associated with a probability distributions. The von Neumann entropy assigns a value to the level of uncertainty about the state of a probabilistic system defined over a vector space. The results of an experiment conducted on a semantic space constructed from nouns and adjective-noun pairs are reported in section 5.

2 Semantics in Vector Spaces

Although many different models exist for constructing vector representations for words on the basis of co-occurrence statistics observed in large language corpora, a simple baseline model consists of the following. A vector of frequencies can be constructed for each word w, in which each vector component corresponds to a context word. The vector for w is then:

$$[\text{count}(c_1),\ \text{count}(c_2),\ \ldots\ \text{count}(c_i),\ \ldots\ \text{count}(c_n)] \tag{1}$$

where each c_i is a context word co-occurring with w.

This type of representation easily induces a variety of metrics for distributional distance between words, which can empirically be justified as a metric of semantic distance. A common metric of lexical distance is the *cosine similarity*, which is given by the equation:

$$\text{sim}(\vec{w_1}, \vec{w_2}) = \frac{\vec{w_1} \cdot \vec{w_2}}{||\vec{w_1}||\ ||\vec{w_2}||} = \cos(\theta) \tag{2}$$

where θ is the angle between $\vec{w_1}$, $\vec{w_2}$, and $||\vec{w}||$ is the length of the vector $\vec{w}$. This quantity is greatest when $\frac{\vec{w_1}}{||\vec{w_1}||} = \frac{\vec{w_2}}{||\vec{w_2}||}$. The division by the lengths of $\vec{w_1}$ and $\vec{w_2}$ corresponds to the fact that the overall frequency of occurrence of a word is not important, only its relative frequency of occurrence with each context, as a proportion of all its contexts. As an intuitive example, consider that while *canine* and *dog* are essentially identical in meaning, they will occur with very different frequencies, a fact which may obscure their underlying similarity. Normalizing the word vectors $\vec{w}$ (giving them length 1 by dividing by their lengths $||\vec{w}||$) eliminates this distinction, leaving only differences in the proportion of contexts corresponding to each c_i. Other common similarity metrics include Eulcidean and City Block distance, and information-theoretic measures such as Hellinger, Bhattacharya, and Kullback-Leibler distance (Bullinaria and Levy, 2007). In addition to modelling the semantic relations between individual words or short phrases obtained from co-occurrence statistics, many researchers have explored the possibility of implementing algebraic operations over such vectors, allowing for the creation of adequate vector representations for expression classes constructed out of more basic expressions.

These projects are largely motivated by the linguistic insight that meaning construction is combinatory and functional, and, moreover, that the types of larger expressions are often quite different than those of their constituent parts. As a consequence, many current models of compositional distributional semantics propose a deep unification of vector representations for atomic words with compositional semantic architecture, generally modelled along the lines of Montague semantics or Lambek syntactic pregroups, guided by a linguistic typology of expressions (Clark et al., 2008, Clark and Pulman, 2007, Coeke et al., 2010, Sadrzadeh et al., 2013, 2014).

Concrete proposals for word and short phrase representations include studies of adjective-noun composition (Baroni and Zamparelli, 2010, Mitchell and Lapata, 2010) based on a variety of semantic composition functions. In Baroni and Zamparelli (2010), adjectives are assumed to be linear maps given by order-2 tensors (matrices) acting on distributionally obtained vectors for nouns. Each adjective matrix $A : N \rightarrow N$ is therefore a function whose inputs and outputs are both nouns, such that the resulting vectors can be compared with those for nouns. Adjective maps are estimated by linear regression over noun vectors, with the adjective matrix optimized to achieve minimum-error mappings for nouns onto their adjective-noun phrase counterparts. This model is the most pertinent to the current research project, but one extension of it bears mentioning. Grefenstette et al. (2013) extend this algorithm to cover other word-classes, for instance those like transitive verbs that take multiple (NP) arguments. For a functional word F with arguments x and y, this is achieved through multi-step regression learning. The steps are: (1) estimate order-2 tensors Fx for each x predicting vectors Fxy for each y. (2) Then,

for each x, estimate an order-3 tensor F predicting Fx. The process may be iterated for order-$n+1$ tensors, although at this point the representations become quite large. Not only is it difficult to compute the results of compositional procedures involving such large objects; in order to estimate the tensor representations for a functional-type lexical item, it is necessary to solve an exponentially increasing number of linear regression problems.

Another class of compositional distributional models found widely in many current systems uses neural network embeddings of words, phrases, and contexts to derive representations that exhibit sensitivity to variations in meaning (Huang et al., 2012, Mikolov et al., 2013). Such models have been extended beyond modeling words or small phrases; some are implemented with compositional architecture (Socher et al., 2012). However, they are orthogonal to the current study because they are not informed by any linguistic semantic typology, and thus they are essentially atheoretical.

2.1 Distributional meaning

There exist many successful implementations of compositional distributional models, and these models have proven successful in a variety of semantic tasks (Grefenstette and Sadrzadeh, 2011). However, as many researchers freely admit, there is little understanding of the correspondence between the distributional representations employed in compositional distributional semantics and the logical and set-theoretic objects used to characterize natural language meaning in the type-logical setting of formal semantics.

Some work in this direction does exist, however. A widespread hypothesis is the *distributional inclusion hypothesis*, which states that the contexts of a hypernym will be a superset of the contexts of a hyponym (Geffet and Dagan, 2005, Roller et al., 2014). If $A \subseteq B$, then any context in which A occurs, B may occur as well. Whereas cosine similarity measures are symmetric, the inclusion relation is asymmetric. Therefore, a variety of measures have been devised to test distributional inclusion and classify words accordingly (Lenci and Benotto, 2012). For instance, Rimell (2014) considers a measure of *topic coherence* as a candidate feature for detecting hypernyms, based on the hypothesis that hyponyms, which are more specific, will occur in some contexts where their hypernyms are highly unlikely to occur, since the hyponyms gravitate towards certain topics–for instance, *beer* may be found in the highly specific context *drunk* where *beverage* is unlikely to occur.

At a highly general level, Coeke et al. (2010) propose a two-dimensional truth-theoretic space for sentences that is related to Montagovian characterizations of sentence meaning. However, I am aware of no concrete implementations of this proposal.

Quite close to the topic of the current investigation into the properties of intersective and non-intersective adjectives, Boleda et al. (2012) investigate the distributional properties of adjective maps modelled within the Baroni and Zamparelli (2010) framework described earlier, distinguishing between first-order and higher-order modification. They compared the cosine similarity between observed adjective-noun vectors and those for their corresponding nouns across three classes– intersective, non-intersective subsective, and intensional (privative) adjectives. They found that the mean cosine similarity between adjective-nouns and nouns was highest for intersectives and lowest for privatives. These findings indicate that classifications of adjectives based on set-theoretic characteristics are plausibly linked with distributional effects.

3 Adjective semantics

According to Huddleston and Pullum (2002), adjectives are defined as a word class whose characteristic function is to modify nouns. Syntatically, they appear in one of three positions. Adjectives may be

1. attributive appearing before a noun as in *the red sky*
2. predicative appearing as a copular complement and modifying a noun phrase via the linking copula, as in *Mark is tall*
3. postpositive appearing after a noun, as in *a man full of his own importance*

Often, postpositive constructions preclude the use of an adjective by itself, requiring a larger phrase. Consider the strangeness of *a man full*. Moreover, the postpositive use of adjectives is quite rare in English. Hence, I focus on the core cases of attributive and predicative adjectives. When an adjective appears in the attributive adjective-noun position, I refer to the compound as an AN.

A naïve first approximation of adjective semantics, based on the intuition that adjectives denote properties, is to simply identify an adjective with a set associated with the property. That is, for every adjective α, the denotation $[\![\alpha]\!]$ of α is a predicate of type $e \to t$. Taking the copula to be a function $(e \to t) \to (e \to t)$ that links a predicate given by the adjective, we have the following denotation for *Mark is tall*:

$$\text{is} \coloneqq \lambda A \lambda x . x \in A \tag{3}$$

$$mark \in Tall \tag{4}$$

Given a denotation for an adjective as a set, we can obtain denotations for their attributive counterparts via a simple type-shifting operator $\mathsf{shift} : \lambda A \lambda N . N \cap A$.

Hence, from the denotation of A, we are immediately given the denotation of $\mathsf{shift}A$, a function from noun denotations to the intersection of the adjective denotation with the noun denotation.

However, if this were the correct denotation for an adjective α, an immediate consequence would be that the attributive uses of an adjective should denote subsets of their predicative uses. However, at least since Montague (1974), it has been known that this treatment of adjective denotations is highly unsatisfactory. While adjectives, like nouns, would seem to denote sets, and indeed largely do so when used in the predicative position, they do not uniformly behave this way when used attributively. Using the above example of a *big flea* being *big* in a sense different from a *big elephant*, Montague generalized to the worst case, arguing that the denotation of an adjective could not be a single set.

> The denotation of an adjective phrase is always a function from properties to properties... The standard denotations of many adjectives—for instance, "green" and "married"—may be taken as intersection functions, that is, functions H such that, for some property P, $H(Q)$ is, for every property Q, the property possessed by a given individual with respect to a given possible world if and only if the individual possesses both P and Q with respect to that possible world. It would be a mistake, however, to suppose that all adjectives could be so interpreted. (Montague, 1974, p. 211)

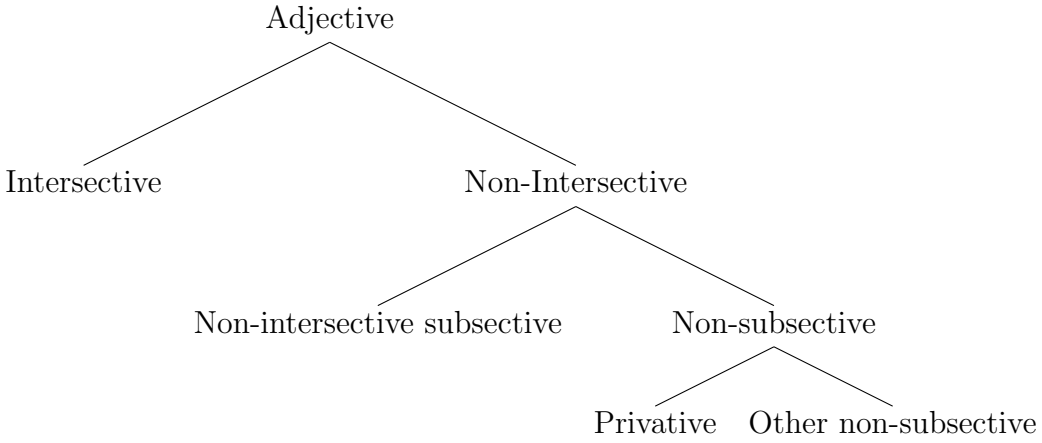
Montague goes on to use the example of a graded adjective, *big*, to show how certain adjectives fail to be intersective: "not all big fleas (indeed, probably no big fleas) are big entities. (A big flea is, roughly, a flea bigger than most fleas, and a big entity an entity bigger than most entities.)"

This characterization highlights the major difference between intersective and non-intersective adjectives: the latter are sensitive to the surrounding context, in particular the context provided by their noun argument, to receive a denotational interpretation. Whereas the set intersected with a noun argument in the case of an intersective adjective is invariant across all uses, the meaning of "big" in any particular instance is determined by the noun it is applied to. This fact of varying meanings depending on the local context is ripe for empirical investigation using the tools of distributional semantics.

3.1 Intersectives and non-intersectives: Basic properties

Before proceeding, it is important to clarify the terminology around different types of adjectives. The typology of adjectives with respect to their inclusion properties can be represented in a partial order, where the "mother of" relation corresponds

to set inclusion.



It is important to note that the non-intersective adjectives broadly understood are not a subset of the subsective adjectives. In fact, any adjective not falling into the class of intersective adjectives may be considered non-intersective. However, when I refer to non-intersective adjectives, I mean the non-intersective subsective kind, as that is the primary distinction of interest in this study. *Subsective* adjectives have the property that, when they are applied to a noun, the resulting denotation is always a subset of the denotation of the noun. For instance, the set of *unfortunate* ηs is always a subset of ηs for any noun η. Compare this with the *intensional* or *privative* adjectives such as *former* or *fake*. The set of *former dancers* is complementary to the set of *dancers*; they have no members in common. In between are more ambiguous cases like *possible*. The set of *possible suitors of Pocahantas* is neither included in, nor complementary to, the set of *suitors of Pocahantas*.

Non-subsective adjectives, including privative/intensional adjectives, will not be studied here due to their sparseness. It is difficult to prepare a large enough sample of such adjectives to study their distributional properties adequately. Hence they are avoided in this project, and are intentionally excluded from the testing dataset discussed in section 5.

Many accounts exist of the set-theoretic properties of intersective and non-intersective adjectives respectively, but I will focus on those that attempt to account the meaning-shifting character of non-intersective adjectives. Considering examples such as

1. *Sam is a giant and a midget*
2. *Sam is a giant midget*
3. *Sam is a midget giant*

Kamp and Partee (1995) argue that, in cases of non-subsective behavior, the adjective is "coerced" by its context–in this case, the head noun–into adopting a meaning compatible with the context. Whereas 1 is simply contradictory without an exotic interpretation of *giant* or *midget*, 2 and 3 are generally accepted to be perfectly acceptable. This suggests that, in modifier-head constructions, the modifier undergoes a reinterpretation relative to the context provided by the head. An even more radical stance is expounded in (Partee, 2007), where it is argued that, in cases of so-called privative adjectives, such as *fake fur*, the meaning of the noun is actually modified in order to accommodate instances of both real and fake fur.

Pustejovsky (1991) argues from within the Generative Lexicon (GL) theory of lexical semantics that adjectives (among other lexical categories) undergo type- and denotation-shifting alterations as a function of the semantic and pragmatic context in which they are employed. In GL, lexical entries are represented as feature-value structures containing information about various types assigned to the lexical entry. In the course of compositional operations, modifiers make use of the information encoded in the lexical entry, altering their own meanings accordingly. Pustejovsky uses the example of the word *good*, whose effect on its argument varies depending on what the argument is. For instance, all of the following are appropriate characterizations of the meaning of *good*. As an eventive predicate, *good* modifies its arguments based on the events that these are intended to or are typically involved in. Hence, based on the lexical specification that the purpose of a knife is to be used in *cutting*, *good knife* derives the meaning *a knife that* ***cuts*** *well*.

All of these characterizations of the relationship between modifiers and heads are akin to the tack taken here for distinguishing intersective and non-intersective adjectives. A variety of linguistic theorists have identified and discussed the tendency of many adjectives, especially the non-intersective ones, to change their meaning as a function of their its context.

3.2 Tests for intersectivity

One test for intersectivity–the predicative descent test–has already been mentioned. However, it is not fully reliable. Consider the pair of sentences

1. *Mary is a skilled acrobat*

2. *The acrobat is skilled*

It is not at all clear that 2 does not have the same meaning as 1, i.e. that *The acrobat is skilled (at acrobatics)*. In fact, given the information that the target of *skilled* is an *acrobat*, it is easy to assign the interpretation *The acrobat does her acrobatics skillfully* to 2. The denotations of predicative uses of an adjective thus appear sensitive to the meanings of the precopular nouns, just as attributives

are sensitive to the meanings of their noun arguments. Consider another instance involving Montagues' example of the big flea.

Marcel, noticing an unusually large flea on his arm:

Whoa! That flea is big.

The sentence does not seem false, even though $[\![\text{big}]\!]$, under the standard assumption that it is a predicate when used in the predicative position, can not normally apply to fleas, however large. It is likely that, in this case, the meaning of *flea* propagates across the copula to affect the interpretation of *big*, such that the sentence has an interpretation like

- *That flea is (a) big (flea)*

which indicates that predicative uses of an adjective are not so semantically distinct from attributive uses as is generally assumed. The predicative descent test assumes that the type of the adjective in a predictive position is a set invariant across uses, and tries to match attributive uses of the adjective to this set. If the match fails on some cases, then the adjective is determined to be non-intersective. However, the modifiability of the meaning of the predicative use of the adjective in a non-null context indicates a weakness in this identification strategy.

There are other reasons to prefer the attributive propagation test over the predicative descent test. For example, the latter cannot accommodate certain adjectives which cannot appear in the predicative position. *Ethnic*, for instance, is normally an exclusively attributive adjective, and appears only recently to have begun being used predicatively. Hence there are many cases where the predicative descent test simply cannot be used.

For all of these reasons, in section 5 describing the experimental implementation, I eschew the predicative descent test as a means of distinguishing between intersective and non-intersective adjectives in the preparation of a dataset, due to the possibility of false positives. Instead, I employ the following operational definition of intersectivity relied on to distinguish intersectives from non-intersectives. An adjective α is intersective if:

$$\begin{array}{ll} x \text{ is an } \alpha\ \eta_1 & \text{(Premise 1)} \\ x \text{ is an } \eta_2 & \text{(Premise 2)} \\ \hline x \text{ is an } \alpha\ \eta_2 & \text{(Conclusion)} \end{array}$$

for arbitrary η_1, η_2. This operational definition may be called the *attributive propagation test*, and it can be shown to be fully equivalent to the set-theoretic definition of intersectivity. This set theoretic definition is that an intersective adjective has the denotation of an intersection of the predicate denoted by its argument and some given set A, invariant across uses of the adjective. This set-theoretic definition means that, for a given intersective adjective α, there exists a set A satisfying (5).

$$\forall \eta_i, x. x \in [\![\alpha]\!][\![\eta_i]\!] \leftrightarrow x \in A \cap [\![\eta_i]\!] \tag{5}$$

Equivalently:

$$\exists A \forall \eta_i . [\![\alpha]\!][\![\eta_i]\!] = A \cap [\![\eta_i]\!] \tag{6}$$

Although below I discuss some qualifications to this assignment of denotation for intersective adjectives, for the moment, this is claimed to be the denotation of an intersective adjective. To show the correspondence between between this intersectivity condition and the test laid out below, we first make a handful of assumptions. We assume that denotation of an arbitrary noun η is a predicate, that is $[\![\eta]\!] = P$ for some P. Assume as well a process transcribing an attributive use of an adjective α from a usage like "x is an α η" into propositions of first-order logic of the form $[\![\alpha]\!][\![\eta]\!](x)$ where the type of $[\![\alpha]\!][\![\eta]\!]$ is a predicate, i.e. of type $e \to t$. Since each $[\![\eta]\!] \in \mathcal{P}(U)$, then every $[\![\alpha]\!]$ is a function $f : \mathcal{P}(U) \to \mathcal{P}(U)$. This transcription procedure will not be spelled out in detail; we suppose it to exist. It follows that we can translate the above operational definition into the following condition on α:

$$\forall \eta_i, x, \eta_j . [\![\alpha]\!][\![\eta_i]\!](x) \wedge [\![\eta_j]\!](x)] \to [\![\alpha]\!][\![\eta_j]\!](x) \tag{7}$$

This is equivalent to:

$$\forall \eta_i, x, \eta_j . [x \in [\![\alpha]\!][\![\eta_i]\!] \wedge x \in [\![\eta_j]\!]] \to x \in [\![\alpha]\!][\![\eta_j]\!] \tag{8}$$

We will assume that α is subsective; that is, $[\![\alpha[\![\eta]\!] \subseteq [\![\eta]\!]$ for any η. However, it need not be assumed that $[\![\alpha]\!][\![\eta]\!]$ is defined for any η, i.e. that every attributive use of the adjective α has a denotation, only that this condition applies for any pair of nouns η_1, η_2 for which α is defined. We will show that, as long as α is subsective, these two conditions are equivalent.

Proof. (6) $\to$ (7) Let α, be any subsective adjective. Suppose that (6) holds. Let η_i be an arbitrary noun, x any individual, and η_j any noun. Suppose $x \in [\![\alpha]\!][\![\eta_i]\!]$ and $x \in [\![\eta_j]\!]$. Then by assumption $x \in A_\alpha \cap [\![\eta_i]\!]$, and so $x \in A_\alpha$. But then also $x \in A_\alpha \cap [\![\eta_j]\!]$, and by assumption $x \in [\![\alpha]\!][\![\eta_j]\!]$ as well. Since η_i, x, and η_j were chosen arbitrarily, this is the case for any η_i, x, η_j. This is just condition (7).
(7) $\to$ (6) Now in the other direction. Suppose (7). Consider the set $A_\alpha = \bigcup_i \{[\![\alpha]\!][\![\eta_i]\!]\}$. Let η_i be any noun and x be any individual. If $x \in [\![\alpha]\!][\![\eta_i]\!]$, then clearly $x \in A_\alpha$. And by the subsectivity of α, $x \in [\![\eta_i]\!]$. Therefore $x \in A_\alpha \cap [\![\eta_i]\!]$.
Suppose now that $x \in A_\alpha \cap [\![\eta_j]\!]$. It follows from the definition of A_α that x is in some set $[\![\alpha]\!][\![\eta_i]\!]$ for some η_i, not necessarily η_j. However, by condition (7), if $x \in [\![\alpha]\!][\![\eta_i]\!]$ and $x \in [\![\eta_j]\!]$, then also $x \in [\![\alpha]\!][\![\eta_j]\!]$. Therefore, $x \in [\![\alpha]\!][\![\eta_j]\!]$. From the fact that every x in $[\![\alpha]\!][\![\eta_i]\!]$ is in $A_\alpha \cap [\![\eta_i]\!]$ and conversely, it follows that they are the same set. Hence (6) holds. □

This shows that (7) and the associated operational test for intersectivity provide necessary and sufficient conditions for intersectivity. In other words, the operational

test for intersectivity precisely corresponds to the notion expressed in set theoretic terms. In particular, if an adjective does not satisfy (7), then it is not intersective, and if it does satisfy (7), it is guaranteed to be intersective.

When does intersectivity fail? This can be deduced directly from condition (5). If $\neg$ (5), then for any set A, $\exists \eta_i, x.x \in [\![\alpha]\!][\![\eta_i]\!] \not\Leftrightarrow x \in A \cap [\![\eta_i]\!]$. That is, for all sets A, there is some individual that falls under the denotation of $[\![\alpha]\!][\![\eta_i]\!]$ but not $A \cap [\![\eta_i]\!]$, or conversely. If A exists, then minimally A must contain every $x \in [\![\alpha]\!][\![\eta]\!]$ for each η; otherwise, there would be some $x \in [\![\alpha]\!][\![\eta]\!]$ that is not in $A \cap [\![\eta]\!]$. But in the case of a subsective non-intersective adjective, there is some individual falling under a predicate $[\![\eta]\!]$ that is this set (the generalized union of denotations for the adjective and its noun arguments), but not in $]\!]\alpha[\![[\![\eta]\!]$. The intuitive idea that there is no unified concept, represented as a set, that a non-intersective adjective corresponds to thus has this precise set-theoretic meaning. Consider a concrete example: *Fred is a* $[\![big]\!][\![animal]\!]$ and *Fred is a* $[\![big]\!][\![elephant]\!]$. Clearly if Frank is a baby elephant, he may be considered a big animal without being at the same time considered a big elephant. Then there is no set that may be identified as the meaning of *big*. Instead, *big* must be considered to be some function whose denotation is dependent on the meaning of the noun it is applied to.

It should be noted that the definition of intersective adjectives proposed as conditions (5) and (6) is not uncontroversial, and there exists literature defending alternative accounts of the semantics of intersective adjectives. For instance, pointing to the availability of predicative uses of non-intersective adjectives such as *Jumbo is small*, Heim and Kratzer (1998) argue for first-order $e \to t$ type for apparently non-intersective adjectives. As an example, they provide the following denotation for *small*:

$$[\![\text{small}]\!] = \lambda x.x\text{'s size is below } c\text{, where } c \text{ is the size standard made salient by the utterance context} \quad (9)$$

(Heim and Kratzer, 1998, p. 71)

Given this typing, Heim and Kratzer can account for the failure of entailment of, for example, *Jumbo is a small animal*$\not\Rightarrow$*Jumbo is a small elephant* by appealing to context updates forced by the noun argument *elephant*, which changes the size parameter dictated by the context. At the same time, they are able to maintain that the adjective has a first-order type like that of (9). However, it seems to me that this commitment to a first-order type for the denotations of context-sensitive adjectives is not tenable, since clearly context update must be a function that has the adjective's noun argument as one of its parameters. Although the details are debatable, it is reasonable to suppose that the contextual size parameter is given as a function of $[\![\text{elephant}]\!]$, so that the updated context c is something resembling

(10).

$$c = a \text{ where } a \text{ is the average of } height(x) \text{ for every } x \text{ in } [\![\text{elephant}]\!] \quad (10)$$

Composing [[small]] with this contextual update function and then abstracting out the noun [[elephant]] returns a function that is *at least* second-order (I am not committed to the idea that the noun denotation is the *only* parameter to the context update function). So while the adjective function may be *called* first-order, in an attributive use, it is really covertly second-order or above, with type $(... \times (e \to t) \times ...) \to (e \to t)$. Such a denotation is sharply different from one satisfying the intersectivity condition laid out above, in that the operation performed by the adjective changes as a function of the noun it is applied to.

Detailed examination of these claim is beyond the scope of this paper. However, I have provided a defense of the more restrictive notion of intersectivity encoded in conditions (5)/(6). Although alternative accounts of intersectivity will be considered, the focus will be on non-context-dependent adjectives. For these context-dependent adjectives identified by Heim and Kratzer, Partee (2007), following Siegel (1976), shows that such adjectives can be singled out by examining their distribution in *as-* and *for*-phrases, a test employed later in the computational experiments reported on here.

It is worth noting that, as long as we recognize the distinction between context-dependent and non-context-dependent adjectives, it is unnecessary to dwell for very long on whether they are to be classed as intersectives. So long as all essential terms are properly defined, the choice of whether to call any class of adjectives *intersective* versus *non-intersective* versus intersective but context-dependent is mostly stipulative and of no inherent interest. It suffices to say that by *intersective*, I mean adjectives that satisfy the intersectivity condition (5)/(6). Adjectives like *small*, with denotations like (9) are instead referred to as *context-dependent*. In Section 6, some evidence from a computational experiment indicates that context-sensitive adjectives have more in common with non-intersective adjectives, in distributional terms, than they do with intersectives, a fact which somewhat strengthens the choice to consider context-sensitive adjectives to be part of the class of non-intersectives.

3.3 Problems with non-intersective adjective denotations

The principle of compositionality, speaking somewhat roughly, is the contention that the meanings of expressions are given by the meanings of their constituent sub-expressions and their means of combination. The characterization of non-intersective adjective meanings as maps between properties, maps that vary widely widely depending on the denotations of their arguments, poses no problems for the compositionality. Non-intersective adjectives are a problem for this principle, and for

linguistic theory in general, because the functions denoted by non-intersective attributives are fundamentally underdetermined.

Put another way, even given all available information about the meanings of the constituent parts of the expression, one still cannot specify the meaning of the expression built up through the combinatorial rules of the language. Hence, the meaning of the whole cannot be determined from the meanings of the parts and their means of combination—the essence of compositionality. Even given a set corresponding to the non-intersective adjective used in a predicative position, its meaning when the type of the expression shifts from a predicate to a function over predicates remains unspecified. The adjective functions, in the case of non-intersectives, have missing parameters.

Non-intersectivity of attributive adjectives poses a problem for compositionality. However, the relevant problem may be more one of the insufficiency of a particular theory rather than in the nature of the adjectives themselves. The issue from the standpoint of a semantic theory is that the particular function corresponding to a non-intersective adjective is not known, once the sets denoted by the adjective and all its possible arguments are given. However, a speaker necessarily *does* possess knowledge of the denotation of the intentional adjectives she uses. So the speaker clearly implements a compositional procedure for combining the adjective with its noun argument. So, again adopting the view that adjectives are functions over sets, we might argue that there is no need to specify the particular collection of sets that an adjective function maps its arguments to, so long as we can state some nontrivial conditions on that interpretation. An example is the condition that the interpretation of a subsective attributive be a subset of its argument.

In addition to the problems they pose for compositionality, non-intersective adjectives are also highly challenging from the point of view of providing an adequate explanatory account of natural language semantics that accounts for the acquisition of the semantics of a language. A semantic theory should describe structures such that they could be learned by a speaker upon an encounter with the normal range of data about meanings. From this point of view, it is clear that one who wishes to learn the meanings of non-intersective adjectives faces a daunting task, since such a learner must internalize a method of deriving, for any arbitrary noun which may be combined with an adjective, what the denotation of the adjective-noun compound will be.

In the worst-case scenario, where the meanings of nouns are simply identified with their denotations, the space of possible meanings for an adjective is massive. Taking the traditional picture of the denotations of adjectives as being functions from sets to sets, then, I lay out some implications for the size of the learning problem involved for acquiring the meaning of an adjective. The learning problem is as follows. An adjective maps a set–the denotation of a noun–into another set. Hence a speaker

with knowledge of the language must have knowledge of the function, which may be a partial function. A function can be characterized as a set of ordered pairs of arguments and images under the function. Hence, learning an adjectival function involves learning a (partial) map from the denotations of nouns in the language back into subsets of the domain of individuals. Specifically, an adjective is a function from the power set of the domain $\mathcal{P}(U)$ back into the same power set. Let us assume a language $\mathcal{L}$ with n nouns comprising a set $\mathcal{N} \subseteq \mathcal{P}(U)$ such that $\mathcal{N} = \{N_i\}$. Furthermore, let each $A_i \in \mathcal{A}$ be an attributive-type adjective. Then learning the meaning of each adjective involves learning a set of functions $\mathcal{A}$ such that:

$$\mathcal{A} \subseteq \{f : \mathcal{N} \xrightarrow{f} \mathcal{P}(U)\} \tag{11}$$

In words, each A_i in $\mathcal{A}$ is a partial endofunction on $\mathcal{P}(U)$. Without any specification of what the adjective maps should look like, the space of possible functions is truly immense. To illustrate this, I will consider the relatively trivial case of subsective adjectives. Each adjective maps a noun to a subset of itself, so the image of each noun N_i is in $\mathcal{P}(N_i)$. There are $|\mathcal{P}(N_i)| = 2^{|N_i|}$ such functions for each noun. Hence, for a language with n nouns, each subsective adjective must be learned assigning to each noun an image in its power set. Counting these possible functions, we find their number is given by:

$$\prod_{i=1}^{n} 2^{|N_i|} = 2^{\sum^i |N_i|} \tag{12}$$

Consider a hypothetical language $\mathcal{L}_0$ with a single adjective A_0 and three nouns N_1, N_2, N_3, each denoting 10 individuals. A learner of this language has to consider 2^{30}, or over 10^9 hypotheses. In the special case of intersective adjectives, there is a shortcut that uniquely identifies the subset of N_i denoted by $A(N_i)$. However, for the non-intersective subsective cases, it is clearly unreasonable to suppose that all of these possibilities are entertained. The situation is even worse with the non-subsective adjectives. Consider the privative adjectives. In these cases, we replace N_i in (12) with its complement set $N_i^{\mathcal{C}}$, which is very large for most nouns.

Remarkably, it is not even possible to make this number smaller when the denotations of the nouns overlap. Consider the sets of bankers and doctors. The intersection of these sets is nonempty. Now, it might be hoped that when an adjective like *skillful* is applied to these sets, it would identify the same subset of each argument set. But not even this condition is true.

Formulated as a learning problem involving the selection of a target set for each application of an adjective to any noun, non-intersective adjectives present immense difficulties for a speaker attempting to learn the semantics of a language, and given the poverty of the stimulus about adjective meanings, it is unlikely that a learning procedure based on assignment of denotations will realistically account

for the mature knowledge of a speaker with command of the denotations of non-intersective adjectives.

These considerations lead to the conclusion that an adequate account of the semantics of non-intersectives must lie within a class of theories that add additional parameters to the representations of word meanings. Many such theories are possible, including type-drive combinatorial theories such as the Generative Lexicon Pustejovsky (1991) or the framework presented in Asher (2011) that directly provide an interface with model-theoretic intepretations. In line with common sense, models differentiating words based on co-occurrence data that all language learners are exposed to gives one surely important set of parameters that language learners employ individuate and relate lexical entries. In addition, the mathematical relations obtaining between these representations can be studied using new sets of tools applicable both novel and traditional semantics topics. The relation between the two broad classes of models is best thought of as complementary–examining related phenomena at different levels of granularity, using different data.

4 A probabilistic model of word meaning

4.1 Basic concepts

The distributional model of word meaning is based on the hypothesis, that even if the meaning of word may not be identified with its patterns of co-occurrence with other words, that such patterns are indicative of semantic distinctions. This is a semantic counterpart of the distributional techniques of analysis employed in syntax and phonology for identifying substitution classes of expressions that indicate their commonality at some level of representation. In other words, distributional data preserve some subset of psychologically relevant relations and distinctions: synonymy, semantic incompatibility, perhaps hypernymy and others. The model of word meaning assumed in this framework may be formulated as follows.

Let Σ be a message-emitting machine that produces symbols μ serially. When Σ emits a sequence of messages $\mu_1\mu_2...\mu_k$ drawn from an alphabet $M = \{\mu_n\}$, we write:

$$\Sigma : \ \mu_1 \ \mu_2 \ ... \ \mu_k \tag{13}$$

Consider the information available to an observer O recording the output of Σ. Such an observer is a position to predict the future distribution of messages by Σ based on its prior output. Such predictions come in the form of conditional probabilities $P(\mu_i|\mu_{i-1})$. A model of Σ thus constructed is essentially a finite state machine. From it, one can readily compute the probability of any sequence of messages emitted by

Σ. This elementary characterization may be extended to provide a way of comparing distributions.

An observer O recording the output of Σ will, given enough data, be able to develop an elementary model of the output of Σ in the form of a probability distribution over the messages of Σ. Such a distribution would minimally include $P(\mu_i|\mu)$ for any message μ and any possible context μ_i. It may also be possible, however, for O to compare distinct messages μ_j, μ_k based on the similarity of distribution of other messages around them–speaking very informally, the patterns of distribution for other messages around the modelled messages μ_j, μ_k. This characterization may be rendered more vivid if we refer to lexical states σ corresponding to every message μ, and encoding the patterns of distribution of messages in various context positions around μ.

Definition 1. *A lexical state σ indexed to a message μ is the state of a message-emitting machine Σ when Σ produces μ.*

Correspondingly, the *state space* of Σ is the set of all states that Σ may be found in.

Given a probability distribution over contexts, lexical states may be compared with one another, such that states having similar patterns of contextual distribution are found to be "close together" in some sense. The experimental meaning of this asserted relation between lexical states is that the expected distribution of outcomes in the vicinity of the messages indexed to these states are similar.

The class of language models assumed in this framework can be characterized as follows. A model of a language $\mathcal{L}$ includes, for any message μ, the probability that any message μ_j will be emitted at distance k to the left/right of μ. A model of σ thus consists of an ensemble of probabilities of emitting each μ_i. Each observation of a context of σ can be conceptualized as the outcome of a random experiment, in which state σ results in the outcome of a context μ_i with probability p_i.

$$P(\mu_i|\sigma) = p_i \tag{14}$$

For arbitrary states σ corresponding to messages μ, an elementary model of σ can be constructed from counts of the production by Σ in state σ producing μ of contexts, say, immediately to the right of μ, a situation depicted in (15).

$$\Sigma^{(\sigma)} : ... \ \mu \ \mu_j \ ... \tag{15}$$

Clearly, for any pair of distinct messages μ_i, μ_j, observations of μ_i and μ_j as contexts of μ in this position are mutually exclusive. That is, $P(\mu_i, \mu_j|\sigma) = 0$. Moreover, the entire vocabulary M spans the space of possible context messages; that is, $\sum_{i=1}^{m} P(\mu_i|\sigma) = 1$. Hence our model of σ is a discrete conditional probability distribution over messages emitted by Σ in state σ.

4.2 Lexical state vectors

Given the above stipulations, how will a model for Σ be represented? The model of word meaning characterized above can be given a compact mathematical representation in vector spaces. The main advantage of this presentation is that it provides a way of computing probabilities for context outcomes, as well as providing similarity relations between lexical states. These will be briefly touched on in section 4.3, but are not the main focus of the exposition. Following (Nielsen and Chuang, 2010), I will use the Dirac notation for writing down state vectors and operators over them.[1] The mathematical presentation follows notational conventions designed for applications in quantum physics and quantum information theory, but the thread of exposition is designed to make clear the formalism's relation to the linguistic structures at issue in the present analysis.

Each lexical state σ is associated with a state vector written as a column vector or "ket" $|\sigma\rangle$ which is a superposition (weighted sum) of basis vectors $\{|\beta_i\rangle\}_i$ with coefficients $\{\gamma_i\}_i$ in the interval $[0, 1]$:

$$|\sigma\rangle = \gamma_1 |\beta_1\rangle + \gamma_2 |\beta_2\rangle + ... + \gamma_n |\beta_n\rangle \tag{16}$$

The set of basis vectors is chosen to be orthonormal. That is, for each i, j:

$$\langle\beta_i|\beta_j\rangle = \begin{cases} 1, & \text{if } i = j. \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

Each lexical state vector is represented by a "ket" $|\sigma\rangle$ in a vector space V, and has a dual vector in the dual vector space V^* denoted by the "bra" $\langle\sigma|$, which is obtained by transposing $|\sigma\rangle$ and taking the complex conjugate of each of is components. If $|\sigma\rangle = \gamma_1 |\beta_1\rangle + \gamma_2 |\beta_2\rangle + ... + \gamma_n |\beta_n\rangle$, then:

$$\langle\sigma| = \gamma_1^* \langle\beta_1| + \gamma_2^* \langle\beta_2| + ... + \gamma_n^* \langle\beta_n| \tag{18}$$

where each γ_i^* is equal to the complex conjugate of γ_i.[2] This distinction is crucial in quantum mechanics, where states have coefficients ranging over the complex numbers. However, the co-occurrence vectors constructed in the contexts of distributional semantics have coefficients ranging over the real numbers. Equivalently, $\gamma^* = \gamma$ for each γ. Hence, in this restricted domain, the distinction between bras and kets can be viewed simply as a distinction between row and column vectors (Nielsen and Chuang, 2010, p. 62).

[1] The essentials of this notation are reviewed here, but more detail may be found in (Nielsen and Chuang, 2010).

[2] The complex conjugate of a complex number $a + bi$ is simply the same number with its imaginary portion negated. That is, $(a + bi)^* = a - bi$.

The inner product $(|\phi\rangle, |\psi\rangle)$ between two vectors is written $\langle\phi|\psi\rangle$. The *norm* of a vector is the square root of its inner product with itself, that is: $norm(|\sigma\rangle) = \sqrt{\langle\sigma|\sigma\rangle}$, and is additionally the definition of the "length" of a given vector. The *normalization condition* for state vectors states that if $|\sigma\rangle$ is a state vector, then $|\sigma\rangle$ has a unit norm, or $\langle\sigma|\sigma\rangle = 1$. Therefore, the state space of Σ is equal to the set of all unit-length vectors in some vector space V.

4.3 Computing probabilities from state vectors

If basis vectors are conceived as indexes over possible emissions of messages, and the product of conjugate coefficients $\gamma_i^*\gamma_i$ is thought of as the probability of the outcome μ_i given state σ, it is clear that $|\sigma\rangle$ defines a probability distribution over possible contexts, which are the outcomes of observations of the lexical state σ. Understood this way, the normalization condition has the meaning that the probability of any outcome is 1 (the sum of probabilities of all disjoint outcomes is 1) Susskind and Friedman (2014). Probabilities for any given outcome μ_i can be computed from the state vectors in the following way.

Let σ be a lexical state and $|\sigma\rangle = \gamma_1 |\beta_1\rangle + \gamma_2 |\beta_2\rangle + ... + \gamma_n |\beta_n\rangle$ its state vector written as a superposition of orthonormal basis vectors $\{|\beta_i\rangle\}_i$ indexed to outcomes $\{\mu_i\}$. Clearly for any vector $|\beta_i\rangle$ in the basis, $\langle\sigma|\beta_i\rangle\langle\beta_i|\sigma\rangle = \gamma_i^*\langle\beta_i|\beta_i\rangle\gamma_i\langle\beta_i|\beta_i\rangle = \gamma_i^*\gamma_i$. Moreover, $\langle\sigma| \left(\sum_i |\beta_i\rangle \langle\beta_i|\right) |\sigma\rangle = \sum_i \langle\sigma|\beta_i\rangle\langle\beta_i|\sigma\rangle = \sum_i \gamma_i^*\gamma_i = 1$ by the normalization condition. So $|\sigma\rangle$ encodes a probability distribution over the outcomes μ_i.

Given this fact, it makes sense to associate each observable $O_i = |\beta_i\rangle \langle\beta_i|$ with an elementary event in the outcome space (Susskind and Friedman, 2014). This leads to the following definition.

Definition 2. *Let μ_i be a (co-occurrence) outcome and $|\beta_i\rangle$ an orthonormal basis element indexed to μ_i. Then*

$$\begin{aligned} P(\mu_i|\sigma) &= \langle\sigma|\beta_i\rangle\langle\beta_i|\sigma\rangle \\ &= \langle\sigma| O_i |\sigma\rangle \end{aligned}$$

where O_i is the observable corresponding to μ_i.

The objects $|\beta_i\rangle \langle\beta_i|$ are the outer products of basis vectors of $|\beta_i\rangle$, and correspond to projections onto one-dimensional subspaces induced by the individual basis elements. The outer product $|\phi\rangle \langle\psi|$ between two vectors $|\phi\rangle = \sum_i^n \gamma_i |\phi_i\rangle$ and $|\psi\rangle = \sum_j^n \delta_j |\psi_j\rangle$ is $\sum_{i,j} \gamma_i^*\delta_j |\phi_i\rangle \langle\psi_j|$ and can be represented as an $m \times n$ matrix M with entries $M_{ij} = \gamma_i^*\delta_j$. Consider an observable O_i to correspond to a random experiment in which the result is 1 if μ_i is observed, and 0 otherwise. Then

$$\langle O_i\rangle = \langle\sigma| O_i |\sigma\rangle = P(\mu_i|\sigma) \tag{19}$$

is the expected value of this experiment. In the distributional semantics setting, then, $\langle O_i \rangle$ gives the expected value of observations (Kartsaklis, 2014, Susskind and Friedman, 2014) of a co-occurrence context μ_i when the signaller Σ is in state σ.

I will now give a concrete construction of lexical state vectors derived from a corpus that satisfy these requirements. Each modelled word μ is associated with a vector $|c\rangle$ with components that are raw counts of the co-occurrence frequency for each word μ_i that co-occurs with it. An indexation of context words μ_i to orthonormal basis elements $|\beta_i\rangle$ is assumed. Therefore each raw frequency vector is given by

$$|c\rangle = \sum_i \text{count}(\mu_i, \mu)\, |\beta_i\rangle \tag{20}$$

where $\text{count}(\mu_i)$ is the frequency of μ_i occurring in a particular contextual position with respect to the target word μ. Since all contexts are recorded,

$$P(\mu_i|\sigma) = \frac{\text{count}(\mu_i, \mu)}{\sum_j \text{count}(\mu_j, \mu)} \tag{21}$$

where $\sum_j \text{count}(\mu_j, \mu)$ is the frequency of occurrence of μ in *any* context. The lexical state vector $|\sigma\rangle$ is given by:

$$|\sigma\rangle = \frac{\sum_i \sqrt{\text{count}(\mu_i, \mu)}\, |\beta_i\rangle}{\sqrt{\sum_j \text{count}(\mu_j, \mu)}} \tag{22}$$

To verify that $|\sigma\rangle$ satisfies our requirements for a state vector, we check that it meets the two conditions laid out above.

1. $|\sigma\rangle$ is normalized.

$$\langle\sigma|\sigma\rangle = \frac{\sum_i \sqrt{\text{count}(\mu_i, \mu)}^2}{\sqrt{\sum_j \text{count}(\mu_j, \mu)}^2} \langle\beta_i|\beta_i\rangle = 1$$

2. For any basis element $|\beta_i\rangle$ indexed to outcome μ_i, $\langle\sigma|\beta_i\rangle\langle\beta_i|\sigma\rangle = P(\mu_i|\sigma)$.

$$\begin{aligned}
\langle\sigma|\beta_i\rangle\langle\beta_i|\sigma\rangle &= \frac{\sqrt{\text{count}(\mu_i, \mu)}}{\sqrt{\sum_j \text{count}(\mu_j, \mu)}} \langle\beta_i|\beta_i\rangle \frac{\sqrt{\text{count}(\mu_i, \mu)}}{\sqrt{\sum_j \text{count}(\mu_j, \mu)}} \langle\beta_i|\beta_i\rangle \\
&= \frac{\text{count}(\mu_i, \mu)}{\sum_j \text{count}(\mu_j, \mu)} \\
&= P(\mu_i|\sigma)
\end{aligned}$$

Hence a state vector produced from co-occurrence statistics in this way defines a probability distribution over all context outcomes for μ.

States can be compared by an analogue of the measurement procedure. An example is given by states in complementary distribution. The state σ can only be identified on the basis of some number of observations large enough to determine the expectation of each observation. However, orthogonal states are perfectly distinguishable. For instance, if a system is known to be either in state σ_a or σ_b where $a \neq b$, an observation $O = |\beta_i\rangle \langle\beta_i|$, $i = a$ or $i = b$, will determine with certainty which of the two states the system is in. This is because

$$\langle\sigma_a|\sigma_b\rangle = \sum_i \gamma_i^a \gamma_i^b \langle\beta_i|\beta_i\rangle = \gamma_i^a \gamma_i^b \tag{23}$$

where obviously $\gamma_i^a \neq 0$ and $\gamma_j^b \neq 0$ for some i, j, since these are normalized state vectors. Clearly then either γ_i^a or γ_i^b must be zero for every i for the products all to be zero, which means that σ_a and σ_b are in complementary distribution. Hence, an observation of any outcome μ_i for which $\langle\sigma_a|\beta_i\rangle \neq 0$ conclusively determines the state to be σ_a, and similarly for σ_b.

Non-orthogonal states are somewhere between the same and different. They occur in overlapping contexts, and so are not completely distinguishable. However, some states are more similar to one another than others, in that there is more overlap between their contexts, and the probability with which they occur in various contexts is similar. Whereas each member of the canonical basis is associated to a canonical observable O_i, each state σ may likewise be associated with a projection onto the one-dimensional subspace it defines, given by $O_\sigma = |\sigma\rangle \langle\sigma|$. Clearly, $\langle\sigma| O_\sigma |\sigma\rangle = \langle\sigma|\sigma\rangle\langle\sigma|\sigma\rangle = 1$, and for any state vector $|\psi\rangle \neq |\sigma\rangle$, $\langle\sigma| O_\psi |\sigma\rangle < 1$. Since state vectors have length 1, this last equation is just the square of the cosine distance of Equation (2), which varies monotonically with cosine distance. Hence this generalization of the observables to arbitrary state vectors defines an appropriate distance ordering between states.

4.4 A mixed state picture of adjective meanings

Given a corpus of sufficient size, one can construct a model of the signaller Σ of the form specified in section 4 on the basis of the co-occurrence statistics of each word μ. Co-occurrence statistics are then directly recoverable from the state space representation of σ corresponding to μ.

However, as seen in section 3, there are good reasons for assigning a sharply different type to adjectives than holds for nouns. Under the compositional semantic model, adjectives are minimally functions from noun-type denotations into noun-type denotations. In the model-theoretic case, these noun-type denotations are

properties corresponding to extensional predicates. In the distributional case, I follow Baroni and Zamparelli (2010) and Mitchell and Lapata (2010) in assuming that noun denotations are state vectors and adjective denotations are maps between state vectors. There are good reasons for assuming this typology. Common nouns and ANs form a syntactic substitution class. Given any instance of a noun occurring in some syntactic context, the noun may be substituted for an AN, preserving grammaticality. Similarly, common nouns and ANs are members of the same semantic class, in that they have the same type of denotation. In particular, both have the denotation of predicates, or functions $e \to t$. Hence, there is convergent evidence that they are expressions of the same basic type. Accordingly, they are assigned the same type in the distributional model, namely, that of lexical states. Adjectives, however, as in type-logical truth-conditional semantics, are assigned the type of a function over noun types. In the truth-conditional contexts, adjectives are modelled as functions of type $(e \to t) \to (e \to t)$. Correspondingly, adjectives are modelled as functions between lexical states, whose particular representations are derived from co-occurence contexts in corpora.

The theoretical model adopted here conceptualizes an adjective as a function over the noun state. We have evidence about the nature of this adjective function from the initial state of the noun and the terminal state represented by the state of the AN. The mapping is thus given for some finite set of a cases, from which it is possible to generalize about the action of the adjective in general. The analytical problem of characterizing the action of an adjective on a noun meaning thus has the form:

$$A := \vec{n} \longrightarrow \vec{an} \tag{24}$$

where A is an unknown process with the known property that it maps from any $\vec{n}$ to a corresponding $A\vec{n} = \vec{an}$.

Each A can therefore be thought of as a preparation process for noun states. A is a function that takes the lexical state of a noun as input, and produces a new lexical state as output. Even if the process itself remains impenetrable, the inputs and outputs may be observed; they are the states of nouns and ANs respectively. The relevant process may be imagined as a black box we wish to analyze. Though the equation mapping each input to each output is unknown, the map itself is known for some set of observed states.

The study of such functions is rendered difficult by the fact that they cannot be directly observed. Instead of directly modelling the adjective function, it is possible to study its properties from a more abstract point of view, without specifying the equation mapping inputs to outputs. The structure of the current study is observational. Given observational data about the input states (noun vectors) and the output states (AN vectors) of the adjective-process, we investigate how adjectives act on the whole spread of noun states they apply to.

This approach to modelling word meaning is, as far as I am aware, unique. Unlike other frameworks, such as (Baroni and Zamparelli, 2010), no assumption is made about the form of the adjective map, other than the stipulation that adjectives map distributions onto distributions. By comparison, Baroni and Zamparelli (2010) directly estimate adjective maps, in particular assuming that adjective functions have the form of linear maps over noun vectors. It is not at all clear, a priori, that this is the right model to learn for an adjective. While I assume that adjectives are maps from noun vectors into other vectors of the same type (living in the same space), I do not estimate the adjective maps directly. I only assume the presence of a hidden adjectival map.

4.5 Density operators: Mixed states

One additional formal device will enable the characterization of the uncertainty associated with adjective meanings. Consider a collection of states $\{\sigma_i\}$ any of which a system may be in when it is observed. This corresponds to the model laid out above for the meanings of adjective-noun pairs. We wish to consider the properties of a statistical ensemble defined by the presence in each case of the modelled adjective modifying the noun. For example, the adjective "big" will occur with the meaning of "big dog", "big elephant", and "big flea", each of which has some probability p_i of occurring, and none of which individually captures the range of variation for the meaning of "big". It is expected, in light of the generalization that the members of a set corresponding to a given adjective α have some property in common, that this fact will have a distributional effect reflected greater proximity between the ANs corresponding to the adjective. These adjectives will tend to be found in a similar region of the overall space. Conversely, non-intersective adjectives, whose meanings differ depending on the argument they are applied to, would exhibit a greater range of variation in their eventual location after being affected by the adjective. Density operators have been proposed for linguistic applications in (Blacoe et al., 2013), but to my knowledge, the first proposal for their use in analyzing polysemy using the von Neumann entropy is (Kartsaklis, 2014).

A model of the range of states that an adjective maps its noun arguments to will therefore consist of distinct observations of ANs, each of which is in a different state. Application of an attributive adjective therefore produces a "mixed state" of AN states, where, depending on which particular sample is drawn (which noun argument the adjective has been applied to), the resulting state of the AN will differ. In this mixed state, there are two types of uncertainty. There is uncertainty about the outcome of a random experiment given that the system is in any given state. In addition, there is uncertainty about what state the system will be in when it is observed. Such mixed states can be characterized, and the probabilities of any given outcome of a random experiment on them calculated, using density operators

conventionally denoted by the letter ρ.

Let $\{\langle|\sigma_i\rangle, p_i\}_i$ be an ensemble of states paired with the probability that the system is in that state. The density operator corresponding to this mixture is:

$$\rho = \sum_i p_i \rho_i = \sum_i p_i |\sigma_i\rangle \langle\sigma_i| \tag{25}$$

A special case of a density operator is the pure state, which can be written as the outer product of a state vector $|\sigma\rangle \langle\sigma|$.In calculations, each ρ_i is represented by a square density matrix whose i, j^{th} entries are the products of the corresponding entries of the state vectors $|\sigma\rangle$, $\langle\sigma|$.

Theorem 1. ***Spectral Theorem*** *Let Q be any normal operator on a vector space V. Then Q is diagonal with respect to some orthonormal basis for V.*[3]

ρ is a Hermitian operator (equal to its own conjugate transpose) and that, in the real numbers, it is represented by a symmetric matrix. The spectral theorem gives a canonical decomposition for Hermitian operators that is based on the eigenvectors of such operators. As a special case of the spectral theorem, a Hermitian operator can be decomposed into a weighted sum of outer products of its eigenvectors, all of which are pairwise orthogonal. This gives the spectral decomposition for ρ (Nielsen and Chuang, 2010).

Definition 3. *Spectral Decomposition Let ρ be an arbitrary density operator, and hence Hermitian. By* (1)*, ρ has a spectral decomposition*

$$\sum_i \lambda_i |\lambda_i\rangle \langle\lambda_i| \tag{26}$$

where the $|\lambda_i\rangle s$ are orthonormal eigenvectors of ρ with corresponding eigenvalues λ_i.[4]

Each $|\lambda_i\rangle \langle\lambda_i|$ are pure states. It is readily seen that the sum of the diagonal elements of each pure state ρ_i is equal to the inner product $\langle\sigma|\sigma\rangle$, and that when the $|\sigma_i\rangle$ are unit vectors weighted with factors p_i summing to 1, the sum of diagonal entries has value 1.

This quantity, called the trace $\mathrm{Tr}[\rho_i]$, is invariant across matrix representations of the density operator, and hence the trace is equal for density operators and

[3]If all eigenvalues for ρ are distinct, then these eigenvectors are orthogonal. If the eigenspace of ρ is degenerate, an orthonormal eigenbasis can be built for the degenerate subspace using the Gram-Schmidt procedure. These results will not be proven, but are well-known and thus assumed. For more detail, the reader is referred to (Nielsen and Chuang, 2010).

[4]For any operator A, $\vec{v}$ is an eigenvector of A if $A\vec{v} = \lambda\vec{v}$ for some scalar λ. In this case, λ is called the eigenvalue corresponding to eigenvector $\vec{v}$.

their spectral decompositions[5]. The measurement statistics for a given observable O corresponding to some message μ given a statistical mixture ρ are given by the trace:

$$\begin{aligned}\langle O\rangle &= \mathrm{Tr}\left[\rho O\right] \\ &= \mathrm{Tr}\left[\sum_i p_i \left|\sigma_i\right\rangle \left\langle\sigma_i\right| O\right] \\ &= \sum_i p_i \left\langle\sigma_i\right| O \left|\sigma_i\right\rangle \\ &= \sum_i p_i \times P(\mu|\sigma_i)\end{aligned}$$

The trace $Tr[\rho O]$ is equivalent to the probability of outcome μ for each state multiplied by the probability of that state (Nielsen and Chuang, 2010). The measurement statistics for a statistical ensemble of states can therefore be recovered from the mixed state expressed as a density operator.

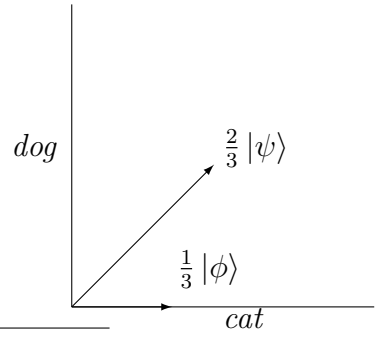
Example 1. *Consider a density operator that is a weighted sum of two pure states* $|\phi\rangle$ *and* $|\psi\rangle$ *represented in a lexical state space with only two basis elements* $|cat\rangle$, $|dog\rangle$*:*

$$|\phi\rangle = |dog\rangle \qquad P(\phi) = \frac{1}{3} \tag{27}$$

$$|\psi\rangle = \frac{|cat\rangle}{\sqrt{2}} + \frac{|dog\rangle}{\sqrt{2}} \qquad P(\psi) = \frac{2}{3} \tag{28}$$

$$\rho = \frac{1}{3}\left|\phi\right\rangle\left\langle\phi\right| + \frac{2}{3}\left|\psi\right\rangle\left\langle\psi\right| \tag{29}$$

The density operator ρ *represents a state that is a weighted mixture of these two pure states.*



[5]Let $\{\beta_i\}_i$ be any orthonormal basis for vector space Φ and A any linear operator on Φ. Then $\mathrm{Tr}[A] = \sum_i \left\langle\beta_i\right| A \left|\beta_i\right\rangle$.

Since these vectors are non-orthogonal, the orthogonal decomposition returns a new pair of vectors $\lambda_1 |e_1\rangle, \lambda_2 |e_2\rangle$ *that are orthogonal and capture the same co-occurrence statistics as* $|\phi\rangle$, $|\psi\rangle$.

4.6 Entropy of statistical mixtures

In addition to their link with co-occurrence statistics for uncertain distributions of lexical states, density operators are associated with a measure of entropy, which makes it possible to measure the degree of uncertainty about what state a given system is in. The application of this method to the analysis of semantic ambiguity is the work of Kartsaklis (2014).

Entropy is a quantity associated with a probability distribution that, depending on the interpretation, measures the uncertainty associated with the outcome of a random experiment–the distribution's level of bias towards some subset of possible outcomes–or expected value of the average number of bits needed to communicate the outcome of a random experiment given an optimal encoding of outcomes. Given random variable X with values x_i such that $P(x_i) = p_i$, the Shannon Entropy $H(X)$ is

$$H(X) = -\sum_i p_i \log(p_i). \tag{30}$$

The Shannon Entropy is minimized when a single outcome has probability 1, and maximized when every outcome has equal probability. A generalization of the Shannon entropy for vector spaces was given by Von Neumann as the formula:

$$S(\rho) = -\mathrm{Tr}[\rho \log(\rho)] \tag{31}$$

Since each density operator is equal to its spectral decomposition, in calculations, this quantity can be given as:

$$S(\rho) = -Tr\left[\sum_i \lambda_i |\lambda_i\rangle \langle\lambda_i|\right] \tag{32}$$

$$= -\sum_i \lambda_i \log(\lambda_i) \tag{33}$$

where $\sum_i \lambda_i |\lambda_i\rangle \langle\lambda_i|$ is the spectral decomposition for ρ. The symmetry between the formulae for the classical and von Neumann entropy is readily seen. In fact, the Shannon entropy is just the von Neumann entropy when the states are orthogonal (Petz, 2001).

The equality implies that the two distributions–the orthogonal decomposition and the original mixture of states and their probabilities–are equivalent in the measurement statistics that they produce. In general, there exist many decompositions

for a given non-pure ensemble ρ that are equivalent in this sense. Different decomposition of mixed states are indistinguishable on the basis of the results of measurements on them. However, the orthogonal decomposition is central in that its summands are pairwise orthogonal, and therefore represent entirely distinct states. The spectral decomposition gives the minimum number of vectors, along with weights between zero and one, needed to generate the co-occurrence statistics encoded in ρ.

For any pure state $\rho_s = |\sigma\rangle \langle\sigma|$, there is only a single eigenvector $|\sigma\rangle$, and it has eigenvalue 1. Hence $S(\rho_s) = \log(1) = 0$. Correspondingly, there is no uncertainty about the state of the system; it is unambiguously in state $|\sigma\rangle$. For mixed states, however, the system may be in any of several states, and moreover, these states may not be orthogonal–that is, they are somewhere between the same and different. However, their distributional properties, by the spectral theorem, can be reduced to the statistics produced by a set of orthogonal vectors. The von Neumann entropy is defined for these vectors.

4.7 Entropy of composite systems

In the experiments reported on here, nouns were modeled as composite systems consisting of one-word left and right contexts. However, primarily for computational reasons due to the large space of outcomes operated over in the characterization of lexical states, it is necessary to decompose systems into smaller systems assumed to be independent. The computation of entropy for joint independent systems, however, is thankfully straightforward.

Joint systems are produced from distinct systems via the tensor product $\otimes$. Let Φ, Ψ be state systems with bases $\{|\phi_i\rangle\}_i^m$ and $\{|\psi_j\rangle\}_i^n$. Then their tensor product $\Phi \otimes \Psi$ has the basis $\{|\phi_i\rangle \otimes |\psi_j\rangle\}_{i,j}^{mn}$, or more compactly, $\{|\phi_i\psi_j\rangle\}_{i,j}^{mn}$. For arbitrary scalar z and $|\phi\rangle \in \Phi$, $|\psi\rangle \in \Psi$, the tensor product satisfies:

1. $z\,|\phi\rangle \otimes |\psi\rangle = |\phi\rangle \otimes z\,|\psi\rangle = z(|\phi\rangle \otimes |\psi\rangle)$
2. $|\phi\rangle \otimes (|\psi_i\rangle + |\psi_j\rangle) = |\phi\rangle \otimes |\psi_i\rangle + |\phi\rangle \otimes |\psi_j\rangle$
3. $(|\phi_i\rangle + |\phi_j\rangle) \otimes |\psi\rangle = |\phi_i\rangle \otimes |\psi\rangle + |\phi_j\rangle \otimes |\psi\rangle$

The tensor product of operators is also defined. If A, B are operators over vector spaces Φ, Ψ, then the operator $A \otimes B$ is an operator over the vector space $\Phi \otimes \Psi$ with the property in (34).[6].

$$(A \otimes B)(|\phi\rangle \otimes |\psi\rangle) = A\,|\phi\rangle \otimes B\,|\psi\rangle \tag{34}$$

[6]For more details about the tensor product, with graphical representations of tensor states in a linguistic context, the reader is referred to Section 3 of (Clark, 2013)

It can be verified that if both $|\phi_i\rangle$ and $|\psi_j\rangle$ are of unit length, then their tensor product $|\phi_i\rangle \otimes |\psi_j\rangle$ is of unit length as well. Hence, the tensor product of two state vectors also defines a state vector, and likewise defines a probability distribution over outcomes. Specifically, the tensor product of two states gives the distribution of joint probabilities of outcomes on the two states when the two distributions are stochastically independent. We can therefore define:

Definition 4. *Let μ_i, μ_j be outcomes indexed to observables O_i, O_j, and let σ, ϕ be independent states.*

$$P(\mu_i, \mu_j|\sigma, \phi) = P(\mu_i|\sigma)P(\mu_j|\phi) = \langle\sigma\phi| \, O_i \otimes O_j \, |\sigma\phi\rangle \tag{35}$$

This picture generalizes to the density operator formalism, where the product of two independent mixed states ρ_1 and ρ_2 is simply obtained via their tensor product $\rho_1 \otimes \rho_2$. However, the representations of joint systems grow quite large. In the linguistic distributional model, each context position is a separate system. Each pair of distinct outcomes for that contextual position is mutually exclusive (orthogonal), and the probability of all outcomes in that position is equal to 1. Hence, for a given state σ of Σ, corresponding to μ, the first right and left contexts are outcomes of distinct systems, and so on for the second right and left contexts, and for the $n+1^{th}$ right and left contexts. The representations for such systems grow exponentially, and thus are difficult to compute over. The following theorem relates the entropy of a joint system to the entropy of the individual systems[7].

Theorem 2. *For a pair of density operators ρ^A, ρ^B, $S(\rho^A \otimes \rho^B) = S(\rho^A) + S(\rho^B)$.*

Proof. Since ρ^A, ρ^B are density operators, they have orthogonal decompositions $\sum_i \lambda_i \rho_i$, $\sum_j \lambda_j \rho_j$ where each $\rho_k = |\lambda_k\rangle \langle\lambda_k|$. It is clear that each $|\lambda_i\rangle \otimes |\lambda_j\rangle$ is an eigenvector of $\rho^A \otimes \rho^B$ with eigenvalue $\lambda_i \lambda_k$, since:

$$\begin{aligned}
(\rho^A \otimes \rho^B)(|\lambda_i\rangle \otimes |\lambda_j\rangle) &= \rho^A |\lambda_i\rangle \otimes \rho^B |\lambda_j\rangle \\
&= \sum_k \lambda_k |\lambda_k\rangle \langle\lambda_k|\lambda_i\rangle \otimes \sum_\ell \lambda_\ell |\lambda_\ell\rangle \langle\lambda_\ell|\lambda_j\rangle \\
&= \lambda_i |\lambda_i\rangle \otimes \lambda_j |\lambda_j\rangle \\
&= \lambda_i \lambda_j |\lambda_i\rangle \otimes |\lambda_j\rangle
\end{aligned}$$

The vectors $|\lambda_i\rangle \otimes |\lambda_j\rangle$ are pairwise orthogonal and form a basis for $A \otimes B$, and so

[7] This is stated as a theorem in (Nielsen and Chuang, 2010, p. 514), but not proven. The proof is original as far as I am aware.

they provide an orthogonal decomposition for $\rho^A \otimes \rho^B$. Therefore,

$$\begin{aligned}
S(\rho^A \otimes \rho^B) &= S\left(\sum_i \lambda_i \rho_i \otimes \sum_j \lambda_j \rho_j\right) \\
&= S\left(\sum_{i,j} \lambda_i \lambda_j \left|\lambda_i\right\rangle \left\langle\lambda_i\right| \otimes \left|\lambda_j\right\rangle \left\langle\lambda_j\right|\right) \\
&= \sum_{i,j} \lambda_i \lambda_j \log(\lambda_i \lambda_j) \\
&= \sum_{i,j} \lambda_i \lambda_j \log(\lambda_i) + \lambda_i \lambda_j \log(\lambda_j) \\
&= \sum_{i,j} \lambda_i \lambda_j \log(\lambda_i) + \sum_{i,j} \lambda_i \lambda_j \log(\lambda_j) \\
&= \sum_i \lambda_i \log(\lambda_i) + \sum_j \lambda_j \log(\lambda_j) \\
&= S(\rho^A) + S(\rho^B)
\end{aligned}$$

□

This means that the joint entropy of two systems is the sum of the entropy for the systems considered separately. This is essential from a computational point of view, since the size of joint systems grows exponentially with the the number of subsystems. In the experiments described in 5, I treat the right and left contexts as independent systems.

5 Experimental implementation

The implementation developed here for testing the effects on entropy of differing word classes in English is based on the formalism developed above that models lexical entries as probabilistic states derived from co-occurrence statistics in corpora. These experiments showed an unexpected, negative relationship between distributional entropy and intersectivity, with intersective adjectives having lower mean entropies than their non-intersective counterparts. Recall that intersective adjectives correspond to a set that is invariant across all attributive uses of the adjective, whereas non-intersective adjectives lack this property. In fact, non-intersective adjectives vary in meaning across predications, as a function of the meanings of their noun arguments. Despite results that contradicted the predictions of the theory, however, significant group differences between intersectives and non-intersectives were brought into sharp relief through measurements of their distributional entropy.

5.1 Corpus

In order to obtain the most accurate estimates of the distributional probabilities of each modeled lexical entry, a large 2.5-billion-word corpus was prepared based on a 2014 dump of Wikipedia concatenated with the British National Corpus (BNC). The Wikipedia corpus was cleaned using tools provided free by LinguaTools (Kolb, 2015). The corpus was lemmatized and tagged for part of speech using the TreeTagger (Schmid, 1995). POS tagging was necessary in order to identify nouns, ANs, and relevant contextual features.

Lemmatizing the corpus is a parametric choice. In this case, it was motivated by a belief that the distributional features relevant to determining semantic representations involve generalizing over derivational morphological variations of the same words. Whether a modelled noun or adjective-noun combination occurs to the left (or right) of *eat*, *eating*, or *ate* does not make a difference for determining its relationship to eating. Pre-processing the data in this way assumes a certain amount of grammatical knowledge among the language users being modelled; however, this does not seem like an unreasonable assumption.

5.2 Word vector models

Because the parts of speech of interest were exclusively attributive adjectives, these were isolated by finding adjacent pairs tagged as adjectives and nouns. Co-occurrence vectors were prepared by collecting all of the contexts of each modelled noun or AN, with a window of 1, and tagging each context as occurring either to the left or to the right of the modelled word. Co-occurrence vectors were constructed separately for nouns and ANs, in order to compare the entropy values for noun and AN lexical states.

All contexts were included in the raw counts, although eventual basis truncation was carried out as explained in section 5.5. However, only content words were separately counted in the model. These included nouns, verbs, other adjectives, and adverbs, but excluded determiners and other function words, since it was was not believed that these context words would contribute much to a model of meaning. For the remaining contexts, a dummy basis element $|\phi_0\rangle$ was included in the model to capture the contribution to the co-occurrence counts of these non-included elements.

Lexical vectors were compiled using two different formulae: normalized co-occurrence vectors, and lexical state vectors as defined in section 4.3. It is standard in distributional semantics to prepare word vectors as normalized versions of co-occurrence count vectors, or *normed co-occurrence vectors* (NCVs). These vectors are given by

$$norm(\overrightarrow{word}) = \frac{\overrightarrow{word}}{\sqrt{\overrightarrow{word} \cdot \overrightarrow{word}}} \tag{36}$$

where $\overrightarrow{word}$ is the raw co-occurrence count vector for *word* and $\overrightarrow{word} \cdot \overrightarrow{word}$ is the dot product of this vector with itself. The computation of lexical state vectors from raw co-occurrence counts followed the formula given by Equation 22, repeated here for convenience:

$$|\sigma\rangle = \frac{\sum_i \sqrt{\text{count}(\mu_i, \mu)}}{\sqrt{\sum_j \text{count}(\mu_j, \mu)}} |\beta_i\rangle \tag{37}$$

This model corresponds to the lexical state vectors discussed in section 4.3, and is preferred to the NCV model of (36) on theoretical grounds, due to the lexical state model's clear relation to co-occurrence probabilities. Moreover, the results reported in section 6 show that the effects on entropy of various word classes show up more prominently in the lexical state model. However, results for both sets of vectors are reported there.

Density operators for each adjective, modelled as a mixed state, were constructed from the state vectors for the ANs corresponding to the same adjective, and separately for the nouns that occurred somewhere as arguments of the given adjective. The probabilities assigned to each noun state in the construction of the noun density matrices were the probabilities of the noun occurring as an argument of the adjective, on the principle that what was being modeled was the action of the adjective on the set of nouns $N = \{\eta_i\}$ occurring as arguments of the adjective in the ratios $\{p_i\}$.

5.3 Joint left-right context model

Co-occurrence counts were gathered for a lexical state model including contexts to the left and right of the target word w in a window of 1, with the counts for the left and right window kept separate. Separate models were therefore compiled for the left and right contexts, with the lexical state for a modelled word being represented in a joint system $\Phi_r \otimes \Phi_\ell$.

In preparing the models for each lexical state $|\sigma_\ell\rangle \otimes |\sigma_r\rangle$, it was assumed that the left and right states were independent. It was therefore assumed that states in the product space $\Phi_\ell \otimes \Phi_r$ could be obtained as tensor products of the single-system states $\sigma_\ell \in \Phi_\ell$ and $\sigma_r \in \Phi_r$, motivating the separate collection of counts for each subsystem.

It was assumed that the subsystems are independent, which is equivalent to the assumption that each product state can be obtained as a tensor product of the single-system states. That is:

$$|\sigma_\ell \sigma_r\rangle = |\sigma_\ell\rangle \otimes |\sigma_r\rangle \tag{38}$$

Following Theorem 2, entropy calculations for the product states are expressed as a

simple sum of the entropies for each state individually, using the formula:

$$S(\rho^{\ell r}) = S(\rho^{\ell} \otimes \rho^{r}) = S(\rho^{\ell}) + S(\rho^{r}) \tag{39}$$

where each $\rho^k = \sum_i^n p_i \left|\sigma_k^i\right\rangle \left\langle\sigma_k^i\right|$ for every lexical state σ_k^i included in the statistical ensemble.

To be explicit about the alternative, it is in principle possible to directly collect observations of the product state $|\sigma_\ell\rangle \otimes |\sigma_r\rangle$ by taking counts of paired context occurrences. That is, for each pair $\langle w_i, w_j\rangle$ of context words corresponding to basis elements $|\beta_i\rangle$, $|\beta_j\rangle$ in Φ_ℓ, Φ_r, collect counts of contexts $\langle w_i, w_j\rangle$ to determine superposition coefficients for product spaces basis elements $|\beta_i\rangle \otimes |\beta_j\rangle$. Clearly, this results in a basis space of size $|\Phi_r| \times |\Phi_\ell|$, and since the joint entropy cannot be decomposed into a sum of independent subsystems, the corresponding density operators the square of that size, a truly immense space in which entropy calculations are clearly not feasible without severe basis truncation.

5.4 Adjective dataset

The sample of adjectives modeled was selected from the 400 adjectives in the corpus occurred most frequently in the attributive position, that is, immediately preceding a noun. The adjectives in this class occurred between 50,000 and 400,000 times, ensuring that sufficient evidence was available about the distribution of the ANs for each adjective to build adequate lexical state models for them. Only a handful of adjectives were excluded from this dataset in principle. Instead, the adjectives were manually coded for a number of semantic features indicated in 1, so that controls for each category could be selectively applied.

The semantic tags covered intersective (I) and non-intersective (N) adjectives, as well as some other classes like non-subsective (O) and intensional/privative (T). Given the relationship of the hypothesis to potential polysemy among uses of the adjective, polysemous words were also tagged with one of two characteristics: strong polysemy and weak polysemy. This ensured that a controlled comparison of non-intersective polysemy with other types of polysemy could be conducted. Unfortunately, the polysemy classifications could not help but be subject to a level of arbitrariness, since there are no clear, universally accepted standards for identifying polysemous words. Therefore, words were tagged in consultation with their dictionary entries in the Merriam-Webster Online edition (Merrian-Webster, 2015). If a word had multiple dictionary entries that were strongly related or very similar, they were not considered polysemous. However, if they had multiple entries that were somewhat related (Pw) or unrelated or only very loosely related (Ps), they were classified as polysemous.

Intersective and non-intersective adjectives were classified on the basis of the attributive propagation test. An adjective α was deemed to be intersective if, given

Tag	I	N	O
Meaning	Intersective	Non-intersective	Non-subsective
Examples	japanese black foreign dead annual christian	federal economic poor racial north powerful	possible potential past future initial
Tag	C	Ps	Pw
Meaning	Context-Dependent	Polysemous (strong)	Polysemous (weak)
Examples	short old young long strong low	classical lead civil critical right free	direct physical historical open global visual

Table 1: Examples of words coded with various semantic tags

the following premises, we have the conclusion:

$$\begin{array}{ll} x \text{ is an } \alpha\ \eta_1 & \text{(Premise 1)} \\ x \text{ is an } \eta_2 & \text{(Premise 2)} \\ \hline x \text{ is an } \alpha\ \eta_2 & \text{(Conclusion)} \end{array}$$

for every η_1, η_2 which is defined for α. As discussed in 3.2, this test is a better criterion for intersectivity than the predicative descent test, in which one asks whether the the adjective always licenses a deduction from the attributive use of the adjective to the predicative use.

Additionally, adjectives ruled non-intersective by the attributive propagation test were coded as context-dependent (C) if they satisfied the criterion proposed by Siegel (1976) and Partee (2007). According to these authors, certain adjectives are only apparently non-intersective–in reality, they are dependent on a context that shifts as a function of, among other things, the noun argument of the adjective. These adjectives can be identified by their distribution in *as-* versus *for*-phrases. While intersective but context-dependent adjectives like *tall* appear in predications like *John is tall for a jockey*, they seem very odd in contexts like *#John is tall as a jockey* with the noncomparative meaning that John is tall when he is considered in his capacity as a jockey. Conversely, true non-intersectives like *skillful*, in this version of the typology, readily appear in *as*-phrases, as in *John is skillful as a jockey.*

The entire dataset of adjectives, along with their associated tags, can be found in the Appendix.

5.5 Basis truncation

Distributional semantics typically works with very large co-occurrence vectors that are intended to capture co-occurrence preferences of different words. In practical applications, it is common to employ dimensionality reduction techniques such as singular value decomposition. However, such transformations lose their direct interpretability as states giving rise to the observed distribution of contexts, and hence they are avoided here. However, for the purposes of computation, it is crucial to realize that the vectors operated over in this study, though quite large, were truncated versions of the observed states. The size of co-occurrence bases for adjective-noun pairs were manageably small, but those for nouns were immense when all co-occurrences were taken into account–on the order of 10^5 on average. However, computing density matrices requires computing outer products for n-dimensional vectors, involving n^2 multiplications and addition of n^2-cell matrices. Given the computing resources available, these operations could not be performed in a reasonable for input vectors of greater than $n = 10,000$.

In order to render the computations feasible, the lexical state vectors were truncated to a basis of $k = 10^4$ basis vectors. That is, for each lexical state vector $|\sigma\rangle = \Sigma_i^n \gamma_i |\sigma_i\rangle$, only the reduced vector $|\sigma_r\rangle$ was employed, with coefficients γ_i giving the weight for each basis element.

$$|\sigma_r\rangle = |\sigma\rangle - \sum_{j=k+1}^{n} \gamma_j |\sigma_j\rangle = \sum_{i=1}^{k} \gamma_i |\sigma_i\rangle \tag{40}$$

However, by itself, this operation distorts the true co-occurrence statistics, since it is equivalent to the false assumption that $\gamma_i = 0$ for $k < i \leq n$, or the stipulation that the distribution of contexts is fully biased in favor of the included contexts.

Given that the full co-occurrence vectors cannot be used in entropy calculations due to computational limitations, a maximally plausible model of the remaining contexts can be obtained by making some simplifying assumptions that render the construction of density matrices possible. I will assume that the set of n observed co-occurring words spans the space of possible outcomes of σ, and correspondingly, that the n basis elements corresponding to these co-occurrence words span the space of possible observations O_i such that $\langle\sigma| O_i |\sigma\rangle \neq 0$. The simplest assumption consistent with the data included in the density operator computations is the maximum entropy principle. This principle holds that the unobserved outcomes (in this case, the wilfully ignored ones) should be distributed in such a way as to reflect maximal uncertainty for the result of a probabilistic experiment on this space of outcomes,

i.e. the distribution of outcomes on this part of the space is fully unbiased. The maximally uncertain distribution is the one in which the probability of all outcomes is equal. Hence, we hold that holds that for each lexical state σ and each pair of basis words w_i, w_j such that $k < i \leq j \leq n$,

$$P(w_i|\sigma) = P(w_j|\sigma) \tag{41}$$

This is just the null hypothesis with respect to the distribution of these words, i.e. that they occur randomly in state σ. However, these contexts should satisfy maximum entropy subject to an additional set of constraints, namely that the probability of each included outcome w_i is equal to the expected value $\langle O_i \rangle$ of its associated projective operator.

$$P(w_i|\sigma) = \langle \sigma|\beta_i\rangle\langle\beta_i|\sigma\rangle = \langle\sigma|\, O_i \,|\sigma\rangle = \langle O_i\rangle \tag{42}$$

The first $n - k$ elements of the basis should therefore be left untouched; they should reflect the probabilities the outcomes they are indexed to. However, the omitted portions should be represented in the vector computations in a way that satisfies the requirement of maximum entropy for those omitted outcomes. This can be done by setting $\gamma_i = \gamma_j$ for all $k < i \leq j \leq n$, that is, for all elements after the cutoff.

Since the probabilities for each omitted w_i must be equal, we have that

$$P(w_i|\sigma) = \frac{1}{(n-k)\sum_{i=k+1}^{n} c_i} \tag{43}$$

This leads to a constraint on lexical state vectors $|\sigma\rangle$ with omitted basis words $\{w_i\}_{k+1,n}$.

Proposition 1. *Let $|\beta_o\rangle = |\beta_{k+1}\rangle + |\beta_{k+2}\rangle + ... + |\beta_n\rangle$ be the sum of basis vectors corresponding to omitted context words $\{w_i\}_{k+1,n}$. Then for any state vector $|\sigma\rangle$, the projection $|\beta_i\rangle\langle\beta_i|\sigma\rangle$ of $|\sigma\rangle$ onto any $|\beta_i\rangle$ such that $k < i \leq n$ is equal to $\gamma\,|\beta_i\rangle$ for some fixed γ.*

$$\begin{aligned}
\sum_{i=k+1}^{n} \langle\sigma|\beta_i\rangle\langle\beta_i|\sigma\rangle &= \sum_{i=k+1}^{n} P(w_i|\sigma) \\
&= \frac{\sum_{i=k+1}^{n} c_i}{\sum_{j=1}^{n} c_j} \\
&= \sum_{i=k+1}^{n} \frac{1}{(n-k)\sum_{i=k+1}^{n} c_i} \\
&= \frac{1}{\sum_{i=k+1}^{n} c_i}
\end{aligned}$$

An immediate corollary is that the projection $|\beta_o\rangle \langle\beta_o|\sigma\rangle = \delta\,|\beta_o\rangle$ for a fixed δ given by $(n-k)\gamma$. This shows that the projection of any state vector onto the subspace spanned by $\{|\beta_i\rangle\}_{k+1,n}$ is always a scalar multiple of a single vector $|\beta_o\rangle$. The significance of this result is mainly pragmatic. It means that, given the assumption of maximum entropy for lexical states with omitted basis elements, the omitted portion of the full co-occurrence vector can be expressed as a scalar multiple of single basis element $|\beta^o\rangle$. Hence, the corresponding density operators can be expressed as sums of outer products of much smaller vectors $\sum_i p_i(|\sigma_i\rangle + \gamma_i^o\,|\beta_o\rangle)(\langle\sigma_i| + \gamma_i^{o*}\,\langle\beta_o|)$, with the value of γ_i^o given by the count of contexts omitted from the model.

These facts enable a dramatic simplification of the entropy calculations, since they amount to the replacement of all omitted basis elements with a single basis element, call it $|\beta_o\rangle$, that projects equally into each one-dimensional subspace defined by the omitted elements of the canonical basis. It is easy to see that, for any $1 \leq i \leq k$, $\langle\sigma_i|\beta_o\rangle = \langle\beta_i|\beta_o\rangle = 0$. Moreover,

$$\begin{aligned}\langle\sigma|\beta_o\rangle\langle\beta_o|\sigma\rangle &= \sum_{i=k+1}^{n} \rangle\sigma|\beta_i\rangle\langle\beta_i|\sigma\rangle = \sum_{i=k+1}^{n} \frac{1}{(n-k)\sum_{i=k+1}^{n} c_i} \\ &= \sum_{i=k+1}^{n} P(\mu_i|\sigma) = P(\mathrm{om})\end{aligned}$$

Therefore for any state vector $|\sigma\rangle$, the coefficient γ of $|\beta_o\rangle$is $\sqrt{P(\mathrm{om})}$.

In practice, it was found that on average, around 98% of left contexts and 96% of right contexts were found to correspond to observations of the 10^4 most frequent co-occurrence words (an instance of "the problem of the long tail"). This suggests that much of the information about the lexical states could be found in the most frequent segment of the basis, with the remaining contexts contributing minimally to the modelled similarity between lexical states, and correspondingly, contributing little to the entropy values. However, it should also be noted that the some information about the meaning of the words being modelled is lost in the truncation. In particular, it is likely these rarer words carry information about the specialized uses of a word that discriminate it from near neighbors.

One alternative choice of models for the omitted basis elements bears mentioning, though it is not implemented here. Assumption (41) is based on maximum entropy at the level of lexical states. However, at the level of density operators, the choice to optimize the entropy of states leads to a minimization of von Neumann entropy for the density operators encoding mixed states, given the constraint that the probability of any omitted context is equal to that of any other. This is due to the fact that the density operator constructed in this way is the lowest-rank possible choice of density operators to represent the statistical ensemble of truncated states $|\sigma\rangle$, adding only one linearly independent component to the eigenbasis of the density

operator ρ. In particular, if the operator corresponding to just the included contexts ρ_i has rank ℓ, then the operator for states $|\sigma^i\rangle + |\sigma^o\rangle$ has rank $\ell + 1$. The entropy values computed by this method should therefore be considered lower bounds on the true entropy values. It is expected that the true entropy values are higher, since the omitted components are highly unlikely to be linearly dependent. However, the true values are prohibitive to calculate.

Maximal von Neumann entropy is given by the density operator for the omitted portion of each summed pure state vector is fully biased in a direction orthogonal to all others. That is, for each pair of pure states $|\sigma^a\rangle$, $|\sigma^b\rangle$ such that $|\sigma^j\rangle = \left(\sum_{i=1}^{k} \gamma_i \left|\beta_i^j\right\rangle\right) + |\beta_o^j\rangle$, $\langle\beta_o^a|\beta_o^b\rangle = 0$. While this alternative was not carried out in the present study, it is important to be aware of it as a modelling choice.

5.6 Linking hypothesis

The link between von Neumann entropy and the set-theoretic characteristics can be summed up in the notion of uncertainty. Since the denotational meanings of non-intersectives are highly mutable, in that they vary with the arguments that they are applied to, it was expected that the distributionally-obtained meaning models for lexical entries should vary as well. Correspondingly, non-intersective ANs should be found scattered throughout the semantic space, projected into various various regions of it by the adjective map depending on which region they started in. The von Neumann entropy provides a means of quantifying the number of independent senses needed to characterize the spread of senses represented in a mixture of lexical states, as well as their relative probabilities of occurring.

From another point of view, we can consider directly what the effect of set intersection is on the entropy of a predicate calculus space. A binary predicate calculus may be considered a special case of a vector space calculus, one in which probabilities are either 1 or 0. To link the distributional hypothesis about entropy to the truth-conditional specification of the properties of different classes of adjectives, I will construct an explicit semantic space that expresses the relevant notions, and show the entropic effects of different adjective classes on this space manifest themselves. Finally, I will tentatively link this construction to the distributional vector space, which is the main empirical domain available to study.

To every distinct atomic predicate P_i is associated a space $\mathcal{B}_i$ spanned by an orthonormal basis $\{|0_i\rangle, |1_i\rangle\}$. A semantic space is a set $\{\bigotimes_i |\sigma\rangle_j \in \mathcal{B}_i \text{ s.t. } {}_j\langle\sigma|\sigma\rangle_j = 1\}_j$. That is, a semantic space is a tensor product of systems, each of which represents an elementary predicate. If the predicate P_i is true, then $|\sigma_i\rangle$ is set to $|1_i\rangle$, and it is set to $|0_i\rangle$ if P_i is false. If the predicate P_i is neither true nor false–if it is undetermined–then $|\sigma_i\rangle$ is some superposition of $|0_i\rangle$ and $|1_i\rangle$, perhaps $\frac{|0_i\rangle}{\sqrt{2}} + \frac{|1_i\rangle}{\sqrt{2}}$.

Any adjective is a function from predicates to predicates. If an adjective A is

intersective, then for some P_i, A sets P_i to True; that is, $A\,|\sigma_i\rangle = |1_i\rangle$, and this its only effect. This is just a direct translation of the set-theoretic definition of an intersective adjective. In this setting, an intersective "flips" P_i to $|1_i\rangle$, where P_i is the intersected set. The density operator for this subsystem is then $\sum_i p_i\,|1_i\rangle\,\langle 0_i|$ for each i. Since each state $|\sigma\rangle$ is by construction part of an independent joint system, by theorem 44, its total entropy can be expressed as a sum of the entropies of the individual subsystems.

$$S(\rho) = S(\rho_1) + S(\rho_2) + ... + S(\rho_i) + ... + S(\rho_n) \tag{44}$$

Moreover, since each $|\sigma_i\rangle$ in the joint system is set to $|1_i\rangle$ by A, the density operator ρ_i is in a pure state. Hence it has entropy 0. Supposing that even one $|\sigma_i\rangle \neq |1_i\rangle$ prior to adjectival modification, the resulting entropy of the whole system will be less than prior to the modification. Hence, whatever the starting state of the system, as long as the intersection is nontrivial–that is, as long as some subsystem did not start out in position $|1_i\rangle$–its entropy declines due to the adjectival modification. This proves that intersective modification reduces entropy in a predicate calculus.

Conversely, consider the model of a non-intersective adjective within this framework. In the intersective case, there is some set of spaces $\{\mathcal{B}_i\}_i$ such that the adjective maps each $|\sigma\rangle_i$ to $|1\rangle_i$. That is: $A\,|\sigma\rangle_i = |1\rangle_i$. The non-intersectivity condition states that there is no set $P = \bigcap_i P_i$ such that the denotation of the adjective is an intersection with the set P_i. Assuming subsectivity, every $|\sigma\rangle_i$ that is equal to either $|0\rangle_i$ or $|1\rangle_i$ is mapped back to itself; informally, determinate results of an observation of σ remain determinate. So the map A is the identity on these elements. As for the rest, their state may be altered by A in an undetermined way, with uncertain results on entropy. But the non-intersectivity condition guarantees that there is no subsystem $|\sigma\rangle)_i$ on which the values for each state affected by the adjective map will match. While this does not guarantee that the reduction in entropy for an intersective adjective will in each case decline to a greater degree than that for a non-intersective, it indicates that intersectives reliably decline in entropy in this boolean model, while this is not the case for non-intersectives.

The linking hypothesis rests on the empirical supposition that distributional patterns are correlated with, if not strictly determined by, the predicates entailed by the associated lexical entry. A strong version of this hypothesis is that there is some function from predicates to distributions such that the distributional contexts available to a given word are fully determined by the predications compatible with it. Various factors make this strong version of the distributional hypothesis unlikely to hold. For instance, the ubiquity of polysemy in language, which is reflected in a one-to-many mapping of words onto logical predicates, means that words can appear in contexts where some predicate they are related to is false, leading to non-complementary occurrence between these (sometimes) contradictory predicates and words.

A weaker version of this hypothesis states that distributional contexts are correlated with predicational attributes, and that this correlation leads to overall trends that can be analyzed with relatively large samples. It is not within the scope of this project to test these alternatives, or to decide the extent of the correspondence, in general, between model-theoretic characteristics of adjectives and their distributional similarities. However, I do evaluate one possible source of variation in the level of observed distributional entropy for adjectives, and link these to the model-thereoretic characterization of adjective meanings in an attempt to bring the two models closer together.

6 Results

The entropy values for both the AN and corresponding noun samples were compiled into a simple measure of the relative entropy quotient (REQ) pre- and post-adjectival modification.

Definition 5. ***Relative entropy quotient*** *The relative entropy quotient for an adjective is given by:*

$$\frac{S(\rho_{AN})}{S(\rho_N)} \tag{45}$$

where ρ_{AN} and ρ_N are, respectively, the density operators for the ANs corresponding to the adjective, and the adjective's noun arguments.

In order to assert a relationship between intersectivity and entropy, it is necessary to reject the null hypothesis that intersectivity has no effect on entropy. To evaluate the null hypothesis, I employ a standard t-test to compare the means of each sample.

The distinction between intersective and non-intersective adjectives is a binary classification, so a difference of means test for two independent samples is appropriate. The two-sample t-test compares the difference between means of two samples as a proportion of sample standard deviation. Table 2 gives the two-sample t-test statistic and associated p-values for several sub-samples of the dataset screened for a number of different features, using the NVC vectors. The test results are given for intersective and non-intersective adjectives, with strongly and weakly polysemous adjectives removed in turn.

The sign of the t-score indicates the direction of the effect. The theory-derived prior motivated by the linking hypothesis connecting entropy to interectivity predicted that entropy would be reduced for intersective adjectives. Contrary to these predictions, entropy was observed to textitincrease for intersective adjectives relative to their non-intersective counterparts. Moreover, this result has high confidence, given that $p \leq .005$ for all three trials, even controlling for polysemy. The robustness

Group 1	Group 2	Excluded	t-score	p-value
N ($n = 197$)	I ($n = 91$)	$\emptyset$	-2.9711	.003
N ($n = 181$)	I ($n = 80$)	Ps	-2.8489	.005
N ($n = 130$)	I ($n = 64$)	Ps, Pw	-2.9863	.003

Table 2: t-statistic and p-value for REQs of intersective and non-intersective sample (NVC model)

of this result is indicated by the persistence of the effect of intersectivity on entropy when polysemous adjectives are removed.

Table 3 displays the same results for the lexical state vectors model. The results from this set of word vectors indicate the same effect as those obtained from the normed NVC vectors, except that the observed effect is even stronger, and the confidence even greater. As with the vectors produced in the NVC model, the effect persists even when polysemous adjectives are removed from the sample, again displaying the robustness of the relationship between intersectivity and entropy. The increased magnitude and significance of the effects of intersectivity on entropy suggest that the lexical state model is superior to the usual NVC method of constructing word vectors, at least with respect to the measurement of differences in adjective meaning.

Group 1	Group 2	Excluded	t-score	p-value
N ($n = 197$)	I ($n = 91$)	$\emptyset$	-4.8737	.00001
N ($n = 181$)	I ($n = 80$)	Ps	-4.3198	.00003
N ($n = 130$)	I ($n = 64$)	Ps, Pw	-4.1889	.00005

Table 3: t-statistic and p-value for REQs of intersective and non-intersective sample (state vectors model)

While the NVC model showed no statistically significant difference between polysemous and non-polysemous adjectives, a substantial difference was found in the lexical state model, again suggesting that this model is more sensitive to the distributional effects of variations in meaning among adjectives. Table 4 shows the

Group 1	Group 2	Excluded	t-score	p-value
Ps ($n = 33$)	Sub ($n = 194$)	Pw	1.0896	.2771
Ps ($n = 33$)	Sub ($n = 194$)	Pw	1.0896	.2771
Ps, Pw ($n = 102$)	Sub ($n = 194$)	$\emptyset$	2.1548	.032

Table 4: t-statistic and p-value for REQs of polysemous and non-polysemous subsective adjectives (lexical state model)

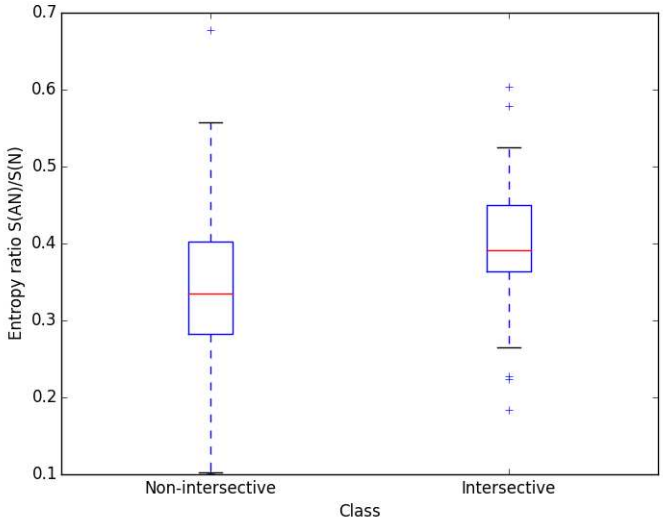


Figure 1: Box and whiskers plot of entropy ratios for intersectives and non-intersectives, controlling for strong and weak polysemy (state vectors model)

t-statistics for the REQ values of polysemous and non-polysemous adjectives in the lexical state model. The tag *Sub* indicates subsective (intersective or non-intersective) adjectives that were *not* polysemous, so the two samples compared in each row of Table 4 were disjoint. In the lexical state model, strong polysemy exhibited a small positive effect on entropy, though this effect was not statistically significant. Especially given that the sample of strongly polysemous adjectives was small ($n = 33$), the slight effect of strong polysemy on distributional entropy may have been produced by chance. However, when strongly *and* weakly polysemous adjectives were compared with non-polysemous subsective adjectives, the results conformed to the predictions delivered by the distributional hypothesis that lexical polysemy, and hence uncertainty about the meaning of any given instance of an adjective, is reflected in greater distributional entropy (Kartsaklis, 2014) ($p = 0.032$). This indicates that, for nouns and adjective-noun compounds, the positions of lexical state vectors in distributional space vary as a function of their spread of their variety of senses.

As a final test of the power of intersectivity to predict entropy levels, context-sensitive adjectives were separated from non-context-sensitive non-intersective adjectives, with the former being found to behave like their non-context-sensitive kin. Recall that context-sensitive adjectives are those for which the meaning of the adjective-noun compound appears to depend crucially on its context, rather than the noun argument (Partee, 2007). In this view, the noun argument is only a

part of the context, and hence the meaning of the adjective-noun compound is an intersection with a set, albeit one whose identity is set by the context. The meaning of such context-sensitive adjectives can therefore be modelled as a function with two arguments, a noun and a context. The denotation of such an adjective is thus:

$$\lambda C.\lambda N.a(C) \cap N \tag{46}$$

where a is a function of type $(e \to t) \to (e \to t)$. Such adjectives can be thought of as "intersective" even if they do not satisfy the attributive propagation test, in that the denotation of such an adjective applied to a noun η is the intersection of a contextually-determined set $a(C)$ and $[\![\eta]\!]$.

As observed earlier, if the contextual argument includes the noun argument, and if the meaning of the context-sensitive adjective shifts as a function of the noun as part of the context, then the meaning of the adjective will indeed shift as a function of the meaning of the noun argument. Hence, context-sensitive "intersectives" are predicted to behave like non-intersective adjectives. The results from this further segmentation of the dataset reported in Table 5 bear out this prediction. The tag C refers to context-context-dependent adjectives. Again, the groups compared are disjoint; if a tag X appears in the second column, it should be read *X and not C.*

Group 1	Group 2	Excluded	t-score	p-value
C ($n = 34$)	N ($n = 163$)	$\emptyset$	-0.512	.6092
C ($n = 28$)	N ($n = 102$)	Ps, Pw	-0.1116	.9113
C ($n = 34$)	I ($n = 91$)	$\emptyset$	-3.9942	.0001
C ($n = 28$)	I ($n = 64$)	Ps, Pw	-3.45899	.0008

Table 5: t-statistic and p-value for REQs of context-dependent adjectives (lexical state model)

In these results, context-sensitive adjectives are clearly seen to exhibit the same properties as non-intersectives, providing further evidence that they do not belong to the latter group. In rows 1 and 2, context-sensitive adjectives are compared with non-intersectives, controlling for polysemy in row 2. In both cases, the results are not statistically significant, showing that context-sensitive are not clearly separable from non-intersective adjectives that are *not* context-sensitive. This group relationship is further confirmed by rows 3 and 4, in which the same test is performed with intersectives. Rows 3 and 4 can be compared with rows 1 and 3 of Table 3, where similar values obtain between non-intersectives and intersectives. The same relationship is found between between context-sensitive adjectives and intersective adjectives as between non-intersective adjectives and intersective adjectives, indicating that context-sensitive adjectives belong in the same group as non-intersective

adjectives. Crucially, this further test displays the robustness of the relationship between distributional entropy and intersectivity.

The fact that non-intersective adjectives, which are more polysemous when considered from an extensional point of view, actually exhibit lower distributional entropy than their intersective counterparts, remains a puzzle that will have to be disentangled. However, this lack of correspondence between predicative entropy, as modelled in Section 5.6, and distributional entropy should not obscure the fact that the model successfully discriminated these two groups based on their levels of distributional entropy. While the observed direction of the effect is not currently understood, it clearly should be investigated further. These results indicates a strong connection between the two models of meaning that, at a superficial level, might not be expected to be connected at all. They also exhibit the capacity for distributional characteristics of lexical entries, even those whose meanings are one-place functions on other lexical categories, to predict rather fine-grained semantic properties.

From a practical point of view, the group differences observed between intersective and non-intersective adjectives in this study indicate the potential for an entropy-based approach in semantic classification tasks. Although probably not serviceable by itself as a single-feature classifier, it is clear that measurements of distributional entropy for adjectives provide a highly informative feature for predicting intersectivity, as well as polysemy. Identifying the inferential properties of lexical entries is essential to constructing computer programs that are able to reason using natural language, and so the identification of adjectives that license different entailments is an important task for computational linguists. Incorporating distributional entropy-based features into semantic classification systems promises to facilitate the automation of discovery processes for semantic types. In addition, the lexical state vectors collected for this study are both more theoretically motivated, and empirically more sensitive to polysemy than the typical normed vectors of distributional semantics, indicating that lexical state vectors could be a superior choice for applications that rely on semantic classification.

7 Conclusion

This study has taken several steps forward towards establishing the correspondence between the model-theoretic characterization of meaning employed in formal semantics, and the statistical representations of word meanings derived from analysis of corpora in distributional semantics. The relationship between these two aspects of word meaning are still poorly understood, but it is clear that progress is possible, and that such points of contact do exist.

In this paper, I motivated an empirical hypothesis about the level of entropy associated with two classes of attributive adjectives. In the model-theoretic charac-

terization, intersective adjectives are distinguished by the invariance of their interpretations across all uses, in the crucial sense that they always perform the same (extensional) operation on their noun arguments: intersection with a designated set. Non-intersective adjectives, by contrast, are highly mutable; their meanings vary across uses, generally as a function of the nouns that they are applied to. While other definitions of intersectivity do exist, this is the fundamental distinction between intersectives and non-intersective adjectives. The former are conjoined, across all uses, with a single property whose denotation always corresponds to the meaning contribution of the adjectives when they are used attributively. In the simplest analysis, intersective adjectives can simply be identified with this property. Non-intersective adjectives cannot be accounted for in this way.

In Section 4, I provided an explicit account of the model of word meaning underlying distributional semantic work, showing its clear connection to co-occurrence statistics derived from corpora of natural language. These statistics are interpreted in light of a model of probabilistic language states that, under observation, yield information about the similarity of distribution and, by hypothesis, the similarity of meaning of two lexical entries, when the state vectors representing them are mathematically compared. While these observable patterns of distribution are not directly identified with meanings, it is suggested that they are connected by some hidden function. Adjectives were modelled as statistical ensembles represented as density matrices, which are simply a representation the uncertainty associated with a lexical state.

From a density matrix representing a statistical mixture of lexical states, it is shown how to obtain a measure of uncertainty, the von Neumann entropy, associated with such statistical ensembles, and how this measure can be exploited to characterize the amount of uncertainty associated with the meaning of an adjective, modelled as a process affecting lexical distributions. These objects are exploited in a computational experiment described in Section 5. The results presented in Section 6 revealed a tight connection between distributional entropy and intersectivity, though the direction of this link was unexpected. Intersectivity was found to strongly predict an *increase* in entropy, in spite of the fact that intersective adjectives, unlike non-intersective adjectives, are model-theoretically monosemous in that they can be effectively identified with a single property.

These results were shown to be highly robust to controls for polysemy. Significantly, they provided confirmation of the exceptional nature of adjectives satisfying the strict definition of intersectivity embodied in the attributive propagation test, over and above alternative definitions of intersectivity. While it is not yet understood why intersective adjectives should have this unexpected property, the property itself is most assuredly observed, and deserves further study.

7.1 Future work

The present study is a first step into importing the concepts of distributional uncertainty into the study of compositional meaning construction. However, there remains much work to be done. While the probabilistic lexical state model of lexical meaning presented here is the only model I know of where co-occurrence statistics are treated as experimental results in a systematic fashion, the parameters of the experimental model employed here are highly limited. In particular, they may suffer from a weakness in feature selection, in that only single-word contexts to the right and left of the target word are considered. This model assumes extremely limited grammatical sophistication in the modelled language users. An immediate extension of this framework would be to investigate the entropy of adjectives in a dependency-parsed corpus. The features derivable from such a corpus would be expected to be much more informative than the one-word contexts modeled here, since they would include, for example, information about what verbs a word occurs as the subject or object of. Expanding the context window is also a possible extension, but dramatically increase the intensity of the involved computations.

Further studies should examine other means of measuring the uncertainty associated with word meanings, especially in a compositional setting. While much computational work has been focused on automatically identifying polysemy, or else on classifying instances of a polysemous word into one of its known senses, little attention has been given to quantifying the amount of semantic variation available to a single lexical sense depending on its immediate context, a topic which is at the heart of lexical *and* compositional semantics. In this thesis, I have shown that measures appropriate for performing the former task are also appropriate to the latter, broadening the range of known correspondences between distributional and set-theoretic features of adjectives.

References

Asher, N. (2011). *Lexical Meaning in Context: A web of words*. Cambridge University Press.

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. 2010 Association for Computational Linguists.

Blacoe, W., Kashefi, E., and Lapata, M. (2013). A quantum-theoretic approach

to distributional semantics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 847–857, Atlanta, Georgia. Association for Computational Linguistics.

Boleda, G., Vecchi, E. M., Cornudella, M., and McNally, L. (2012). First-order vs. higher-order modification in distributional semantics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1223–1233. Association for Computational Linguistics.

Bullinaria, J. and Levy, J. (2007). Extracting semantic representaitons from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Clark, S. (2013). Type-driven syntax and semantics for composing meaning vectors. *Quantum Physics and Linguistics: A Compositional, Diagrammatic Discourse*, pages 359–377.

Clark, S., Coecke, B., and Sadrzadeh, M. (2008). A compositional distributional model of meaning. In *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, pages 133–140.

Clark, S. and Pulman, S. (2007). Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction*, pages 52–55.

Coeke, B., Sadrzadeh, M., and Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. *CoRR*, abs/1003.4394.

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. Association for Computational Linguistics.

Grefenstette, E., Dinu, G., Zhang, Y.-Z., Sadrzadeh, M., and Baroni, M. (2013). Multi-step regression learning for compositional distributional semantics. *arXiv preprint arXiv:1301.6939*.

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. Association for Computational Linguistics.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Blackwell Oxford.

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. Association for Computational Linguistics.

Huddleston, R. and Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press.

Kamp, H. and Partee, B. (1995). Prototype theory and compositionality. *Cognition*, 57(2):129–191.

Kartsaklis, D. (2014). *Compositional Distributional Semantics with Compact Closed Categories and Frobenius Algebras*. PhD thesis, University of Oxford.

Kolb, P. (2015). linguatools.

Lenci, A. and Benotto, G. (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 75–79. Association for Computational Linguistics.

Merrian-Webster (2015).

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Mitchell, J. and Lapata, M. (2010). Composition in distributional semantics. *Cognitive Science*, 34:1388–1429.

Montague, R. (1974). English as a formal language. In Thomason, R., editor, *Formal Philosopy: selected papers of Richard Montague*, pages 188–222. Yale University Press.

Neumann, J. V. (1955). *Mathematical Foundations of Quantum Mechanics*. Princeton university press.

Nielsen, M. A. and Chuang, I. L. (2010). *Quantum Computation and Quantum Information*. Cambridge University Press.

Partee, B. (2007). Compositionality and coercion in semantics: The dynamics of adjective meaning. *Cognitive foundations of interpretation*, pages 145–161.

Petz, D. (2001). Entropy, von neumann and the von neumann entropy. In *John von Neumann and the Foundations of Quantum Physics*, pages 83–96. Springer.

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, 17:409–441.

Rimell, L. (2014). Distributional lexical entailment by topic coherence. *EACL 2014*, page 511.

Roller, S., Erk, K., and Boleda, G. (2014). Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of the Twenty Fifth International Conference on Computational Linguistics (COLING-14), Dublin, Ireland*.

Sadrzadeh, M., Clark, S., and Coecke, B. (2013). The frobenius anatomy of word meanings i: Subject and object relative pronouns. *Journal of Logic and Computation*, page ext044.

Sadrzadeh, M., Clark, S., and Coecke, B. (2014). The frobenius anatomy of word meanings ii: Possessive relative pronouns. *Journal of Logic and Computation*, page exu027.

Schmid, H. (1995). Treetagger: A language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27.

Siegel, M. E. A. (1976). *Capturing the Adjective*. PhD thesis, UMass Amherst.

Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics.

Susskind, L. and Friedman, A. (2014). *Quantum Mechanics: The theoretical minimum*. Basic Books.

8 Appendix: Adjective dataset with semantic tags ($n = 300$)

short C N
current N Pw
old C N
good Pw N

personal Ps
human I
young C N
long C N
important C N
private Pw I
average C N
full N Ps
common C N Ps
total N
popular N Pw
japanese I
modern N
official N
late Ps
natural Ps
external N Pw
australian I
notable N
regular N Pw
recent C N
top Ps N
black I Pw
economic N
female I
canadian I
italian I
real N Pw
western I Ps
free I Ps
historical I Pw
financial N
religious I Pw
legal Pw
medical I
commercial N Pw
white I
significant N Ps
indian I
little C N

musical N Pw
regional N Pw
administrative N
civil N Ps
foreign I
central N Pw
chinese I
traditional N Pw
federal N Pw
live I Ps
numerous N
independent N Pw
square I Ps
russian I
open N Pw
strong C N Pw
annual I
low C N Pw
cultural N Pw
senior Pw
spanish I
successful N
heavy C N
southern N
wide C N
northern N
physical I Pw
historic N Pw
jewish I
famous N
academic N Pw
greek I
active N Pw
key N
secondary N
minor N
whole N
right N Ps
red I
lead N Ps

irish I
direct N Pw
scientific N Pw
christian I
big C N
present N
eastern N
future O
rural I
critical N Ps
complete N
industrial N Pw
male I
digital I Pw
basic N Ps
technical N Pw
light C Pw N
ancient C N
limited N
dutch I
urban I
literary N Pw
soviet I
close N Pw
nuclear N Pw
standard N
upper N
electric I Ps
african I Pw
polish I
environmental N Pw
possible O
poor C N Pw
true N Pw
online N Pw
prominent N
positive N Ps
double N Ps
global N Pw
extensive N Pw

chief N
classical Ps I
contemporary I Ps
racial N
west N
executive N
royal N Pw
swedish I
domestic N Pw
municipal N
electronic I
permanent N
presidential N Pw
internal N
actual N
democratic I Pw
comic N
unique N
normal C N
native N
serious N Ps
agricultural N
south N Pw
naval N
scottish I
green I
daily I Pw
blue I
simple Pw N
east N
past O
north N
dark C N
principal N
nearby I
hard Ps
formal Ps
korean I
gold I
front N Pw

fictional I
assistant N
electoral N
extra N
brief N
alternative N Pw
powerful C N
criminal I Pw
mexican I
defensive I Ps
considerable N
potential O
bad N Pw
residential I Pw
norwegian I
visual N Pw
related N Ps
ethnic N
solo I
huge C N
provincial N Pw
secret N Pw
catholic I
conservative N Pw
junior N
maximum N
mobile I
wooden I Pw
parliamentary N
deep C N
corporate N
negative N
mental N
complex N
olympic N
mixed N Pw
tropical I
coastal I
artistic N
turkish I

latin I
outstanding N
armed I
legislative N
prime N
swiss I
israeli I
brazilian I
temporary N
sound I Pw
architectural N
amateur N Pw
unincorporated I
yellow I
hot C N
spiritual N
effective N
severe N Pw
fine N Pw
electrical N
immediate N
selected N
danish I
left N
medieval I Pw
creative N
clinical N
dry N Ps
founding N
inner N
portuguese I
massive C N
weekly I
classic N
vocal N
silver I
substantial N
rare N
typical N
strategic N

underground I Pw
muslim I
mass I
proper C N
marine N
colonial N
advanced N
rear N
grand C N
austrian I
animated I Ps
wild C N
rapid C N
martial N
constitutional N
beautiful N
narrow C N
detailed C N
vast C N
experimental C N
islamic I
clear C N Ps
unknown N
broad C N
operational N
asian I
distinct N
acoustic N
retail N
communist I
extreme N
straight Ps I
competitive N
hungarian I
indigenous N
romantic I Ps
honorary N
vertical N
diplomatic N
frequent N

civilian I Pw
mechanical N Ps
imperial N
excellent N
eldest N
solid Ps
outer N
solar N
extended N
moral N
constant N Pw
arab I
egyptian I
welsh I
first-class N
dead I