Doctoral School in Cognitive and Brain Sciences
XXVI Cycle

Doctoral Thesis

Distributional semantic phrases vs. semantic distributional nonsense: Adjective Modification in Compositional Distributional Semantics

Eva Maria Vecchi

Advisors: Prof. Roberto Zamparelli
Prof. Marco Baroni

December 2013

This thesis is dedicated to my eternally supportive family, for without their love, encouragement and patience I would have been lost.

Acknowledgements

There are so many people to acknowledge for the sweat and tears that has been put into the work of this thesis, it is hard to know where to begin. A whole chapter should be written just to thank my supportive and extremely patient advisors: Roberto Zamparelli, whose creative insight and intelligence has forced open doors in my research that I never thought possible, and Marco Baroni, whose work ethic, curiosity and determination has encouraged me to push myself to always learn, question and explore more each day. They have each taught me so much during these years, and I am a better researcher and person for this. I thank them both for their encouragement, advice and friendship throughout my Ph.D. career.

I would like to thank my Oversight Committee, particularly Raffaella Bernardi and Marco Marelli, for their time and invaluable suggestions.

Thank you to Gemma Boleda and Louise McNally for their contribution to my work, my scientific interest and growth. It has been an incredible pleasure working with you.

A very special thanks are dedicated to both Andrew Anderson, who has been a rock for me through it all, and Elia Bruni for his infinite patience and advice during these three years (even when I never get it right: "ventolaio"!).

I would also like to acknowledge the members of the COMPOSES team and the CLICers for all the support they provided; my friends and colleagues who made this experience so memorable and full of laughs, adventures and dancing; my family for their love and eternal patience; and Giorgio Membretti for his unconditional support.

Abstract

In this thesis, I study the ability of compositional distributional semantics to model adjective modification. I present three studies that explore the degree to which semantic intuitions are grounded in the distributional representations of adjective-noun phrases, as well as provide insight into various linguistic phenomena by extracting unsupervised cues from these distributional representations. First, I investigate degrees of adjective modification. I contrast three types of adjectival modifiers – intersectively used color terms, subsectively used color terms, and intensional adjectives – and test the ability of different composition strategies to model their behavior. Next, I propose an approach to characterize semantic deviance of composite expressions using distributional semantic methods. I present a set of simple measures extracted from distributional representations of words and phrases, and show that they are more significant in determining the acceptability of novel adjective-noun phrases than measures classically employed in studies of compound processing. Finally, I use compositional distributional semantic methods to investigate restrictions in adjective ordering. Specifically, I focus on properties distinguishing adjective-adjective-noun phrases in which there is flexibility in the adjective ordering from those bound to a rigid order. I explore a number of measures extracted from the distributional representation of such phrases which may indicate a word order restriction. Overall, this work provides strong support for compositional distributional semantics, as it is able to generalize and capture the complex semantic intuition of natural language speakers for adjective-noun phrases, even without being able to rely on co-occurrence relations between the constituents.

Publications

Parts of this thesis (ideas, figures, results and discussions) have appeared previously in the following publications:

Vecchi, E.M., Baroni, M. & Zamparelli, R. (2011). (Linear) maps of the impossible: Capturing semantic anomalies in distributional space. In *P*roceedings of the ACL Workshop on Distributional Semantics and Compositionality, 1–9, Portland, OR.

Boleda, G., Vecchi, E.M., Cornudella, M. & McNally, L. (2012). First-order vs. higher-order modification in distributional semantics. In *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*, 1223–1233, Jeju Island, Korea.

Vecchi, E.M., Marelli, M., Zamparelli, R. & Baroni, M. (2013). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *In Review.*

Vecchi, E.M., Zamparelli, R. & Baroni, M. (2013b). Studying the recursive behaviour of adjectival modification with compositional distributional semantics. In *Proceedings of the 2013 Conference on EMNLP*, 141–151, Seattle, WA.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, I study the ability of compositional distributional semantics to model adjective modification. In particular, I am interested in how such statistical approaches to meaning representation are able to approximate both the intuition of natural language speakers and the knowledge we have gained throughout generations of theoretical research on this topic.

The general aim of this research is twofold. First, I investigate how compositional distributional semantics of adjective-noun phrases is able to capture linguistic phenomena that have been concluded in previous literature. For example, a coherent model of adjective-noun semantics should be able to handle the difference between the degree of modification of the word *white* in the phrases *white shirt*, *white wine* and *white lie*. Further, strong evidence has been provided for the effects of adjective semantics on ordering restrictions in recursive modification. Ideally, models that properly model the semantics of adjective-noun phrases should encompass inherent properties that support this evidence.

Second, I hope to look at the other side of the coin, namely, study how compositional distributional semantics can provide insight to our understanding of natural language. In other words, I want to understand what specific semantic properties, which can be extracted from the distributional representation of phrases in a relatively painless and efficient manner, affect what we, as natural language speakers, just "understand". As a simple example, consider the intuitive difference between the phrases *sophisticated senator* and *legislative onion*. Clearly, an English speaker would respond that the latter is quite odd. That said,

even though they seem to share no properties, I would like to point out that both phrases are never found (or unattested) in an extremely large and comprehensive corpus, although their constituents are all found extremely frequently in the same corpus. The idea of what makes a novel phrase acceptable or unacceptable is interesting in many respects, and many fields can benefit from such knowledge (see Chapter 4 for more discussion). My goal is to exploit the properties of the distributional representations of such phrases to be able to gain a better understanding of this issue.

Semantics is the cognitive faculty that allows us to use language to reason and communicate about states of the world and of our minds (Chierchia & McConnell-Ginet, 2000). A number of fields have long sought to devise an artificial system endowed with human-like capabilities to understand and use natural language semantics. Researchers working on the development of this topic have explored a number of approaches to establish a system capable of this task. Although the approaches have varied a bit, the aims of such a system are consistent. First, it should be able to model semantic representations for naturally occurring text on a large scale. Second, the system should understand and incorporate the compositional aspect of natural language in order to combine words to construct and interpret sentences that have not yet been encountered by the system. Finally, it should agree with and even emulate human behaviour and intuition on a variety of semantic tasks.

Such a system that has the ability to retrieve and manipulate meaning could benefit a multitude of fields. In Cognitive Science, for example, it could aid in tasks such as Memory Retrieval, Categorization, Problem Solving, Reasoning and Learning. While in Theoretical Linguistics, it could promise a wider coverage of semantic analysis and free it of unrealistic idealizations. Natural Language Processing (NLP) could benefit from such a system in tasks like Question-Answering, Information Retrieval and Machine Translation.

One approach to semantic representations, Semantic Networks (Collins & Quillian, 1969), aims to represent concepts as nodes in a graph. The semantic relationships are denoted by the edges in the graph, while word meaning is expressed by the number and type of connections to other words in the graph. In this representation, we can model word similarity as a function of path length

– specifically, shorter paths for semantically related words.

However, this approach has been criticized as accounting for an idealized representation that is not completely in sync with real-world usage (Mitchell & Lapata, 2010). This criticism is based on the fact that the represenations are hand-coded by human modelers who a priori determine which relationships are most relevant in representing meaning. Attempts ot create semantic networks from word association norms (Steyvers & Tenenbaum, 2005) have aimed to overcome this, however they can only represent a small fraction of the vocabulary of an adult speaker.

An alternative to semantic networks exploit the idea that word meaning can be described in terms of feature lists (Smith & Medin, 1981). Specifically, each word is represented by a distribution of numerical values over a feature set. Although norming studies have the potential of revealing which dimensions of meaning are psychologically salient, working with such data gives rise to a number of difficulties such as the varying number and type of attributes generated, the large degree of freedom in the way responses are coded and analyzed, and asking multiple subjects to create a representation for each word limits the studies to a small-sized lexicon.

Another approach to natural language semantics, based on Montague Grammar (Montague, 1974), is built on predicate logic and lambda calculus. Such formal approaches are truth-conditional and model theoretic; that is, the meaning of a sentence is represented by its truth conditions which are expressed in terms of truth relative to some model of the world. In other words, a language is mapped onto a *set of possible worlds*, and other semantical notions are defined as *functions* on individuals and possible worlds. The meanings of referring expressions are taken to be entities/individuals in the model and predicates are functions from entities to truth-values (i.e. the meanings of propositions). These functions can also be characterised in an 'external' way in terms of sets in the model – this extended notion of reference is usually called denotation.

In this framework, the core semantic notion is the sentence, not the word. As a result, this approach has the advantage of focusing on the meaning of *function* words (such as determiners and conjunctions), which remains a weakness in other approaches. In addition, this approach allows composition to be car-

ried out syntactically (see below for further discussion on Formal Semantics and Compositionality). However, a drawback of this approach is that it lacks fine-grained representations for *content* words (such as adjectives and nouns), which is a centerpiece and advantage in other models. Such an approach also relies on ambiguous inputs, leading to expensive computations to compute all possible readings.

Distributional Semantic Models An approach to semantic representation that has gained a lot of attention in recent work, Distributional Semantic Models (DSMs), is based on the assumption that word meaning can be learned from the linguistic environment. According to the "distributional hypothesis of meaning" (Harris, 1968), words that are similar in meaning tend to occur in contexts of similar words, and thus meaning is susceptible to distributional analysis. DSMs aim to capture meaning *quantitatively* in terms of co-occurrence statistics from large collections of text, or corpora. Words are represented as vectors in a high-dimensional space, where each component corresponds to some co-occurring contextual element (Turney & Pantel, 2010).

An example implementation of this approach is the Hyperspace Analog to Language model (HAL, Lund & Burgess, 1996). In this model, each word is represented by a vector where each element of the vector corresponds to a weighted co-occurrence value of that word with some other word. Latent Semantic Analysis (LSA, Landauer & Dumais, 1997) also derives a high-dimensional space for words while using co-occurrence information between words and the passages they occur in. Both models are pioneer data-driven and wide-coverage DSM systems.

One major advantage of such an approach is that meaning is represented geometrically. Assuming similar words tend to occur in similar contexts, the distributional vectors point in similar directions, and therefore geometric distance approximates similarity in meaning (Bullinaria & Levy, 2007; Grefenstette, 1994; Lund & Burgess, 1996; Padó & Lapata, 2007; Schütze, 1997). In addition, these models are *unsupervised*, meaning they do not require manually labeled examples of target outputs to be trained, and *general-purpose*, in that a model is extracted once from a corpus (as a co-occurrence matrix) and can be used in a large variety of different semantic tasks (Baroni & Lenci, 2010). Semantic Space Models (and

its extensions, such as probabilistic topic models (Blei *et al.*, 2003; Griffiths *et al.*, 2007)) have also proved successful at simulating a wide range of psycholinguistic phenomena, for example, semantic priming (Griffiths *et al.*, 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996), word categorization (Laham, 2000), reading times (Griffiths *et al.*, 2007; McDonald, 2000), and judgments of semantic similarity (McDonald, 2000) and association (Denhière & Lemaire, 2004; Griffiths *et al.*, 2007). However, approaches that model semantic meaning this way are not naturally compositional, and most often vectors are combined by approaches that are insensitive to word order and syntactic structure.

Formal semantics and compositionality Speakers of a natural language are able to understand infinitely many sentences with different meanings. In fact, we are able to understand and produce sentences that have never before been heard or expressed based on our knowledge of the language. This productive capacity has been accredited to the compositional nature of natural language, a crucial property that allows us to derive the meaning of a complex linguistic constituent from the meaning of its immediate syntactic subconstituents (Frege, 1892; Partee, 2004).

Above, I introduced an approach to semantic representation based on predicate logic and lambda calculus. Logic-based frameworks in Formal Semantics (FS, Montague, 1974) are founded on the premise that there exists no theoretically relevant difference between artificial (formal) and natural (human) languages. In consequence, we can model logical structures of natural languages by means of universal algebra and mathematical (formal) logic. This framework is parallel to a syntactic system in which simple structures are put together into complex structures (Categorical Grammar) complex meanings are also constructed from simple meanings. FS aims to obtain first-order logic (FoL) representations of the meaning of a phrase compositionally through function application following the syntactic structure.

Frege's principle of *compositionality* (Frege, 1892) states whole meaning of a phrase can be described according to the functional interdependency of the meanings of its well-formed parts. Partee (1995) refines the principle further by taking into account the role of syntax: The meaning of the whole is a function

of the meaning of the parts and of the way they are syntactically combined. In other words, each syntactic operation of a formal language should have a corresponding semantic operation. This concept is illustrated in examples (1) and (2) from Landauer *et al.* (1997).

(1) It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.

(2) That day the office manager, who was drinking, hit the problem sales worker with the bottle, but it was not serious.

Compositionality is a matter of degree rather than a binary notion since linguistic structures range across levels of compositionality (Nunberg *et al.*, 1994). In simple cases, the meaning of an expression can be considered fully compositional, such as attributive adjective-noun phrases in which the meaning is the intersection of the meaning of the adjective and the noun, such as *red car*. Syntactically fixed expressions, such as *take advantage*, are only partly compositional because the constituents can still be assigned separate meanings. Certain idioms, such as *kick the bucket*, or multiword expressions, such as *by and large*, are considered much less compositional since their meaning cannot be distributed across their constituents.

Compositional distributional semantics

One the one hand, FS semantic representations in terms of logical formulas are able to represent and account for compositionality, however they are not well suited to modeling similarity quantitatively as they are based on discrete symbols. On the other hand, DSMs can easily measure similarity but they are not naturally compositional. As a result, current research in Computational Linguistics and Cognitive Science attempts to incorporate compositionality in DSMs (Baroni & Zamparelli, 2010; Erk & Padó, 2008; Guevara, 2010; Mitchell & Lapata, 2008). Following the insights gained from FS, the principle of compositionality, and current implementations of DSMs, this work aims at modeling the compositional phenomena in natural language semantics in a natural and linguistically relevant manner.

Semantic representations of single words can be represented as vectors in high-dimensional DSMs. By exploiting the geometric nature of these representations, given two independent vectors v_1 and v_2 in the space, we can then combine the independent vectors to produce a semantically compositional result v_3. Attempts in this task have explored a number of possible operations to combine these vectors, described in detail below. We can measure the success of such approaches in terms of their ability to model semantic properties of simple phrases, in tasks such as phrase similarity (Baroni & Zamparelli, 2010; Erk & Padó, 2008; Grefenstette & Sadrzadeh, 2011b; Mitchell & Lapata, 2010), textual entailment, semantic plausibility analysis (Vecchi *et al.*, 2011), and sentiment analysis (Socher *et al.*, 2011).

For the experiments presented in this thesis, we focus on various composition models in recent literature. The models and their parameter settings used in these experiments are described in detail in Section 2.2.

Outline

The thesis is structured as follows. In Chapter 2, I present a general framework for the experimental design which is then implemented in all experiments presented here. In this chapter, I describe the construction of the semantic space (Section 2.1) as well as the parameter tuning of the space (Section 2.1.4). In addition, I provide a general description of the implementation of each composition model (Section 2.2) and, again, provide details about the parameter estimation for each model (Section 2.2.1).

Chapter 3 introduces an experiment in which I use distributional semantic methods to detect differences in degrees of modification. I introduce the study with an overview of adjective semantics (Section 3.2). I then present the materials and methodology used to explore the research question (Section 3.3), and finally I provide an analysis of the results of the experiments (Section 3.4) and discussion (Section 3.5).

In Chapter 4, I present work that aims to detect semantic deviance in novel (unattested) phrases using unsupervised cues extracted from the generated distributional representation of the phrase. I first provide an introduction to the

question of "unattestedness" and semantic deviance (Section 4.1) and a description of related work on the topic (Section 4.2). I then present two studies to explore the issue. In Section 4.3, I introduce a pilot study which tests the feasibility of detecting semantic deviance and introduces preliminary measures. I present an analysis for these results (Section 4.3.3) and open the door for a more extensive study (Section 4.3.4). Section 4.4 introduces the extended study, providing a comparison to previous psycholinguistic analysis of acceptability of novel phrases. In this study, I expand the plausibility dataset to cover phrases containing nearly 700 adjectives (Section 4.4.1), and expand on the preliminary measures for semantic deviance (Section 4.4.2). I then analyze the results (Section 4.4.3) and conclude with a discussion (Section 4.4.4).

Chapter 5 is a study of the behavior of adjectives in recursive modification. I apply compositional models recursively (Section 5.3.2) to generate distributional representations of complex adjective-noun phrases, and extract information using these representation to gain a better understanding of adjective ordering restrictions. I construct an evaluation set of recursive adjective phrases (Section 5.3.4) and introduce measures to detect order restrictions (Section 5.3.3). The results are analyzed in detail (Section 5.4) and I close the chapter with a discussion and ideas for future work (Section 5.5).

Chapter 6 provides a general discussion and conclusions. I also provide a number of ideas for applications of the work presented here as well as future steps in the direction of the goals of this thesis.

Chapter 2

General experimental design

2.1 Semantic space

Our initial step was to construct a *semantic space* for our experiments, consisting of a matrix where each row represents the meaning of an adjective, noun or AN as a distributional vector. I first introduce the source corpus, then the vocabulary of words and ANs that I represent in the space, and finally the procedure adopted to build the vectors representing the vocabulary items from corpus statistics, and obtain the semantic space matrix. I work here with a "vanilla" semantic space (essentially, following the steps of Baroni & Zamparelli 2010), since our focus is on the effect of different composition methods given a common semantic space. In addition, Blacoe & Lapata (2012) found that a vanilla space of this sort performed best in their experiments.

2.1.1 Source corpus

I use as a source corpus the concatenation of the Web-derived ukWaC corpus (`http://wacky.sslmit.unibo.it/`), a mid-2009 dump of the English Wikipedia (`http://en.wikipedia.org`) and the British National Corpus (`http://www.natcorp.ox.ac.uk/`). The corpus has been tokenized, POS-tagged and lemmatized with the TreeTagger (Schmid, 1995), and it contains about 2.8 billion tokens. I extract all statistics at the lemma level, meaning that I consider only the canonical form of each word ignoring inflectional information, such as plural-

ization and verb inflection.

2.1.2 Semantic space vocabulary

The words and phrases in the semantic space must of course include the items that I need for our experiments (adjectives, nouns and ANs used for model training, as input to composition and for evaluation). Moreover, in order to study the behavior of the test items I are interested in (that is, model-generated AN vectors) within a large and less ad-hoc space I also include many more adjectives, nouns and ANs in our vocabulary not directly relevant to our experimental manipulations.

I first populate our semantic space with a core vocabulary containing the 8K most frequent nouns and the 4K most frequent adjectives from the corpus. In order to compare our experimental procedure to standard similarity judgment datasets, I included any adjective and noun used in Rubenstein & Goodenough (1965) and Mitchell & Lapata (2010). The vocabulary was then extended to include a large set of ANs (119K cumulatively), for a total of 132K vocabulary items in the semantic space.

To create the ANs needed to run and evaluate the experiments described below, I focused on adjectives which are very frequent in the corpus so that they generally be able to combine with many classes of nouns. I therefore define a **target vocabulary** containing the 700 most frequent adjectives and the 4K most frequent nouns in the corpus. Before generating the ANs, I manually controlled the target adjectives and nouns for problematic cases —adjectives such as *above*, *less*, or *very*, and nouns such as *cant*, *mph*, or *yours* – often due to parsing errors in the corpus. The ANs were generated by crossing the target nouns with the filtered 663 target adjectives and the filtered 3,910 target nouns, producing a set of 2.59M generated ANs.

I include those ANs that occur at least 100 times in the corpus in our vocabulary, which amounted to a total of 128K ANs. Of these ANs, 60% were randomly selected and used for training, circa 3% (10 ANs per target adjective) were used for the phase of parameter tuning described in Section 2.2.1 (this will be referred to as the *development set* in what follows); the rest was reserved to test the mod-

els. In addition, I included the set of 25 ANs used in Mitchell & Lapata (2010) in our vocabulary. To add further variety to the semantic space, I included a less controlled second set of 3.5K ANs randomly picked among those that are attested at least 100 times in the corpus and are formed by the combination of any of the adjectives and nouns in the core vocabulary.

2.1.3 Semantic space construction

For each of the items in our vocabulary, I first build 10K-dimensional vectors by recording the item's sentence-internal co-occurrence with the top 10K most frequent content words (nouns, adjectives, verbs or adverbs) in the corpus. I built a rank of these co-occurrence counts, and excluded from the dimensions any element of any POS whose rank was from 0 to 300 (the effect was to exclude any grammaticalized element from serving as a contextual dimension). The raw co-occurrence counts were then transformed into (positive) Pointwise Mutual Information (pPMI) scores, an association measure that closely approximates the commonly used Log-Likelihood Ratio while being simpler to compute (Baroni & Lenci, 2010; Evert, 2005). Specifically, given a row element r (here, the adjectives, nouns or ANs in the semantic space), a column element c (in this case, the 10K most frequent content words), and a join distribution $P(r, c)$, then

$$pmi(r, c) = \log \frac{P(r, c)}{P(r)P(c)} \tag{2.1}$$

$$ppmi(r, c) = pmi(r, c) \text{ if } pmi(r, c) \geq 0 \text{ else } 0 \tag{2.2}$$

Next, I reduce the full co-occurrence matrix applying the Non-negative Matrix Factorization (NMF) operation, a technique of dimensionality reduction that reduces a co-occurrence matrix into a lower dimensionality approximation with nonnegative factors. See Lee & Seung (2000) for references and discussion. I reduced in this way an original 12K-by-10K matrix composed of just the core vocabulary to a 12K-by-300 matrix. This step is motivated by the fact that I will estimate linear models to predict the values of each dimension of an AN from the dimensions of the components. I thus prefer to work in a smaller and denser

space. I then mapped the remaining 119K ANs in the semantic space to the 300 vectors of the NMF solution.

2.1.4 Semantic space parameter tuning

As a sanity check, I verify that I obtain state-of-the-art-range results on various semantic tasks using this reduced semantic space. Below, I explore additional methods of count-frequency transformation and dimensionality reductions found in the literature to confirm that our parameter settings are indeed optimal.

In the literature, transforming the raw co-occurrence counts to a measure of association between words has shown to be a very effective for sparse frequency counts (Baroni & Lenci, 2010; Dunning, 1993; Padó & Lapata, 2007). A number of transformations have been applied in recent studies of compositional distributional semantics (Baroni & Zamparelli, 2010; Boleda *et al.*, 2012; Vecchi *et al.*, 2013b), including (positive) Local Mutual Information (pLMI) and (positive) Log Weighting (pLOG). Given a row element r, a column element c, and a count of cooccurrence $count(r, c)$ (as for pPMI in Equation 2.2), I transform the count frequency with pLMI as shown in Equation 2.3, and I obtain the pLOG by simply taking the log of the count frequency, as shown in Equation 2.4.

$$plmi(r, c) = ppmi(r, c)count(r, c) = \log \frac{P(r, c)}{P(r)P(c)} count(r, c) \tag{2.3}$$

$$plog(r, c) = log(r, c) \text{ if } pmi(r, c) \geq 0 \text{ else } 0 \tag{2.4}$$

In addition to NMF, another common approach often used in dimensionality reduction is Singular Value Decomposition (SVD), a technique of that approximates a sparse co-occurrence matrix with a denser lower-rank matrix of the same size. See Turney & Pantel (2010) for references and discussion. This technique is used in LSA and related distributional semantic methods (Landauer & Dumais, 1997; Rapp, 2003; Schütze, 1997).

In order to evaluate the semantic space used in the experiments described in this thesis, I implemented a series of experiments to ensure state-of-the-art quality of the space. In Table 2.1, I report three quality evaluation experiments. I

first consider the correlation between the distance of noun vectors in the semantic space (described by their cosine distance) and human similarity judgments, based on the dataset provided in Rubenstein & Goodenough (1965) consisting of 65 noun pairs rated by 51 subjects on a 0-4 similarity scale. For example, the nouns *food* and *rooster* resulted in a low similarity rating, and this should therefore correlate to being further from each other in the semantic space than, say, *gem* and *jewel*.

Similarly, I compare the distance between word vectors in the semantic space and similarity judgments provided in the MEN dataset (Bruni *et al.*, 2012, `http://clic.cimec.unitn.it/~elia.bruni/MEN`). The MEN test dataset consists of 773 word pairs[1] (adjectives and nouns), randomly selected from words that occur at least 700 times in the freely available ukWaC and Wackypedia corpora combined (size: 1.9B and 820M tokens, respectively) and at least 50 times (as tags) in the opensourced subset `http://www.cs.cmu.edu/~biglou/resources/` of the ESP game dataset `http://en.wikipedia.org/wiki/ESP_game`. Each pair was randomly matched with a comparison pair and rated in this setting by participants of a crowdsourcing experiment using CrowdFlower `http://crowdflower.com/`. Each word pair was rated against 50 comparison pairs, thus obtaining a final score on a 50-point scale.

Finally, I consider a similar evaluation based on the correlation between distance in the semantic space and human similarity ratings of AN phrases, presented in the study of Mitchell & Lapata (2010) in which 72 AN phrases were judged on a 1-7 similarity scale. Again, phrases like *national government* and *cold air* obtained low similarity scores from the participants, and thus their AN vectors should have a lower cosine score than the vectors for the phrases *certain circumstance* and *particular case*.

Based on the results of these quality evaluation experiments, reported in Table 2.1, both the full and pPMI-transformed semantic spaces obtain state-of-the-art results. The best performing semantic space across the board is the space in which the raw cooccurrence counts are transformed with pPMI and the full

[1] Of the 1,000 word pairs in the MEN test set, our semantic space covered 773 of these data points. The coverage should be noted when comparing with state-of-the-art results reported in Table 2.1.

Weighting	*Reduction*	R&G	MEN	M&L
SoA		0.82	0.69	0.43
-	-	0.77	0.72	0.36
PPMI	SVD_{300}	0.72	0.69	0.38
	SVD_{50}	0.68	0.67	0.36
	NMF_{300}	**0.81**	**0.76**	**0.40**
	NMF_{50}	0.69	0.68	**0.40**
PLMI	SVD_{300}	0.70	0.70	**0.40**
	SVD_{50}	0.54	0.55	0.28
	NMF_{300}	0.70	0.68	0.06
	NMF_{50}	0.50	0.55	0.13
PLOG	SVD_{300}	0.40	0.38	0.32
	SVD_{50}	0.39	0.38	0.30
	NMF_{300}	0.62	0.63	**0.40**
	NMF_{50}	0.46	0.51	0.31

Table 2.1: **Semantic space parameter tuning.** The correlation scores (Spearman's ρ) between human similarity judgments of nouns (in the case of the R&G dataset), a mix of adjectives and nouns (in the case of the MEN dataset) or AN phrases (in the case of the M&L dataset) and the cosine distance between the vectors in the specified semantic space. The first row reports the state-of-the-art for each evaluation experiment based on the results reported in Baroni & Lenci (2010), for R&G, in Bruni *et al.* (2012), for MEN, and in Mitchell & Lapata (2010), for M&L. The second row reports the results of the *raw* semantic space, i.e., no transformation of the cooccurrence counts and in the 10K-dimension space. Results are provided for three weighting transformations (*ppmi*, *plmi*, *plog*), two dimensionality reduction approaches (*svd*, *nmf*) and two reduced sizes (*50*, *300*). The best results are in bold.

12K-by-10K space is reduced to 12K-by-300 with NMF.

2.2 Composition models

I focus on six composition functions proposed in recent literature with high performance in a number of semantic tasks. I first consider methods proposed by Mitchell & Lapata (2010) in which the model-generated vectors are simply obtained through component-wise operations on the constituent vectors. Given input vectors $\vec{u}$ and $\vec{v}$, Mitchell & Lapata derive two simplified models from these general forms. The first of which is the simplified additive model (ADD), given by Equation 2.5, and can be extended to the weighted additive model (W.ADD) in which a composed vector is obtained as a weighted sum of the two component vectors, Equation 2.6, where α and β are scalars.

$$\vec{c} = \vec{u} + \vec{v} \tag{2.5}$$

$$\vec{c} = \alpha\vec{u} + \beta\vec{v} \tag{2.6}$$

Next, a simplified multiplicative (MULT) approach that reduces to component-wise multiplication, where the i-th component of the composed vector is given by: $p_i = u_i v_i$, generalized by Equation 2.7.

$$\vec{c} = \vec{u} \odot \vec{v} \tag{2.7}$$

Mitchell & Lapata extend the multiplicative approach to a basis-independent composition which is based solely on the geometry of **u** and **v**, referred to here as the dilation method (DL):

$$\vec{c} = (\vec{u} \cdot \vec{u})\vec{v} + (\lambda - 1)(\vec{u} \cdot \vec{v})\vec{u} \tag{2.8}$$

where $\vec{v}$ is dilated along the direction of $\vec{u}$ by a factor λ. Here, the intuition is that the action of combining two words can result in specific semantic aspects becoming more salient, hence an action of dilation which stretches $\vec{v}$ differentially

to emphasize the contribution of $\vec{u}$.

Mitchell & Lapata evaluate the simplified models on a wide range of tasks ranging from paraphrasing to statistical language modeling to predicting similarity intuitions. Both simple models fare quite well across tasks and alternative semantic representations, also when compared to more complex methods derived from the equations above. Given their overall simplicity, good performance and the fact that they have also been extensively tested in other studies (Baroni & Zamparelli, 2010; Erk & Padó, 2008; Guevara, 2010; Kintsch, 2001; Landauer & Dumais, 1997), I re-implement here the ADD, W.ADD, MULT and DL models. In addition to finding that the MULT, W.ADD and DL models perform best overall, Mitchell & Lapata (2010) observed that the DL models performed consistently well across all representations.

Mitchell & Lapata (as well as earlier researchers) do not exploit corpus evidence about the $\vec{c}$ vectors that result from composition, despite the fact that it is straightforward (at least for short constructions) to extract direct distributional evidence about the composite items from the corpus (just collect co-occurrence information for the composite item from windows around the contexts in which it occurs). Here, I also consider the full extension of the additive model (F.ADD), presented in Guevara (2010) and Zanzotto *et al.* (2010), such that the component vectors are pre-multiplied by weight matrices before being added, Equation 2.9:

$$\vec{c} = \mathbf{W}_1\vec{u} + \mathbf{W}_2\vec{v} \tag{2.9}$$

The main innovation of Guevara (2010), who focuses on adjective-noun combinations (AN), is to use the co-occurrence vectors of corpus-observed ANs to train a supervised composition model. Guevara adopts the full additive composition form from Equation (2.9) and he estimates the $\mathbf{W}_1$ and $\mathbf{W}_2$ weights (concatenated into a single matrix, that acts as a linear map from the space of concatenated adjective and noun vectors onto the AN vector space) using partial least squares regression. The training data are pairs of adjective-noun vector concatenations, as input, and corpus-derived AN vectors, as output. Guevara compares his model to the ADD and MULT models of Mitchell & Lapata. Corpus-extracted ANs are nearer, in the space of corpus-extracted and model-generated test set ANs, to the

ANs generated by his model than to those from the alternative approaches. The ADD model, on the other hand, is best in terms of shared neighbor count between corpus-extracted and model-generated ANs.

Finally, I consider the lexical function model (LFM), first introduced in Baroni & Zamparelli (2010), in which attributive adjectives are treated as functions from noun meanings to noun meanings. This is a standard approach in Montague semantics Thomason (1974), except noun meanings here are distributional vectors, not denotations, and adjectives are (linear) functions learned from a large corpus. In this model, composed vectors are generated by multiplying a function matrix $\mathbf{U}$, representing the adjective at hand, with a component (noun) vector, Equation 2.10.

$$\vec{c} = \mathbf{U}\vec{v} \tag{2.10}$$

In Baroni & Zamparelli (2010), they show that the model significantly outperforms other vector composition methods, including ADD, MULT and F.ADD, in the task of approximating the correct vectors for previously unseen (but corpus-attested) ANs.

2.2.1 Composition model estimation

Parameters for W.ADD, DL, F.ADD and LFM were estimated following the strategy proposed by Guevara (2010) and Baroni & Zamparelli (2010), recently extended to all composition models by Dinu *et al.* (2013b). Specifically, I learn parameter values that optimize the mapping from the noun to the AN as seen in examples of corpus-extracted N-AN vector pairs, using least-squares methods for all models except LFM. All parameter estimations and phrase compositions were implemented using the DISSECT toolkit (Dinu *et al.*, 2013a, `http://clic.cimec.unitn.it/composes/toolkit`), with a training set of 74,767 corpus-extracted N-AN vector pairs, ranging from 100 to over 1K items across the 663 adjectives. Table 2.2 reports the results attained by our model implementations on the Mitchell & Lapata AN similarity data set.

For the LFM, the weights of each of the 300 rows of the weight matrix are the coefficients of a linear equation predicting the values of one of the dimensions of the AN vector as a linear combination of the 300 dimensions of the component

Model	ρ	*M&L*	*Parameter*
CORP	0.40	0.43	
ADD	0.34	0.37	
W.ADD	0.35	0.44	$\alpha = 0.31$, $\beta = 0.46$
MULT	0.31	0.46	
DL	0.32	0.44	$\lambda = 1.59$
F.ADD	0.35	–	
LFM	0.38	–	

Table 2.2: **Composed space quality evaluation.** Correlation scores (Spearman's ρ, all significant at $p<0.001$) between cosines of corpus-extracted (CORP) or model-generated AN vectors and phrase similarity ratings collected in Mitchell & Lapata (2010), as well as best reported results from Mitchell & Lapata (M&L).

noun. The linear equation coefficients were estimated separately for each adjective using Ridge regression with generalized cross-validation (GCV) to automatically choose the optimal Ridge parameter for each adjective (Golub *et al.*, 1979). For each adjective, the training N-AN vector pairs chosen were those available in the training set.

As a quality control, I verified that the composition models with the parameter settings chosen in the previous step obtained state-of-the-art results in a phrase similarity task presented in Mitchell & Lapata (2010). In this study, the authors asked participants to rate the similarity between pairs of AN phrases that encompassed a range of 3 similarity levels (high, medium and low similarity). They then tested the ability of composition functions to model these human judgments by looking at the correlation of the human similarity scores for the AN pairs with the cosine distance of their model-generated vectors. I replicated this experiment with each of the composition models. Table 2.2 shows that we obtain similar correlation scores to those reported in Mitchell & Lapata (2010). Further, I find that the LFM performs best in comparison to other composition models.

Chapter 3

Degrees of adjective modification in distributional semantics

3.1 Introduction

One of the most appealing aspects of so-called distributional semantic models (see Turney & Pantel (2010) for a recent overview) is that they afford some hope for a non-trivial, computationally tractable treatment of the context dependence of lexical meaning that might also approximate in interesting ways the psychological representation of that meaning (Andrews *et al.*, 2009). However, in order to have a complete theory of natural language meaning, these models must be supplied with or connected to a compositional semantics; otherwise, we will have no account of the recursive potential that natural language affords for the construction of novel complex contents.

In the last 4-5 years, researchers have begun to introduce compositional operations on distributional semantic representations, for instance to combine verbs with their arguments or adjectives with nouns (Baroni & Zamparelli, 2010; Erk & Padó, 2008; Grefenstette & Sadrzadeh, 2011b; Mitchell & Lapata, 2010; Socher *et al.*, 2011)[1]. Although the proposed operations have shown varying degrees of success in a number of tasks such as detecting phrase similarity and paraphrasing,

[1]In a complementary direction, Garrette *et al.* (2011) connect distributional representations of lexical semantics to logic-based compositional semantics.

it remains unclear to what extent they can account for the full range of meaning composition phenomena found in natural language. Higher-order modification (that is, modification that cannot obviously be modeled as property intersection, in contrast to first-order modification, which can) presents one such challenge, as we will detail in the next section.

The goal of this chapter is twofold. First, we examine how the properties of different types of adjectival modifiers, both in isolation and in combination with nouns, are represented in distributional models. We take as a case study three groups of adjectives: 1) color terms used to ascribe true color properties (referred to here as **intersective** color terms), as prototypical representative of first-order modifiers; 2) color terms used to ascribe properties other than simple color (here, **subsective** color terms), as representatives of expressions that could in principle be given a well-motivated first-order or higher-order analysis; and 3) **intensional** adjectives (e.g. *former*), as representative of modifiers that arguably require a higher-order analysis. Formal semantic models tend to group the second and third groups together, despite the existence of some natural language data that questions this grouping. However, our results show that all three types of modifiers behave differently from each other, suggesting that their semantic treatment needs to be differentiated.

Second, we test how five different composition functions that have been proposed in recent literature fare in predicting the attested properties of nominals modified by each type of adjective. The model by Baroni & Zamparelli (2010) emerges as a suitable model of adjectival composition, while multiplication and addition shed mixed results.

The chapter is structured as follows. Section 3.2 provides the necessary background on the semantics of adjectival modification. Section 3.3 presents the methods used in our study. Section 3.4.1 describes the characteristics of the different types of adjectival modification, and Section 3.4.2, the results of the composition operations. The chapter concludes with a general discussion of the results and prospects for future work.

3.2 The semantics of adjectival modification

Accounting for inference in language is an important concern of semantic theory. Perhaps for this reason, within the formal semantics tradition the most influential classification of adjectives is based on the inferences they license (see Parsons (1970) and Kamp (1975) for early discussion). We very briefly review this classification here.

First, so called intersective adjectives, such as (the literally used) *white* in *white dress*, yield the inference that both the property contributed by the adjective and that contributed by the noun hold of the individual described; in other words, a white dress is white and is a dress. The semantics for such modifiers is easily characterized in terms of the intersection of two first-order properties, that is, properties that can be ascribed to individuals.

On the other extreme, intensional adjectives, such as *former* or *alleged* in *former/alleged criminal*, do not license the inference that either of the properties holds of the individual to which the modified nominal is ascribed. Indeed, such adjectives cannot be used as predicates at all:

(1) ??The criminal was former/alleged.

The infelicity of (1) is generally attributed to the fact that these adjectives do not describe individuals directly but rather effect more complex operations on the meaning of the modified noun. It is for this reason that these adjectives can be considered higher-order modifiers: they behave as properties of properties. Though rather abstract, the higher-order analysis is straightforwardly implementable in formal semantic models and captures a range of linguistic facts successfully.

Finally, subsective adjectives such as (the non-literally-used) *white* in *white wine*, consitute an intermediate case: they license the inference that the property denoted by the noun holds of the individual being described, but not the property contributed by the adjective. That is, white wine is not white but rather a color that we would probably call some shade of yellow. This use of color terms, in general, is distinguished primarily by the fact that color serves as a proxy for another property that is related to color (e.g. type of grape), though the

color in question may or may not match the color identified by the adjective on the intersective use (see Gärdenfors (2000) and Kennedy & McNally (2010) for discussion and analysis). The effect of the adjective, rather than to identify a value for an incidental COLOR attribute of an object, is often to characterize a subclass of the class described by the noun (white wine is a kind of wine, brown rice a kind of rice, etc.).

This use of color terms can be modeled by property intersection in formal semantic models only if the term is previously disambiguated or allowed to depend on context for its precise denotation. However, it is easily modeled if the adjective denotes a (higher-order) function from properties (e.g. that denoted by *wine*) to properties (that denoted by *white wine*), since the output of the function denoted by the color term can be made to depend on the input it receives from the noun meaning. Nonetheless, there is ample evidence in natural language that a first-order analysis of the subsective color terms would be preferable, as they share more features with predicative adjectives such as *happy* than they do with adjectives such as *former*.

The trio of intersective color terms, subsective color terms, and intensional adjectives provides fertile ground for exploring the different composition functions that have been proposed for distributional semantic representations. Most of these functions start from the assumption that composition takes pairs of vectors (e.g. a verb vector and a noun vector) and returns another vector (e.g. a vector for the verb with the noun as its complement), usually by some version of vector addition or multiplication (Erk & Padó, 2008; Grefenstette & Sadrzadeh, 2011b; Mitchell & Lapata, 2010). Such functions, insofar as they yield representations which strengthen distributional features shared by the component vectors, would be expected to model intersective modification.

Consider the example of *white dress*. We might expect the vector for *dress* to include non-zero frequencies for words such as *wedding* and *funeral*. The vector for *white*, on the other hand, is likely to have higher frequencies for *wedding* than for *funeral*, at least in corpora obtained from the U.S. and the U.K. Combining the two vectors with an additive or multiplicative operation should rightly yield a vector for *white dress* which assigns a higher frequency to *wedding* than to *funeral*.

Additive and multiplicative functions might also be expected to handle subsective modification with some success because these operations provide a natural account for how polysemy is resolved in meaning composition. Thus, the vector that results from adding or multiplying the vector for *white* with that for *dress* should differ in crucial features from the one that results from combining the same vector for *white* with that for *wine*. For example, depending on the details of the algorithm used, we should find the frequencies of words such as *snow* or *milky* weakened and words like *straw* or *yellow* strengthened in combination with *wine*, insofar as the former words are less likely than the latter to occur in contexts where *white* describes wine than in those where it describes dresses. In contrast, it is not immediately obvious how these operations would fare with intensional adjectives such as *former*. In particular, it is not clear what specific distributional features of the adjective would capture the effect that the adjective has on the meaning of the resulting modified nominal.

Interestingly, recent approaches to the semantic composition of adjectives with nouns such as Baroni & Zamparelli (2010) and Guevara (2010) draw on the classical analysis of adjectives within the Montagovian tradition of formal semantic theory (Montague, 1974), on which they are treated as higher order predicates, and model adjectives as matrices of weights that are applied to noun vectors. On such models, the distributional properties of observed occurrences of adjective-noun pairs are used to induce the effect of adjectives on nouns. Insofar as it is grounded in the intuition that adjective meanings should be modeled as mappings from noun meanings to adjective-noun meanings, the matrix analysis might be expected to perform better than additive or multiplicative models for adjective-noun combinations when there is evidence that the adjective denotes only a higher-order property. There is also no *a priori* reason to think that it would fare more poorly at modeling the intersective and subsective adjectives than would additive or multiplicative analyses, given its generality.

In this chapter, we present the first studies that we know of that explore these expectations.

3.3 Methodology

3.3.1 Evaluation material

We built two datasets of adjective-noun phrases for the present research, one with color terms and one with intensional adjectives.[1]

Color terms. This dataset is populated with a randomly selected set of adjective-noun pairs from the space presented above. From the 11 colors in the basic set proposed by Berlin & Kay (1969), we cover 7 (*black*, *blue*, *brown*, *green*, *red*, *white*, and *yellow*), since the remaining (*grey*, *orange*, *pink*, and *purple*) are not in the 700 most frequent set of adjectives in the corpora used. From an original set of 412 ANs, 43 were manually removed because of suspected parsing errors (e.g. *white photograph*, for *black and white photograph*) or because the head noun was semantically transparent (*white variety*). The remaining 369 ANs were tagged independently by the second and fourth authors of Boleda *et al.* (2012), both native English speaker linguists, as **intersective** (e.g. *white towel*), **subsective** (e.g. *white wine*), or **idiomatic**, i.e. compositionally non-transparent (e.g. *black hole*). They were allowed the assignment of at most two labels in case of polysemy, for instance for *black staff* for the person vs. physical object senses of the noun or *yellow skin* for the race vs. literally painted interpretations of the AN. In this chapter, only the first label (most frequent interpretation, according to the judges) has been used. The κ coefficient of the annotation on the three categories (first interpretation only) was 0.87 (conf. int. 0.82-0.92, according to Fleiss *et al.* (1969)), observed agreement 0.96.[2] There were too few instances of idioms (17) for a quantitative analysis of the sort presented here, so these are collapsed with the subsective class in what follows.[3] The dataset as used here consists of 239 intersective and 130 subsective ANs.

[1] Available at `http://dl.dropbox.com/u/513347/resources/data-emnlp2012.zip`. See Bruni *et al.* (2012) for an analysis of the color term dataset from a multimodal perspective.

[2] Code for the computation of inter-annotator agreement by Stefan Evert, available at `http://www.collocations.de/temp/kappa_example.zip`.

[3] An alternative would have been to exclude idiomatic ANs from the analysis.

Intensional adjectives. The intensional dataset contains all ANs in the semantic space with a pre-selected list of 10 intensional adjectives, manually pruned by one of the authors of Boleda *et al.* (2012) to eliminate erroneous examples and to ensure that the adjective was being intensionally used. Examples of the ANs eliminated on these grounds include *past twelve* (cp. accepted *past president*), *former girl* (probably *former girl friend* or similar), *false rumor* (which is a real rumor that is false, vs. e.g. *false floor*, which is not a real floor), or *theoretical work* (which is real work related to a theory, vs. e.g. *theoretical speed*, which is a speed that should have been reached in theory). Other AN pairs were excluded on the grounds that the noun was excessively vague (e.g. *past one*) or because the AN formed a fixed expression (e.g. *former USSR*). The final dataset contained 1,200 ANs, distributed as follows: *former* (300 examples), *possible* (244), *future* (243), *potential* (183), *past* (87), *false* (44), *apparent* (39), *artificial* (36), *likely* (18), *theoretical* (6).[1] Table 3.1 contains examples of each type of AN we are considering.

Intersective	*Subsective*	*Intensional*
white towel	white wine	artificial leg
black sack	black athlete	former bassist
green coat	green politics	likely suspect
red disc	red ant	possible delay
blue square	blue state	theoretical limit

Table 3.1: Example ANs in the datasets.

[1] *Alleged*, one of the most prototypical intensional adjectives, is not considered here because it was not among the 700 most frequent adjectives in the space. We will consider it in future work.

3.4 Results

3.4.1 Corpus-extracted vectors

We began by exploring the empirically **corpus-extracted vectors** for the adjectives (A), nouns (N), and adjective-noun phrases (AN) in the datasets, as they are represented in the semantic space. Note that we are working with the AN vectors directly harvested from the corpora (that is, based on the co-occurrence of, say, the phrase *white towel* with each of the 10K words in the space dimensions), without doing any composition. AN vectors obtained by composition will be examined in the following section. Though corpus-extracted AN vectors should not be regarded as a gold standard in the sense of, for instance, Machine Learning approaches, because they are typically sparse[1] and thus the vectors of their component adjective and noun will be richer, they are still useful for exploration and as a comparison point for the composition operations (Baroni & Lenci, 2010; Guevara, 2010).

Figure 3.1 shows the distribution of the cosines between A, N, and AN vectors with intersective uses of color terms (IE, white box), subsective uses of color terms (S, lighter gray box), and intensional adjectives (I, darker gray box).

In general, the similarity of the A and N vectors is quite low (cosine < 0.2, left graph of Figure 1), and much lower than the similarities between both the AN and A vectors and the AN and N vectors. This is not surprising, given that adjectives and nouns describe rather different sorts of things.

We find significant differences between the three types of adjectives in the similarity between AN and A vectors (middle graph of Figure 3.1). The adjective and adjective-noun phrase vectors are nearer for intersective uses than for subsective uses of color terms, a pattern that parallels the difference in the distance between component A and N vectors. Since intersective uses correspond to the prototypical use of color terms (a *white dress* is the color white, while *white wine* is not), the greater similarity for the intersective cases is unsurprising – it suggests that in the case of subsective adjectival modifiers, the noun "pulls" the AN

[1]The frequency of the adjectives in the datasets range from 3.5K to 3.7M, with a median frequency of 109,114. The nouns range from 4.9K to 2.5M, with a median frequency of 148,459. While the frequency of the ANs range from 100 to 18.5K, with a median frequency of 239.

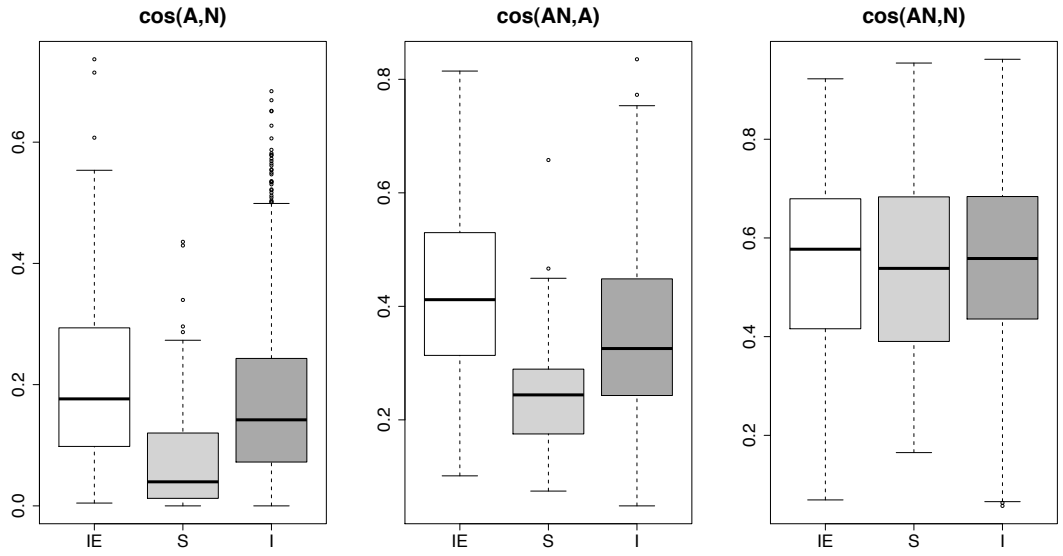

Figure 3.1: Cosine distance distribution in the different types of AN. We report the cosines between the component adjective and noun vectors (cos(A,N)), between the corpus-extracted AN and adjective vectors (cos(AN,A)), and between the corpus-extracted AN and noun vectors (cos(AN,N)). Each chart contains three boxplots with the distribution of the cosine scores (y-axis) for the intersective (IE), subsective (S) and intensional (I) types of ANs. The boxplots represent the value distribution of the cosine between two vectors. The horizontal lines in the rectangles represent the first quartile, median, and third quartile. Larger rectangles correspond to a more spread distribution, and their (a)symmetry mirrors the (a)symmetry of the distribution. The lines above and below the rectangle stretch to the minimum and maximum values, at most 1.5 times the length of the rectangle. Values outside this range (outliers) are represented as points.

further away from the adjective than happens with the cases of intersective modification. This is compatible with the intuition (manifest in the formal semantics tradition in the treatment of subsective adjectives as higher-order rather than first-order, intersective modifiers) that the adjective's effect on the AN in cases of subsective modification depends heavily on the interpretation of the noun with which the adjective combines, whereas that is less the case when the adjective is used intersectively.

As for intensional adjectives, the middle graph shows that their AN vectors are quite distant from the corresponding A vectors, in sharp contrast to what we find

with both intersective and subsective color terms. We hypothesize that the results for the intensional adjectives are due to the fact that they cannot plausibly be modeled as first order attributes (i.e. being *potential* or *apparent* is not a property in the same sense that being *white* or *yellow* is) and thus typically do not restrict the nominal description *per se*, but rather provide information about whether or when the nominal description applies. The result is that intensional adjectives should be even weaker than subsectively used adjectives, in comparison with the nouns with which they combine, in their ability to "pull" the AN vector in their direction. Note, incidentally, that an alternative explanation, namely that the effect mentioned could be due to the fact that most nouns in the intensional dataset are abstract and that adjectives modifying abstract nouns might tend to be further away from their nouns altogether, is ruled out by the comparison between the A and N vectors: the A-N cosines of the intensional and intersective ANs are similar. We thus conclude that here we see an effect of the *type of modification* involved.

An examination of the average distances among the nearest neighbors of the intensional and of the color adjectives in the distributional space supports our hypothesized account of their contrasting behaviors. We predict that the nearest neighbors are more dispersed for adjectives that cannot be modeled as first-order properties (i.e., intensional adjectives), than for those that can (here, the color terms). We find that the average cosine distance among the nearest ten neighbors of the intensional adjectives is 0.74 with a standard deviation of 0.13, which is significantly lower (t-test, $p<0.001$) than the average similarity among the nearest neighbors of the color adjectives, 0.96 with astandard deviation of 0.04.

Finally, with respect to the distances between the adjective-noun and head noun vectors (right graph of Figure 1), there is no significant difference for the intersective vs. subsective color terms. This can be explained by the fact that both kinds of modifiers are subsective, that is, the fact that a white dress is a dress and that white wine is wine.

In contrast, intensional ANs are closer to their component Ns than are color ANs (the difference is qualitatively quite small, but significant even for the intersective vs. intensional ANs according to a t-test, p-value = 0.015). This effect, the inverse of what we find with the AN-A vectors, can similarly be explained

by the fact that intensional adjectives do not restrict the descriptive content of the noun they modify, in contrast to both the intersective and subsective color ANs. Restriction of the nominal description may lead to significantly restricted distributions (e.g. the phrase *red button* may appear in distinctively different contexts than does *button*; similarly for *green politics* and *politics*), while we do not expect the contexts in which *former bassist* and *bassist* appear to diverge in a qualitatively different way because the basic nominal descriptions are identical, though further research will be necessary to confirm these explanations.

Finally, note that, contrary to predictions from some approaches in formal semantics, subsective color ANs and intensional ANs do not pattern together: subsective ANs are closer to their component As, and intensional ANs closer to their component Ns. This unexpected behavior underscores the fact highlighted in the previous paragraph: that the distributional properties of modified expressions are more sensitive to whether the modification restricts the nominal description than to whether the modifier is intersective in the strictest sense of term.

We now discuss the extent to which the different composition functions account for these patterns.

3.4.2 Model-generated vectors

Since intersective modification is the point of comparison for both subsective and intensional modification, we first discuss the model-generated vectors for the intersective vs. subsective uses of color terms, and then turn to intersective vs. intensional modification.

Intersective and subsective modification with color terms. To adequately model the differences between intersective and subsective modification observed in the previous section, a successful composition function should not only generate AN vectors that approximate the corpus-extracted AN vectors; it should also yield a significantly smaller distance between the adjective and AN vectors for intersectively used adjectives, whereas it should yield no significant difference for the distances between the noun and AN vectors.

Table 3.2 provides a summary of the results with the corpus-extracted data

(CORP) and the composition functions discussed in Section 2.2. The median rank of corpus-observed equivalent (ROE) is provided as a general measure of the quality of the composition function. It is computed by finding the cosine between the model-generated AN vectors and all rows in the semantic space and then determining the rank in which the corpus-extracted ANs are found.[1] The remaining columns report the differences in standardized (z-score) cosines between the vector built with each of the composition functions and the corpus-extracted AN, A, and N vectors. A positive value means that the cosines for intersective uses are higher, while a negative value means that the cosines for subsective uses are higher. The first row (CORP) contains a numerical summary of the tendencies for corpus-extracted ANs explained in the previous section. This is the behavior that we expect to model.

model	*ROE*	Δ*:AN*		Δ*:A*		Δ*:N*	
CORP	-	-		1.13	*	.08	
ADD	134	.75	*	.90	*	.90	*
W.ADD	161	.53	*	.91	*	.89	*
MULT	1,106	.66	*	1.05	*	.62	*
DL	800	.19		.92	*	-.78	*
F.ADD	195	.50	*	.91	*	.09	
LFM	542	.39		1.04	*	.51	*

Table 3.2: Intersective vs. subsective uses of color terms. The first column reports the rank of the corpus-observed equivalent (ROE), the rest report the differences (Δ) betwen the intersective and subsective uses of color terms when comparing the model-generated AN with the corpus-extracted vectors for: AN, adjective (A), noun (N). See text for details. Significances according to a t-test: * for $p< 0.001$.

One composition function comes close to modeling the corpus-observed behavior: F.ADD. In this case, we find that the function yields higher similarities for AN-A for the intersective than for the subsective uses of color terms, and a very slight difference for the distance to the head noun. The MULT and LFM models approximate the corpus-observed behavior best with respect to the distance from

[1]The ROE is provided as a general guide; however, recall that the ROE was taken into account to tune the λ parameter in the dilation model, and that the ANs of the color dataset were included when training the matrices for the LFM model.

to the component adjective. Although they are unable to capture the observed, and expected, effect in the distance from the head noun, there is an asymmetry that we would expect between these measure in both composition models. The ADD and W.ADD functions perform very well in terms of ROE (median 134). This suggests that, for adjectival modification, providing a vector that is in the middle of the two component vectors (which is what normalized addition does), or slightly skewed towards the head in the case of W.ADD, is a reasonable approximation of the corpus-extracted vectors. However, precisely because the resulting vector is in the middle of the two component vectors, these functions cannot account for the asymmetries in the distances found in the corpus-observed data. One might expect that a non-normalized version of ADD could not account for these effects because the adjective vector, being much longer (as color terms are very frequent), would totally dominate the AN, resulting in no difference across uses when comparing to the adjective or to the noun.

The DL model shows a strange pattern, as it yields a strongly significant negative difference in the AN-N distance. This is likely a result of the intuitive choice of the adjective vector as $\vec{u}$ and the noun vector as $\vec{v}$ in composition (see Equation 2.8). A post-hoc analysis showed that if we were to reverse the assignment (i.e., the adjective vector as $\vec{v}$ and the noun vector as $\vec{u}$), we find that the results are quantitatively identical, however reversed, i.e., Δ:A= $-.78$ and Δ:N= .92. The MULT model is by far the worst function in terms of ROE, which can be attributed to the sparsity of the model-generated vectors after point-wise multiplication of NFM-reduced component vectors.

All composition functions except for DL and LFM find intersective uses easier to model. This is shown in the positive values in column Δ:AN, which mean that the similarity between corpus-extracted and model-generated AN vectors is greater for intersective than for subsective ANs. This is consistent with expectations. The subsective uses are specific to the nouns with which the color terms combine, and the exact interpretation of the adjective varies across those nouns. In contrast, the interpretation associated with intersective use is consistent across a larger variety of nouns, and in that sense should be predominantly reflected in the adjective's vector. Although this follows our expectations, it is not necessarily a positive feature of these composition functions. The exception in this respect are the DL

		MULT	F.ADD	LFM
green stone	IE	green background	old wall	green marble
		white ground	white stone	red roof
		blue wave	red tower	white stone
		white cross	red stone	yellow stone
		blue ground	stone	green tile
red ball	IE	low cross	other ball	white triangle
		free kick	red ball	blue square
		free header	yellow ball	black colour
		low shot	blue ball	black cross
		own net	red	blue ring
blue shark	S	blue fish	common dolphin	common dolphin
		shark	white shark	whale
		small shark	great shark	green frog
		blue shark	blue shark	blue shark
		dolphin	white whale	great shark
green future	S	environmental asset	strong future	green transport
		local biodiversity	future	green policy
		conservation	long-term future	sustainable alternative
		green infrastructure	positive news	green issue
		biodiversity	long future	green future

Table 3.3: Examples of nearest neighbors for color terms according to the three composition models in intersective (IE) vs. subsective (S) color terms: MULT, F.ADD and LFM.

and LFM functions. In the case of LFM, the weights for each adjective matrix are estimated in relation to the noun vectors with which the adjective combines, on the one hand, and the related corpus-extracted AN vectors, on the other; thus, the basic lexical representation of the adjective is inherently reflective of the distributions of the ANs in which it appears in a way that is not the case for the adjective representations used in the other composition models. And indeed, DL and LFM are the only functions that show no difference in difficulty (distance) between the model-generated and corpus-extracted AN vectors for intersective vs. subsective ANs.

The three composition functions that "best" account for the corpus-extracted patterns in color terms are F.ADD, MULT and LFM. However, an examination of the nearest neighbors of the model-generated ANs suggest that LFM captures

the semantics of adjective composition in this case to a larger extent than both F.ADD and MULT. Consider the difference in nearest neighbors of intersective and subsective color terms in Table 3.3.

Intensional modification. Table 3.4 contains the results of the composition functions comparing the behavior of intersective color ANs and intensional ANs. The tendencies in the ROE are as in Table 3.2, so we will not comment on them further (note the very poor performance of MULT, though). As noted above, we expect more difficulty in modeling intensional modification vs. other kinds of modification, however this is verified in the results for only the ADD and MULT models (cf. the positive values in second column), and only slightly for W.ADD. While we find that the LFM model is able to approximate corpus-observed vectors for intensional modification easier than for intersective uses of color terms. This points to a qualitative difference between subsective and intensional adjectives that could be evidence for a first-order analysis of subsective color terms. (See Boleda *et al.* (2013) for an extended study on detecting intensional modification using compositional distributional semantics.)

model	*ROE*	Δ*:AN*		Δ*:A*		Δ*:N*	
CORP	-	-		.51	*	-.03	
ADD	196	.28	*	.26	*	-.26	*
W.ADD	202	.18		.27	*	.26	*
MULT	1,287	.47	*	.34	*	.13	
DL	598	.01		.26	*	-.25	*
F.ADD	337	-.01		.30	*	.14	
LFM	530	-.56	*	.64	*	-.14	

Table 3.4: Intersective vs. intensional ANs. Information as in Table 3.2.

A good composition function should provide a large positive difference when comparing the AN to the A, and a small negative difference (because the effect is not significant in the corpus-observed data) when comparing the AN to the N. The functions that best match the corpus-observed data are again LFM, F.ADD and MULT. ADD and DL show the predicted pattern, but to a much lesser degree (cf. smaller differences in column Δ:A).

Again, LFM seems to be capturing relevant semantic aspects of composition

	MULT	F.ADD	LFM
artificial leg	total replacement	leg	artificial joint
	artificial joint	weak leg	active patient
	orthopaedic	human arm	artificial limb
	active patient	hard ground	artificial heart
	other joint	entire body	advanced procedure
former job	assistant	permanent job	former worker
	senior	new job	strong rumor
	manager	high job	former manager
	coordinator	previous job	current boss
	principal	high pay	former colleague
possible damage	omission	physical damage	possible consequence
	misconduct	additional damage	potential consequence
	failure	potential consequence	potential loss
	formal action	possible consequence	serious consequence
	negligent	serious damage	potential hazard

Table 3.5: Examples of nearest neighbors for intensional terms according to the three composition models: MULT, F.ADD and LFM.

with intensional adjectives, as seen in Table 3.5..

3.5 Discussion

The present research provides evidence for treating adjectives as matrices or functions, rather than vectors, although simple operations on vectors such as ADD and W.ADD (for their excellent approximation to observed vectors) still account for some aspects of adjectival modification. The MULT model, in contrast, struggles to approximate adjectival modification (as seen in the poor ROE scores) likely due to the sparse, or even zero, vectors that result after point-wise multiplication of NMF-reduced component vectors. This is a serious drawback of the MULT model.

Our results also show that LFM and F.ADD in general perform better than other models. We consider F.ADD very attractive in principle because it generalizes across adjectives and is thus more parsimonious. Part of the flaws of F.ADD are due to limitations of our implementation, as we trained the matrices on only

2.5K ANs, while our semantic space contains more than 170K ANs. However, the linguistic literature and the present results suggest that it might be useful to try a compromise between LFM and F.ADD, training one matrix for each subclass of adjectives under analysis.

Beyond the new data it offers regarding the comparative ability of the different composition functions to account for different kinds of adjectival modification, the study presented here underscores the complexity of modification as a semantic phenomenon. The role of adjectival modifiers as restrictors of descriptive content is reflected differently in distributional data than is their role in providing information about whether or when a description applies to some individual. Formal semantic models, thanks to their abstractness, are able to handle these two roles with little difficulty, but also with limited insight. Distributional models, in contrast, offer the promise of greater insight into each of these roles, but face serious challenges in handling both of them in a unified manner.

Chapter 4

Capturing semantic deviance

4.1 Introduction

A prominent approach for representing the meaning of a word in Natural Language Processing (NLP) is to treat it as a numerical vector that codes the pattern of co-occurrence of that word with other expressions in a large corpus of language (Sahlgren, 2006; Turney & Pantel, 2010): the meaning of the word *painting*, for instance, could be characterized in terms of its proximity with *artist*, *museum*, *colorful*, *abstract*, etc. This approach to semantics (sometimes called *distributional semantics*) naturally captures collocations, scales well to large lexicons and does not require words to be manually disambiguated (Schütze, 1997). Until recently, however, this method had been almost exclusively limited to the level of content words (nouns, adjectives, verbs), and had not directly addressed the problem of *compositionality* (Frege, 1892; Partee, 2004), the crucial property of natural language which allows speakers to derive the meaning of a complex linguistic constituent from the meaning of its immediate syntactic subconstituents. Together with a generative syntactic component (Chomsky, 1957), this principle is responsible for the productivity of natural language, which allows speakers to produce and understand sentences they have never encountered before.

To address this serious shortcoming, several recent proposals have strived to extend distributional semantics with a component that also generates vectors for complex linguistic constituents, using compositional operations in the vec-

tor space (Baroni & Zamparelli, 2010; Blacoe & Lapata, 2012; Grefenstette & Sadrzadeh, 2011a; Guevara, 2010; Mitchell & Lapata, 2010; Socher *et al.*, 2012). All these approaches manage to construct distributional semantics representations for novel phrases, starting from the corpus-derived vectors for their lexical constituents. Since their output is naturally graded, these methods also promise to address the fact that compositionality is a matter of degree (Nunberg *et al.*, 1994), ranging from fully compositional cases, as in those attributive adjective-noun phrases whose meaning is the intersection of the meaning of the noun and adjective (e.g. *rented car*, *wooden spoon*), to syntactically fixed expressions such as *take advantage*, *cut a deal*, where the meaning of some of their subparts can still be recognized in the final meaning, to idioms and multi-word expressions (*kick the bucket*, *red herring*, *by and large*), whose meaning cannot be distributed at all across their constituents. Despite these latter cases, language is still largely compositional, providing an open space for speakers to create novel but understandable complex linguistic expressions.

Yet, linguistic creativity has its limits: as native speakers we have the clear intuition that not all of the infinitely many possible syntactically well-formed strings are equally semantically acceptable. Chomsky's classic example in (1) was devised precisely to show that syntax and semantics can diverge.

(1) *Colorless green ideas sleep furiously*

Our knowledge of compositionality tells us that here the lexical semantics of the words *colorless, green* and *ideas* did not combine properly. The result is a semantically deviant phrase which cannot be used in 'normal' contexts (e.g. non-metalinguistic ones—see below for some qualifications), and therefore will not be found in corpora, not even very large ones, since corpora largely document actual, normal language use.

Of course, the fact that a complex expression is not found in a corpus can be due to a variety of reasons, which can be quite difficult to tell apart: pure chance, the fact that the expression, though understandable, is ungrammatical, that it uses a rare or very complex structure, describes false facts or nonexistent entities, or, finally, the fact that it nonsensical. One criticism aimed at corpus

linguistics from the generative linguistic community was precisely that (crude) statistical approaches could not distinguish between these various possibilities (cf. Chomsky's famous remark that "I live in Dayton, Ohio" is not less grammatical, nor indeed, less meaningful, that "I live in New York", despite being far less frequent).

In this study we show that it is possible to use compositional distributional methods to distinguish the unattestedness due to nonsensicality from all the other cases, in the domain of simple noun phrases. Specifically, we show that distributional measures such as vector length, neighbor density and cosine distance can reliably predict the extent to which a novel adjective-noun combination—one never found in a corpus and never seen by the system in the training phase—makes sense. Moreover, we show that these distributional measures improve over shallow, word-based measures like word length or word frequency. Finally, we show that this result holds across a variety of compositional methods proposed in the literature, though some are of course better then others in various subtasks.

To put the problem in context, consider the difference between two adjective-noun phrases in (2) which are not attested in a large corpus of English.

(2) a. *grooved tangerine*
b. *residential steak*

Although it may be the case that you have never considered that a tangerine could have grooves, such an object is easy to imagine and it can be understood in out-of-the-blue contexts. On the other hand, *residential steak* describes an object that is quite hard to imagine. In what sense can a steak be *residential*? Perhaps in none, perhaps in too many: in the context of a man who always and only eats steak when he is in his residence, *his usual residential steak* makes sense. Notice, however, that now the adjective is used only as a proxy for a larger description (*eaten when in residence*). Out of the blue, *residential steak* is semantically very odd, *grooved tangerine* is not (though it might be factually strange, whence its absence). In truly semantically deviant cases, different speakers would probably not even agree on how to paraphrase the expressions, if given in isolation.

Beyond these intuitions, we still do not have a precise linguistic account of

what it means for a linguistic expression to be "nonsensical", nor a clear relation between this notion and that of being unattested in a corpus: semantic deviance remains a difficult and understudied phenomenon. In formal denotation-based semantics, for instance, a 'meaningless sentence' could perhaps be characterized as one which is false in any imaginable situation (say, in any epistemically accessible possible world). However, this approach would still be unable to determine the degree or even the motivation for the deviance, and could not predict when a novel string will be nonsensical. Moreover, there are many necessarily false expressions such as (3) which do not feel nonsensical, but simply false.

(3) *17 is not a prime*

Thus, the task of distinguishing between unattested but **acceptable** and unattested but **semantically deviant** linguistic expressions is not only a way to address a criticism about the limits of corpus linguistics, but also an interesting linguistic task, whose solution could have an impact on the theoretical and computational linguistic community as a whole, and shape our future treatments of semantic deviance.

In this study, we apply methodologies drawn from psycholinguistics, formal semantics and distributional semantics to model our intuitions about the semantic acceptability of novel linguistic expressions. Our specific goal is to automatically detect semantic deviance in attributive adjective-noun (AN) expressions using a small number of simple, unsupervised cues. The choice of ANs as our testbed is motivated by two facts: first of all, ANs are common, small constituents containing no functional material; and secondly, ANs have already been studied extensively in compositional distributional semantics (Baroni & Zamparelli, 2010; Boleda *et al.*, 2012; Guevara, 2010; Mitchell & Lapata, 2010; Vecchi *et al.*, 2011, 2013a,b). In order to carry out a large–scale study on semantic deviance in AN phrases, we first construct a large set of ANs which are not found in very large corpus of English and which are judged either semantically acceptable or deviant in a crowdsourcing experiment (this dataset can be downloaded from `www.evavecchi.com`). We then estimate semantic representations of these unattested ANs by applying some composition functions popular in com-

positional distributional semantics. Finally, we asses the effect of a number of variables in the ability to model the intuitions of semantic acceptability for novel phrases, on the basis of the acceptability judgments we collected. Since, to our knowledge, this is the first attempt to automatically model semantic acceptability computationally, we did not know a priori which features could be best for the task. Therefore, we used a variety or metrics, some taken from the cognitive and psycholinguistic literature on lexical processing and in particular compound processing, others designed by us and based on the distributional representation we are using. Evaluating the effectiveness of these measures, we show that distributional semantics techniques go beyond semantically shallow wordform-based measures previously tested in psycholinguistic studies.

The unsupervised method we introduce for measuring and estimating the semantic deviance of phrases can be applied to a number of NLP tasks, such as metaphor analysis, the collection of better estimates for language modeling and a measure of plausibility in machine translation tasks.

Outline This section is structured as follows. Section 4.4.1 describes the design of the experiments discussed in this study, including the datasets used and the parameter-tuning phase for the composition methods. The measures we tested are described in Section 4.4.2, and our approach to data analysis is laid out in Section 4.4.2. We present the results of our experiments in Section 4.4.3 in three parts: (i) the ability of word-based measures to model the acceptability of novel AN phrases (our baseline); (ii) the measures extracted from estimated distributional representations of ANs which improve the ability to model semantic acceptability; and (iii) a detailed analysis of the performances of each composition function. Finally, Section 4.4.4 contains a discussion of the conclusions drawn from this study, as well as a number of issues that we would like to address in future research.

4.2 Related work

The question of when a complex linguistic expression is semantically deviant has been addressed since the 1950's in various areas of linguistics. In computa-

tional linguistics, the possibility of detecting semantic deviance has been seen as a prerequisite to access metaphorical/non-literal semantic interpretations (Fass & Wilks, 1983; Zhou *et al.*, 2007). In psycholinguistics, it has been part of a wide debate on the point at which context can make us perceive a 'literal' vs. a 'figurative' meaning (Giora, 2002). In theoretical generative linguistics, the issue is part of an ongoing discussion on the boundaries between syntax and semantics.

(4) *Colorless green ideas sleep furiously*

For instance, despite Chomsky's (1957) claim that (4) is syntactically flawless, the unacceptability of this case could also be regarded as a violation of very fine-grained syntactic *selectional restrictions* on the arguments of verbs or modifiers, on the model of **much computer* (arguably a failure of *much* to combine with a noun +COUNT). However, it has been observed that the features at issue (say, +SOLID, required by *carve*, (5-a)) cannot be on a par with well-established syntactic features such as NUMBER, since the former can act at arbitrary distance: (5-b) is as deviant as (5-a) (Delfitto & Zamparelli, 2009). A semantic account seems preferable.

(5) a. ??Sabrina carved a gas$_{-solid}$
b. ??Sabrina carved carved something which a US lab proved to have identical physical properties as an rare element found in gaseous$_{-solid}$ state only.

The spirit of the selectional approach persists in Asher (2011), who proposes a detailed system of *semantic* types, far beyond individuals (e) and truth values (t). Unacceptable phrases like *residential steak* can now be excluded by type incompatibility. Reducing Asher's proposal to a "cartoon" version for illustration purposes, we might have types such as $< e$-that-are-dwellings $>$ and $< e$-that-you-eat-cooked $>$. Defining *steak* and *residential* as in (6), *residential* would not accept *steak* as a possible input.

(6) a. **steak**: $< e$-thay-you-eat-cooked, $t >$
b. **residential**: $<< e$-that-are-dwellings, $t >, < e$-that-are-dwellings, $t >>$

Note that Asher (2011) also incorporates a theory of *type coercion*, in which a particular interpretation of a word or phrase is coerced from the context, designed to account for the shift in meaning seen in, e.g., (7) (*lunch* as food or as an event).

(7) Lunch was delicious but took forever.

A practical problem with this approach is that a full handmade specification of the types that determine semantic compatibility is a very expensive and time-consuming enterprise, and it should be done consistently across the whole content lexicon. Moreover, it is unclear how to model the intuition that *naval fraction*, *musical North* or *institutional acid* sound odd, in the absence of very particular contexts, while (7) sounds quite natural: whatever the nature of coercion, we do not want it to run so smoothly that any combination of A and N (or V and its arguments) becomes meaningful and completely acceptable. Evidently, a cognitively plausible model should account for gradient acceptability judgments. Consider for instance the expressions in (8), all of which are unattested in a large corpus, and which have received descending acceptability ratings in the crowdsourcing experiment described in Section 4.4.1.

(8) a. creative apprentice
b. ?southern ghost
c. *careful dark

It is clear that while (8-a) and (8-c) represent the binary extremes of acceptability, (8-b) is neither here nor there; it is clearly an odd expression, yet we would not want to consider it as deviant as (8-c).

It is important to note that in this research we talk about "semantically deviant" expressions, but we do not exclude the possibility that such expressions are interpreted as metaphors, or via a 'proxy' association like the 'eaten-when-in-residence' steak. In fact, distributional measures are desirable models to account for this, since they naturally lead to a gradient notion of semantic anomaly.

Further, we would like to emphasize the goal of detecting semantic deviance, which is not entirely synonymous to the notion of plausibility. Previous work has aimed at predicting human plausibility judgments of adjective-noun combi-

nations or verb-relation-argument triples using various computational, corpus-driven models (Lapata *et al.*, 1999, 2001; Padó *et al.*, 2007). Unlike the studies presented in this chapter, the authors used, above all, word frequencies and/or bigram co-occurrence frequencies to determine how plausibility is driven by a strong or weak lexicalized and collocational nature of the phrases. In these experiments, however, we focus on distinguishing between phrases that are unattested in a large corpus due to poor coverage or rareness from those that are unattested because the combination would produce a semantically unacceptable adjective-noun phrase. In addition, in relation to the studies mentioned above, this work provides strong support for compositional distributional semantics, as it is able to generalize and capture the complex semantic intuition of natural language speakers for bigrams, even without being able to rely on co-occurrence relations between the constituents.

4.2.1 Semantic processing of word combinations

Psycholinguists have traditionally studied the processing of word combinations by focusing on compound words with nominal constituents. Their studies have shown that constituent representations are accessed when a compound is read, and that many variables influence this process. Semantic transparency, for example, has been shown to affect the amount of cross-activation between the constituent representations and the compound representation. Sandra (1990) found that the recognition of transparent compounds was aided by prior exposure to a semantically related word, but opaque compounds were not. Likewise, Zwitserlood (1994) examined whether exposure to compound words affects the ease of processing semantic associates of either the first or second constituents. Semantically related words were faster to process following transparent and partially opaque compounds, but not following fully opaque compounds.

Moreover, most studies have demonstrated that word frequency is one of the most robust factors driving processing speed: Words with a high frequency of occurrence are processed faster and more accurately than words with a low frequency of occurrence (Gardner *et al.*, 1987; Gordon, 1983; Hasher & Zacks, 1984). In addition, the frequencies of occurrence of the constituents of complex words

and compounds have been shown to have an effect on lexical processing (Andrews *et al.*, 2004; Juhasz *et al.*, 2003; Pollatsek *et al.*, 2000). Researchers have also explored the effect of *family size*, i.e., the number of distinct phrase types of which the word can be part (for instance, the number of distinct nouns a given adjective can be seen to modify in a corpus). De Jong *et al.* (2002) showed that constituent family-size facilitates the lexical processing of compounds in both Dutch and English: the higher the family size of a constituent, the easier it is to process the compound. These effects are not necessarily independent: Kuperman *et al.* (2009) observed simultaneous effects of compound frequency, left constituent frequency, and family size early (i.e., before the whole compound has been scanned) and also observed the effects of right constituent frequency and family size that emerged subsequently to the compound frequency effect. In addition to these variables, a study carried out in Bertram & Hyönä (2003) provides evidence that string length modulates the access to constituents during the lexical processing of compound words. Specifically, the authors found that in the case of long compounds, it is more likely that the constituents are used for processing (possibly through a compositional procedure), while in the case of short compounds there is probably a direct access to the lexical representations of the compound.

However, all these studies have investigated the processing of familiar word combinations, while the problem of how novel word combinations are elaborated has been mostly overlooked. The few studies on the the topic have focused on the role of relational information. For example, research with novel phrases indicates that the time required to interpret a modifier-noun phrase is affected by the availability of the relation used to link the two constituents, suggesting that the processing of novel phrases is affected by the availability of relations associated with the modifier (Devereux & Costello, 2006; Gagné, 2002; Gagné & Shoben, 2002).

In addition, most of these works on novel phrases have focused on novel noun-noun compounds, not adjective-noun combinations. To our knowledge only two studies have focused on this construction: Mullaly *et al.* (2010) explored how alternative senses of ambiguous adjectives impacted interpretation and sense/nonsense judgments in word combinations, while Schmidt *et al.* (2006) proposes a mathematical model to distinguish sensible yet unlikely adjective-

noun phrases from nonsensical phrases. The latter model incorporates the 'M constraint' (the assumption that categories of objects are organized in a strict hierarchy, and that predicates must span subtrees of the hierarchy, Sommers, 1971) in order to acquire predicability trees given observations of what is true in the world, and nothing else. This study showed that the distinction between sense and nonsense is a statistically learnable one; however, the model remains purely theoretical since it has yet to be applied to real-world datasets. So, while most studies on this issue have provided evidence on how novel compounds are processed and how variables such as relational properties and family size play an important role in lexical processing, the attempts to model or predict the choice of acceptability of novel phrases are for the most part untested, providing little information as to which variables influence acceptability. Although our goal to model semantic acceptability differs from that of investigating factors that affect lexical processing, the studies described above have provided interesting insight into a set of word-based measures that might also have an effect in acceptability judgments. Therefore, in our study we will also consider the impact of variables which have been shown to affect processing of (novel) phrases.

4.3 Experiment 1: Pilot Study

4.3.1 Simple indices of semantic deviance

We consider here a few simple, unsupervised measures to help us distinguish the representation that a distributional composition model generates for a semantically anomalous AN from the one it generates for a semantically acceptable AN. In both cases, we assume that the AN is not already part of the model semantic space, just like you can distinguish between *parliamentary tomato* (odd) and *marble iPad* (OK), although you probably never heard either expression.

We hypothesize that, since the values in the dimensions of a semantic space are a distributional proxy to the meaning of an expression, a meaningless expression should in general have low values across the semantic space dimensions. For example, a *parliamentary tomato*, no longer being a vegetable but being an unlikely parliamentary event, might have low values on both dimensions characterizing vegetables and dimensions characterizing events. Thus, our first simple measure of semantic anomaly is the vector length (***vlength***) of the model-generated AN. We hypothesize that anomalous AN vectors are shorter than acceptable ANs.

Second, if deviant composition destroys or randomizes the meaning of a noun, as a side effect we might expect the resulting AN to be more distant, in the semantic space, from the component noun. Although even a *marble iPad* might have lost some essential properties of iPads (it could for example be an iPad statue you cannot use as a tablet), to the extent that we can make sense of it, it must retain at least some characteristics of iPads (at the very least, it will be shaped like an iPad). On the other hand, we cannot imagine what a *parliamentary tomato* should be, and thus cannot attribute even a subset of the regular tomato properties to it. We thus hypothesize that model-generated vectors of deviant ANs will form a wider angle (equivalently, will have a lower ***cosine***) with the corresponding N vectors than acceptable ANs.

Next, if an AN makes no sense, its model-generated vector should not have many neighbours in the semantic space, since our semantic space is populated by nouns, adjectives and ANs that are commonly encountered in the corpus, and should thus be meaningful. We expect deviant ANs to be "semantically

isolated", a notion that we operationalize in terms of a (neighborhood) ***density*** measure, namely the average cosine with the (top 10) nearest neighbours. We hypothesize that model-generated vectors of deviant ANs will have lower density than model-generated acceptable ANs.

Finally, since length, as already observed Vecchi *et al.* (2011), is strongly affected by independent factors such as input vector normalization and the estimation procedure, we test ***entropy*** as a measure of vector quality, introduced in Lazaridou *et al.* (2013). The intuition is that meaningless vectors, whose dimensions contain mostly noise, should have high entropy.

4.3.2 Methodology

Evaluation materials Our goal is to study what happens when compositional methods are used to construct a distributional representation for ANs that are semantically deviant, compared to the AN representations they generate for ANs they have not encountered before, but that are semantically acceptable.

In order to assemble these lists, we started from the set of 3.5M unattested ANs described in Section 2.1.2 above, focusing on 30 randomly chosen adjectives. For each of these, we randomly picked 100 ANs for manual inspection (3K ANs in total). Two authors went through this list, marking those ANs that they found semantically highly anomalous, no matter how much effort one would put in constructing metaphorical or context-dependent interpretations, as well as those they found completely acceptable (so, rating was on a 3-way scale: deviant, intermediate, acceptable). The rating exercise resulted in rather low agreement (Cohen's κ=0.32), but we reasoned that those relatively few cases (456 over 3K) where both judges agreed the AN was odd should indeed be odd, and similarly for the even rarer cases in which they agreed an AN was completely acceptable (334 over 3K). We thus used the agreed deviant and acceptable ANs as test data.

Of 30 adjectives, 5 were discarded for either technical reasons or for having less than 5 agreed deviant or acceptable ANs. This left us with a **deviant AN test set** comprising of 413 ANs, on average 16 for each of the 25 remaining adjectives. Some examples of ANs in this set are: *academic bladder*, *blind pronunciation*, *parliamentary potato* and *sharp glue*. The **acceptable** (but unattested) **AN**

test set contains 280 ANs, on average 11 for each of the 25 studied adjectives.[1] Examples of ANs in this set include: *vulnerable gunman*, *huge joystick*, *academic crusade* and *blind cook*.

Experimental procedure Using each composition method, we generate composite vectors for all the ANs in the two (acceptable and deviant) evaluation sets (see above). We then compute the measures that might cue semantic deviance discussed in Section 4.3.1 above, and compare their values between the two AN sets. In order to smooth out adjective-specific effects, we z-normalize the values of each measure across all the ANs sharing an adjective before computing global statistics (i.e., the values for all ANs sharing an adjective from the two sets are transformed by subtracting their mean and dividing by their variance). We then compare the two sets, for each composition method and deviance cue, by means of two-tailed Welch's t tests. We report the estimated t score, that is, the standardized difference between the mean acceptable and deviant AN values, with the corresponding significance level. For all our cues, we predict t to be significantly larger than 0: Acceptable AN vectors should be *longer* than deviant ones, they should be *nearer* – that is, have a higher cosine with – the component N vectors and their neighbourhood should be *denser* – that is, the average cosines with their top neighbours should be higher than the ones of deviant ANs with their top neighbors.

4.3.3 Results

The results of our experiments are summarized in Table 4.1. We see that all models – except DL provide significant results in the expected direction for the ***vlength*** and ***cosine*** tests. We are able to capture the distinction between acceptable and deviant ANs in terms of ***density*** only with the ADD and F.ADD models, again in the expected direction. While the results of the ***entropy*** test are rather significant only for MULT and LFM, however in opposite directions.

First, we find that all composition models are able to capture the difference between the acceptable and deviant phrases with respect to the ***vlength*** and

[1]The evaluation sets can be downloaded from `http://www.vecchi.com/eva/resources.html`.

	VLENGTH		COSINE		DENSITY		ENTROPY	
model	*t*	*sig.*	*t*	*sig.*	*t*	*sig.*	*t*	*sig.*
ADD	8.92	*	8.92	*	7.73	*	-2.50	
W.ADD	8.92	*	8.87	*	1.68		-2.31	
MULT	8.03	*	7.59	*	1.49		7.75	*
DL	7.37	*	-7.49	*	-2.30		1.53	
F.ADD	9.29	*	4.04	*	7.87	*	0.72	
LFM	9.31	*	10.00	*	-0.80		-8.86	*

Table 4.1: *t* scores for difference between acceptable and deviant ANs with respect to 4 cues of deviance: ***vlength*** of the AN vector, ***cosine*** of the AN vector with the component noun vector, ***density***, measured as the average cosine of an AN vector with its nearest 10 neighbours in semantic space, and ***entropy***. For all significant results, $p < 0.01$.

cosine measures. In Baroni & Zamparelli (2010), the LFM model performed far better than ADD and MULT in approximating the correct vectors for unseen ANs. On this (in a sense, more metalinguistic) task, again we see that LFM outperforms all models tested with respect to these measures (as seen in the high *t* scores in Table 4.1).

The high scores in the ***vlength*** analyses across all models, especially the component-wise models, are an indication that semantically acceptable ANs tend to be composed of *similar* adjectives and nouns, i.e., those which occur in similar contexts and we can assume are likely to belong to the same domain, which sounds plausible. The high results for the ***cosine*** measure is encouraging, albit not entirely surprising. The behavior of the DL model for this measure is likely a reflection of the high emphasis placed on the noun, which is a characteristic of the implementation of this composition function (see Eq. 2.8).

The behavior of the ***entropy*** measure is quite puzzling, since it provides contradictory results in the two models for which there is a significant difference between acceptable and deviant ANs: MULT and LFM. In the case of the MULT model, higher entropy scores correlate with acceptable ANs, while in the case of LFM higher entropy scores result in deviant ANs. Table 4.2 provides a better look at the results of for these two models, listing the highest/lowest entropy scores for each model, specifying deviant ANs with an (*).

The examples provided in Table 4.2 demonstrate that indeed there is a contra-

model			ENTROPY
MULT	HIGH	huge glimpse	2.45
		spectacular cameraman	2.39
		sharp guess	2.37
		industrial groundwork	2.24
		religious parliamentarian	2.20
	LOW	*academic bowel	0.00
		*institutional deer	0.00
		*printed avenue	0.03
		*optional chemist	0.04
		*reasonable pen	0.05
LFM	HIGH	*exact crab	5.54
		*huge nanotechnology	5.48
		*sharp waterway	5.48
		*blind clay	5.47
		*reasonable pen	5.45
	LOW	academic communications	4.89
		naval damage	4.99
		coastal mosquito	5.00
		residential clubhouse	5.02
		printed icon	5.05

Table 4.2: Examples of the highest/lowest scores of the ***entropy*** measure for the two significant models: MULT and LFM. Deviant ANs are marked with an (*).

dictory effect in both models. It seems the range of entropy is much greater for the MULT model, while AN vectors generated with LFM are in general highly entropic (although the difference between acceptable and deviant ANs is significant).

To gain a better understanding of the neighborhood ***density*** test we performed a detailed analysis of the nearest neighbors of the AN vectors generated by all composition models. For each of the ANs, we looked at the top 10 semantic-space neighbors generated by each of the three models, focusing on two aspects: whether the neighbor was a single A or N, rather than AN, and whether the neighbor contained the same A or N as the AN is was the neighbor of (as in *blind regatta / blind athlete* or *biological derivative / partial derivative*). The results are summarized in Table 4.3.

In terms of the properties we measured, neighbor distributions are quite simi-

model	*status*	A *only*	N *only*	A_1=A_2	N_1=N_2
ADD	accept	20.9	31.1	14.7	15.1
	deviant	18.0	36.3	14.8	14.0
W.ADD	accept	12.7	35.2	4.1	18.2
	deviant	12.1	44.8	3.7	15.4
MULT	accept	18.8	36.5	0.8	0.4
	deviant	15.2	39.9	0.4	0.2
DL	accept	10.4	46.1	0.0	18.1
	deviant	11.1	54.5	0.0	15.2
F.ADD	accept	1.1	5.9	4.1	11.9
	deviant	1.7	8.3	6.3	9.4
LFM	accept	6.8	2.7	19.9	0.1
	deviant	7.4	1.8	21.1	0.0

Table 4.3: Percentage distributions of various properties of the top 10 neighbours of ANs in the acceptable (2800) and deviant (4130) sets for each model. The last two columns express whether the neighbor contains the same Adjective or Noun as the target AN.

lar across acceptable and deviant ANs. One interesting finding is that the system is quite 'noun-driven': particularly for the ADD and W.ADD models (where we can imagine that some As with low dimensional values do not shift much the noun position in the multidimensional space). On the other hand, the LFM is the model that is most driven by the adjective. The DL model, by construction, will favor the meaning of the noun, which is seen clearly in these results, while the MULT model seems to be drawn most to component elements in the space. With respect to the last two columns, it is interesting to observe that matching As are frequent for deviant ANs even in LFM, a model which has never seen A-vectors during training. Further qualitative evaluations show that in many deviant AN cases the similarity is between the A in the target AN and the N of the neighbor (e.g. *academic bladder / honorary lectureship*), while the opposite effect seems to be much harder to find.

4.3.4 Discussion

The main aim of this study was to propose a new challenge to the computational distributional semantics community, namely that of characterizing what happens,

distributionally, when composition leads to semantically anomalous composite expressions. The hope is, on the one hand, to bring further support to the distributional approach by showing that it can be both productive and constrained; and on the other, to provide a more general characterization of the somewhat elusive notion of semantic deviance – a notion that the field of formal semantics acknowledges but might lack the right tools to model.

Our results are very preliminary, but also very encouraging, suggesting that simple unsupervised cues can significantly tell unattested but acceptable ANs apart from impossible, or at least deviant, ones. Although, somewhat disappointingly, the model that has been shown in a previous study (Baroni & Zamparelli, 2010) to be the best at capturing the semantics of well-formed ANs turns out to be worse than simple addition and multiplication.

Future avenues of research must include, first of all, an exploration on the effect on each model when tested in the non-reduced space where computationally possible, or using different dimensionality reduction methods. A preliminary study demonstrates an enhanced performance of the MULT method in the full space.

Second, we hope to provide a larger benchmark of acceptable and deviant ANs, beyond the few hundreds we used here, and sampling a larger typology of ANs across frequency ranges and adjective and noun classes. To this extent, we are implementing a crowd-sourcing study to collect human judgments from a large pool of speakers on a much larger set of ANs unattested in the corpus. Averaging over multiple judgments, we will also be able to characterize semantic deviance as a gradient property, probably more accurately.

Next, the range of cues we used was quite limited, and we intend to extend the range to include more sophisticated methods such as 1) combining multiple cues in a single score; 2) training a supervised classifier from labeled acceptable and deviant ANs, and studying the most distinctive features discovered by the classifier; 3) trying more complex unsupervised techniques, such as using graph-theoretical methods to characterize the semantic neighborhood of ANs beyond our simple density measure.

Finally, we are currently not attempting a typology of deviant ANs. We do not distinguish cases such as *parliamentary tomato*, where the adjective does not ap-

ply to the conceptual semantic type of the noun (or at least, where it is completely undetermined which relation could bridge the two objects), from oxymorons such as *dry water*, or vacuously redundant ANs (*liquid water*) and so on. We realize that, at a more advanced stage of the analysis, some of these categories might need to be explicitly distinguished (for example, *liquid water* is odd but perfectly meaningful), leading to a multi-way task. Similarly, among acceptable ANs, there are special classes of expressions, such as idiomatic constructions, metaphors or other rhetorical figures, that might be particularly difficult to distinguish from deviant ANs. Again, more cogent tasks involving such well-formed but non-literal constructions (beyond the examples that ended up by chance in our acceptable set) are left to future work.

4.4 Experiment 2: Detecting semantic deviance using unsupervised measures

4.4.1 Experimental Setup

Composition models. The experiment was carried out across all compositional methods discussed in Section 2.2. The DL, W.ADD, F.ADD and LFM models include a variety of parameters which were estimated following the strategy proposed by Guevara (2010) and Baroni & Zamparelli (2010), recently extended to all composition models by Dinu *et al.* (2013b). Specifically, I learn parameter values that optimize the mapping from the noun to the AN as seen in examples of corpus-extracted N-AN vector pairs, using least-squares methods, or Ridge Regression in the case of LFM. All parameter estimations and phrase compositions were implemented using the DISSECT toolkit (Dinu *et al.*, 2013a, `http://clic.cimec.unitn.it/composes/toolkit`), with a training set of 74,767 corpus-extracted N-AN vector pairs, ranging from 100 to over 1K items across the 663 adjectives.

Dataset of Plausibility Judgments

Our goal is to study whether estimated distributional representations of unattested ANs (ANs which have never been seen in our large corpus and for which we have no distributional information) that are semantically deviant can recognized as such. In order to do this, we collected an evaluation dataset of human plausibility judgments on unattested ANs through a crowdsourcing experiment on CrowdFlower (CF, `http://www.crowdflower.com`) (Callison-Burch & Dredze, 2010; Munro *et al.*, 2010). As a first step, we defined a test set by extracting a random sample of 30K **unattested** ANs that resulted from the set described in Section 2.1 above (in which 1.42M were attested, and 1.17M were unattested). We reasoned that since adjective-noun is a simple and very frequent construction and the corpus we are working with is very large, the fact that our ANs are not attested should be due to one of the last two factors mentioned in the Introduction: they describe objects that are odd, rare or nonexistent (say, grooved tangerines, platinum screws or Martian senators), or the combination of A and N does not yield a comprehensible meaning. Of course, since both categories are fuzzy and probably partially overlapping, we had to put some care in the design of the test; if we were to ask participants to judge the acceptability of each AN using an absolute method such as the standard Likert scale (1-7), we might expect most ANs to remain at the lower-end of the scale. The distinction between 'odd because unfamiliar, yet acceptable', and 'semantically deviant' would not emerge. Thus, we designed the task in such a way that the participants were forced to make a binary choice on which of two ANs presented together made more sense. This way, we were able to analyze which variables significantly effected the choice of a more acceptable AN (see Section 4.4.2 for details on the analysis of the data).

We constructed a set of AN_x–AN_y pairs in which each of the test ANs were seen 5 times in position x and 5 times in position y without repetition of pairs, resulting in a collection of 150K pairs to be judged. The CF contributors were presented AN_x–AN_y pairs and asked to decide which of the two AN phrases makes *more* sense in each pair; for example, given the ANs *exact egg* and *Danish workplace*, the contributors would probably select the latter as the phrase that makes more sense (c.f. Appendix A.2.1 for a preview of the task as presented

to the contributors). We requested participants to be native speakers of English and only accepted judgments coming from an English-speaking country. Since the dataset is composed of unattested ANs and the pairs were constructed "blindly", it was likely that pairings consisting of two strange or incomprehensible ANs could arise. To address this possibility, contributors were also explicitly told to at least mark the one AN that seemed *less* strange. In addition, we instructed them to judge each AN regardless of which noun may follow it, i.e. as a complete phrase: for instance, *blind starch* would likely be judged unacceptable, regardless of the acceptability of *blind starch producer*).

CF offers a system of quality control, called Gold Standard Data, to determine the accuracy and trustworthiness of the participants. By pre-establishing the correct answers to a small set of data prior to collecting judgments, the system can then calculate the quality of a participant's performance and reject them if their accuracy drops below 70%. This gold data acts as hidden tests that are randomly shown to the participants as they complete the task. Although we cannot guarantee that non-native English speakers did not take part in the study, this system tried to ensure that only the data of speakers with a good command of English and sufficient motivation were retained. We therefore included a total of 180 "gold standard" items consisting of the acceptable vs. deviant ANs used in the Vecchi *et al.* (2011) plausibility experiment. To construct this dataset, we first extracted a randomly sampled set of 3K ANs unattested in our corpus, focusing on the 30 most frequent adjectives. Two authors went through this list, marking those ANs that they found semantically highly anomalous, no matter how much effort one would put in constructing metaphorical or context-dependent interpretations, as well as those they found completely acceptable (so, rating was on a 3-way scale: deviant, intermediate, acceptable). The rating exercise resulted in rather low agreement (Cohen's κ=0.32), but we reasoned that those relatively few cases (456 over 3K) where both judges agreed that the AN was odd should indeed be odd; similarly for the even rarer cases in which they agreed that an AN was completely acceptable (334 over 3K). We thus selected only the cases of agreement as our "gold" deviant and acceptable AN cases. At the end of this process, we selected a random sample of 90 acceptable and 90 deviant ANs and included them in the CF test set in the format AN_x–AN_y, where each pair

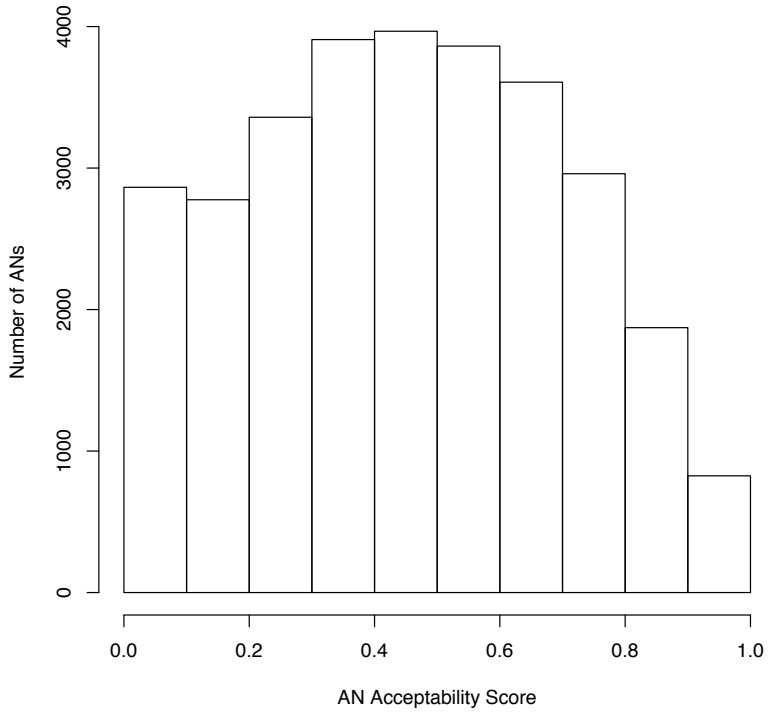

Figure 4.1: **Distribution of acceptability judgments.**

contained one acceptable and one deviant AN, in random order.

The resulting dataset used for human evaluation consisted of the 150K binary judgments, which had to determine which of the left-hand or right-hand AN was more sensical (or at least "less strange"). In total, we had 30K distinct ANs. The gold items were not included in the evaluation material. We can quantify a general score of acceptability on an AN-by-AN basis by computing how often the AN was chosen as the more acceptable phrase with respect to the number of times the AN was seen by participants. The general scores of acceptability follow a normal distribution, as seen in Fig. 4.1. The full evaluation dataset is publicly available and can be downloaded from `www.evavecchi.com`.

4.4.2 Methodology

Measures of Semantic Deviance

Our general goal is to determine which linguistically-motivated factors are involved in the choice of one unattested AN over another. In order to do so, we considered a number of unsupervised measures that could explain the plausibility judgments collected in the CF experiment described in Section 4.4.1.

Word-based measures. Psycholinguistic studies on compound processing give evidence that the family size (***family***) of a constituent, i.e., the number of times a word appears as a constituent of distinct compounds, plays a role in lexical processing (De Jong *et al.*, 2002). Elements that are associated with a large variety of lexical elements have high productivity, while words which only appear in combination with few other elements have low productivity. We hypothesize that highly productive adjectives and nouns correspond to a more flexible semantics; as a result, they should be found more often with acceptable ANs. For our purposes, the family size of adjective and nouns can be defined here as the number of times any given adjective or noun is seen in distinct corpus-attested AN phrases. Our prediction, then, is that high family size of component elements will yield higher acceptability of a novel AN phrase.

A potential measure we also considered was the raw frequency (***fq***) of the component elements in the source corpus. However, the results when using raw frequency were similar to those seen with family size; the two measures turned out to be very highly correlated[1], so for the experiments described here we only used family size.

In a number of lexical processing studies, string length (***slength***) has been known to influence word processing (Baayen *et al.*, 2006; New *et al.*, 2006). Further, the results from Bertram & Hyönä (2003) show that word length affects the processing of compounds. Here, we consider the effect that this variable may have on the choice of acceptability of novel phrases. In what follows, we consider the effect of the string length of component adjectives and nouns for each AN,

[1]The Spearman correlation between adjective family size and raw frequency is 0.67, and the Spearman correlation between noun family size and raw frequency is 0.71.

measured in letters. We hypothesize that longer component words might generally be more abstract, and may therefore be more flexible when integrating new modification. Denominal adjectives, for instance, are often relatively long, and can be very unspecified with respect to the relation that connects the noun root they contain with the AN head (see e.g. *industrial pollution* vs. *industrial site* vs. *industrial process*). Thus, we hypothesize that longer component adjectives and nouns should yield more acceptable ANs.

Distributional semantic measures. DSMs provide an apt framework to exploit the contextual information of phrases to detect deviance of novel phrases. Intuitively, we can expect acceptable phrases to share distributional qualities with sensical (attested) words and phrases already present in a large semantic space, while deviant phrases might fail to correspond to such distributions. Further, DSMs offer a way to quantify semantics in geometric terms, and so we can use them to define objective geometric measures of deviance. In Vecchi *et al.* (2011), we introduced a preliminary set of variables that exploit the geometric nature of these semantic representations to detect deviance in model-generated ANs. In this study, we consider these variables but also test additional measures extracted from the distributional semantic representation of the ANs and their component parts.

If deviant composition destroys or randomizes the meaning of a noun, as a side effect we might expect the resulting AN to be further away in meaning from the component noun. Although a *marble iPad* might have lost some essential properties of iPads (it could for example be an iPad statue you cannot use as a tablet), to the extent that we can make sense of it, it must retain at least some characteristics of iPads (at the very least, it will be shaped like one). On the other hand, we probably cannot converge on one good interpretation for *legislative onion* (laws written in layers? legislations that make you weep? food prescribed by a vegetarian dictator?), and thus cannot attribute it even a subset of the regular onion properties. For these reasons, we hypothesize that model-generated vectors of less acceptable ANs will be farther from component Ns as represented in the semantic space, forming a wider angle with the component N vectors, thus corresponding to lower ***cosine*** scores for less acceptable ANs (cf

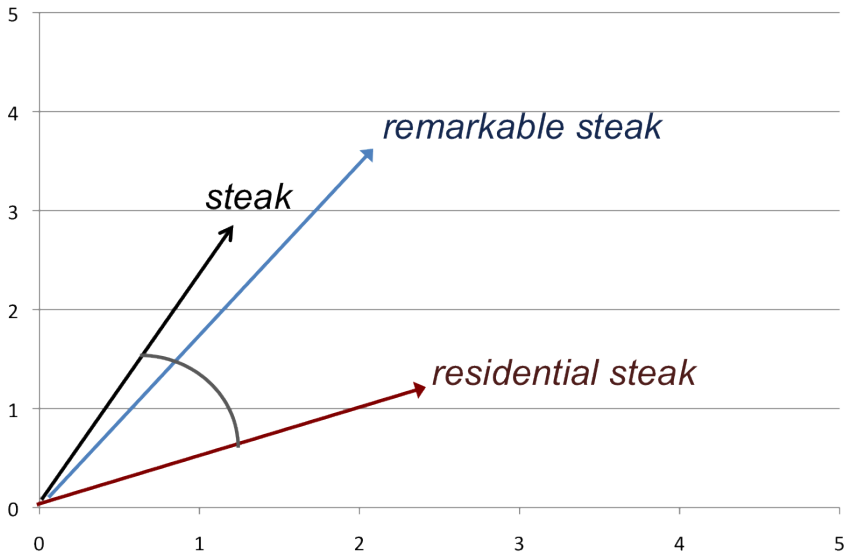

Figure 4.2: **Prediction for cosine.**

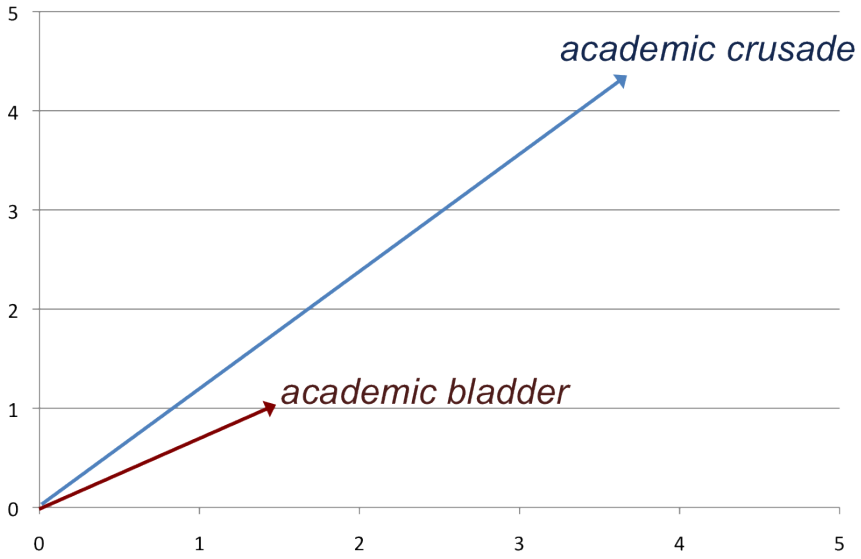

Figure 4.3: **Prediction for vector length.**

Fig. 4.2).

Next, we hypothesize that, since the values in the dimensions of a semantic space are a distributional proxy to the meaning of an expression, a meaningless expression should in general have low values across the semantic space dimensions. Thus, we predict the vector length (***vlength***) of a model-generated AN vector to be a significant factor in the choice of acceptable/unacceptable ANs: the shorter the vector the more likely the AN will be considered less acceptable (cf Fig. 4.3).

In Vecchi *et al.* (2011), we proposed a measure that reflected neighborhood ***isolation*** (previously entitled "density") based on the expectation that model-generated vectors of deviant ANs might have few neighbors in the semantic space, since our space is populated by nouns, adjectives and ANs that are frequently attested in our corpus and should thus be meaningful. This measure was calculated by simply taking the average of the cosines between the model-generated AN vector and its (top 10) nearest neighbors, expecting deviant ANs to be more isolated than acceptable ANs, corresponding to a lower average cosine score. Indeed, *smooth insecurity*, *printed capitalist* and *blind multiplier* were found in a more isolated neighborhood (average cosine score <0.55) than the more acceptable *cultural extremist*, *spectacular sauce* and *coastal summit* (average cosine score >0.75).

In this study, we expanded on this intuition and hypothesized that there may be a certain lack of coherence between the model-generated vector of deviants ANs and its nearest neighbors in our semantic space. Specifically, we expected that model-generated vectors for deviant ANs will share a neighborhood with elements that are not even similar amongst themselves, as they will not inhabit an area of space inhabited by coherent discourse topics. We predicted that ANs with a higher average similarity between all neighbors, or a higher neighborhood ***density***, would correspond to more acceptable ANs (cf Fig 4.4). Similarly to the isolation measure, we can operationalize this notion by taking the average of the cosines between each element in the neighborhood, which includes the (top 10) nearest neighbors as well as the model-generated AN. Though in theory the two measures are independent, in practice we found that the effects of the isolation and the density measures were highly correlated for all composition models[1].

[1]Spearman correlations between neighborhood isolation and neighborhood density for each

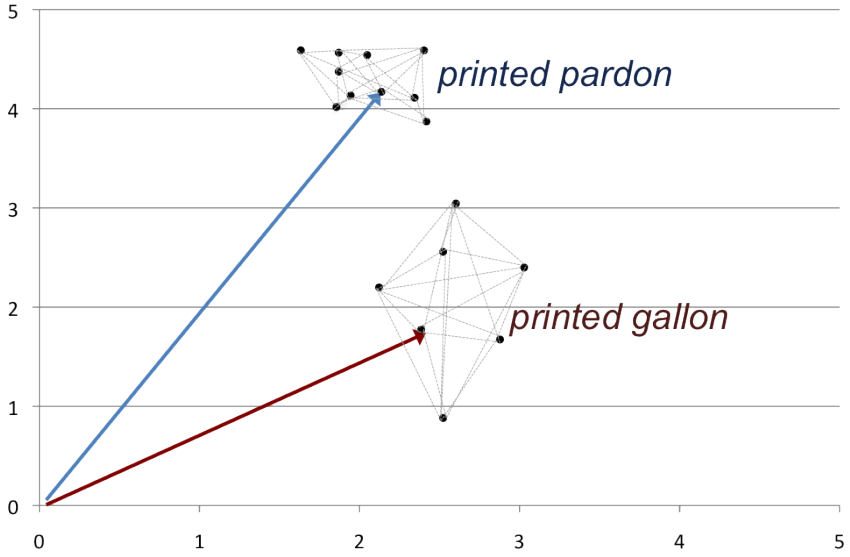

Figure 4.4: **Prediction for density.**

Thus, we report only the results for the density measure introduced here, since it is a more comprehensive description of the effect of neighborhood similarity.

Finally, since length, as already observed Vecchi *et al.* (2011), could be affected by independent factors such as input vector normalization and the estimation procedure, we test ***entropy*** as a measure of vector quality, introduced as a measure of plausibility in Lazaridou *et al.* (2013). The intuition provided by Lazaridou *et al.* is that meaningless vectors, whose dimensions contain mostly noise, should have a uniform distribution, yielding high entropy. While an acceptable AN vector, like *terrorist exchange* in Fig 4.5, should highlight the emphasis on a limited number of specific semantic contexts, resulting in a lower entropy score.

In a post-hoc analysis to better understand the behavior of the density measure (see Section 4.4.3), we also consider whether the acceptability of the AN is affected by the degree to which the component adjective transforms the meaning of the head noun in ANs, as seen in our semantic space. We hypothesize that adjectives that alter the meaning of nouns strongly in a uniform direction are less flexible, and therefore less acceptable in AN combinations not already attested in the corpus. For example, ANs containing adjectives such as *legal* or *nuclear*

composition model: ADD: 0.591; W.ADD: 0.875; MULT: 0.697; DL: 0.851; LFM: 0.885.

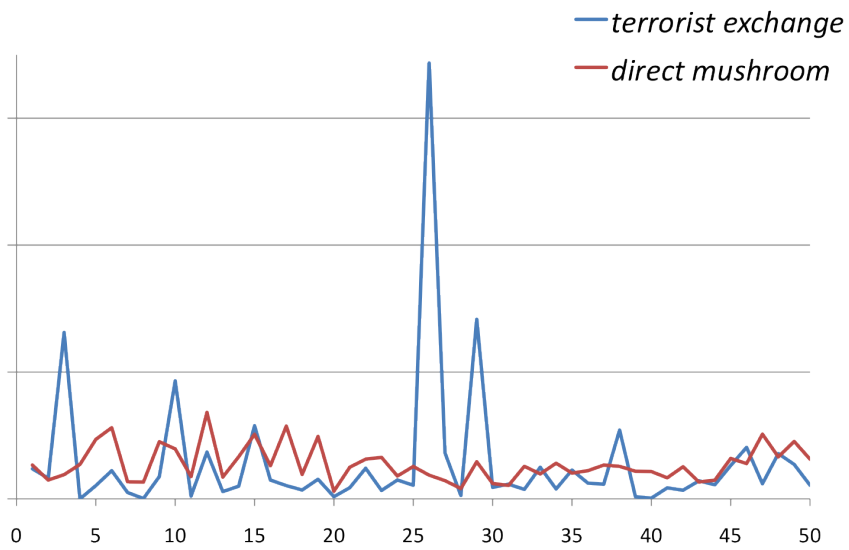

Figure 4.5: **Prediction for entropy.**

attribute quite specific properties to nouns and transform their meanings to a specific context, resulting in a restriction to nouns to which such properties can be attributed. Based on corpus-extracted vectors, we can compute an adjective ***densification*** measure that reflects the modification strength by determining the amount in which the adjective "pulls" nouns to a dense neighborhood in the semantic space, cf Fig. 4.6. Specifically, we compute a log ratio between the average density for a randomly selected set of 40 ANs per adjective (average cosines between all vectors) and the average density of their component nouns. Adjectives with a higher densification factor reflect a strong modification that "pulls" nouns to a dense area in the space (as seen with the adjective *nuclear* in Fig. 4.6), while a low densification factor implies the adjective has a weaker impact on the transformation of the head noun (as with the adjective *standard* in Fig. 4.6). The intuition behind this measure is that restrictions on the acceptability of an AN reflect the modification strength of the component adjective, and therefore may overlap with the N-AN cosine measure. However, the two measures differ significantly in that adjective densification describes the degree to which, given a component adjective, nouns are pulled into a denser nucleus in the semantic space when combined with it. Moreover, this measure applies on an adjective–

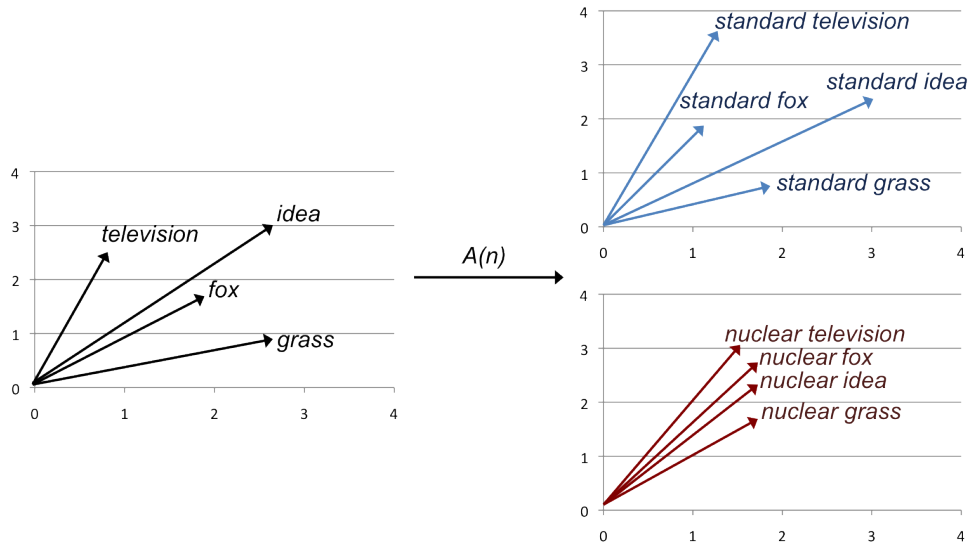

Figure 4.6: **Prediction for adjective densification.**

by–adjective basis—computing a single densification score per adjective—while the cosine measure only applies to ANs.

Data analysis

We estimate the effect of each measure on the participants' judgments by means of logit mixed effects models (Jaeger, 2008). This was aimed at testing how much the different measures would increase the likelihood of choosing one AN *over* the other. The dependent dichotomic variable was whether a participant was more likely to chose the first or the second element in the AN–AN pair. As proposed by Baayen *et al.* (2008), we introduced random intercepts of participants as well as data items (adjectives and nouns) in order to account for the random variance associated to judgers.

In the word-based model, we included 8 measures as independent variables: component adjective and noun family size for both the left- and right-hand AN, as well as component adjective and noun string length, again for both the left- and right-hand AN. Subsequent analyses were implemented in order to test the contribution of distributional semantic measures. Excluding densification, these measures were calculated based on the model-generated AN vectors for each com-

position function, resulting in a total of 14 tested variables. We introduce each measure for both the left- and right-hand ANs as additional independent variables to the word-based models in order to test a possible position effect. Moreover, we tested whether the introduction of each variable to the model significantly improved its goodness of fit, i.e., whether the result of the likelihood ratio test comparing the goodness-of-fit of the model before and after introducing the parameter was significant, cf Baayen *et al.* (2008).

4.4.3 Results

Word-based models

We find that simple word-based variables (described in Section 4.4.2) are significant factors when choosing which AN makes more sense, c.f. Table 4.4. These findings are consistent with results in previous psycholinguistic studies, and confirm the reliability of the plausibility data collected, since they are in line with previous studies on the processing of compounds, particularly with respect to component family size (see Section 4.2.1 for a description of these previous studies). Let's consider them in turn.

First, we found that the string length of the component adjectives and nouns significantly affect the acceptability of an AN: longer adjectives and nouns result in more acceptable ANs. As discussed above, this might be due to longer adjectives and nouns being more abstract, or establishing a more underspecified relation with each other (in particular, in the case of denominal adjectives; note, however, that we found the length of the noun to have a slightly stronger effect than the adjective on the choice of which AN makes more sense). Conversely, adjectives and nouns that apply to common, concrete objects tend to be shorter (think of *cat*, *sky*, *raw* and *big*). It is also possible that these results are due to an attention-capturing effect, in the sense that more attention is needed to evaluate the acceptability of longer strings. String length is the most significant among the word-based measures in the choice of acceptability.

Next, we find that more productive adjectives and nouns, i.e. those with a higher family size, yield more acceptable ANs. This result is quite intuitive, since we can expect adjectives and nouns with a high family size to be highly

Measure	*Estimate*	*Pr (> |z|)*
$A_L family$	-3.149e-04	≈ 0 ***
$A_R family$	3.823e-04	≈ 0 ***
$N_L family$	-1.803e-03	≈ 0 ***
$N_R family$	1.876e-03	≈ 0 ***
$A_L slength$	-6.967e-02	≈ 0 ***
$A_R slength$	7.137e-02	≈ 0 ***
$N_L slength$	-1.084e-01	≈ 0 ***
$N_R slength$	1.037e-01	≈ 0 ***

Table 4.4: **Word-based measures.** Results of the logit mixed effects models run on the CrowdFlower data using only word-based measures. The results include the effect of the ***family*** and ***slength*** of the component adjectives and nouns on the choice of acceptable ANs. For each measure, the polarity of the estimate indicates the likelihood of choosing the left-hand (L, negative) or right-hand (R, positive) AN as the more acceptable AN with respect to the variable. A larger estimate (absolute value) reflects a stronger effect on the choice of AN. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 .

productive, therefore less restrictive when combining with words to create new phrases. This measure, as with string length, has a stronger effect with respect to the component noun rather than the adjective. The unbalanced behavior between the effect of the adjective and noun family size may be due to a difference in family size distribution: nouns generally have a smaller family size (ranging between 6 and 660), while adjectives have a larger and broader distribution (ranging from 588 to 3892) which may dampen the effect. An additional factor influencing this effect could be the large number of nouns (3.9K) in comparison to adjectives (663) in our set of ANs.

Improvement on word-based models brought about by distributional semantic variables

The results for the word-based measures show that traditional psycholinguistic measures indeed have an effect on the processing of novel AN compounds. From here, we test whether the measures extracted from our distributional semantic representations improve the ability to predict the acceptability judgments of the unattested AN phrases.

Table 4.5 shows the results of the likelihood ratio test comparing the goodness-of-fit of the model using the word-based measures (string length and family size for the component elements) before and after introducing each distributional semantic measure. The goodness of fit improves most (i.e., high log likelihood and chi-squared values) with respect to the cosine from the component noun for all composition functions. The W.ADD, DL and LFM models are overall the best at improving the fit of the data for all measures. The only irregularity we find is that the MULT model does not improve the fit with respect to the density measure.

Overall, we find that measures extracted from distributional vectors significantly improve the fit of the plausibility data over simple word-based variables. This tells us that the choice of acceptability of novel phrases is semantically motivated and more complex than simple productivity, as tested in previous psycholinguistic studies using word-based measures.

Distributional semantic measures and composition models

Having shown that distributional semantic measures can explain the data beyond what traditional word-processing measures can account for, we will now focus more specifically on how the distributional measures alone can explain the data, and compare the different composition functions.

We find that vector length and cosine are the strongest and most consistent indicators of plausibility for all composition functions. First, all functions support our hypothesis that longer AN vectors results in more acceptable phrases. This suggests that each model is able to capture the intuition that a novel AN is more likely to be acceptable if the component adjective and noun have a more similar distribution in the source corpus, i.e., many common contexts lengthen the vector significantly.Next, the results in Table 4.6 show that a higher cosine between the model-generated AN and the corpus-extracted component noun vectors yields more acceptable AN phrases. This result implies that ANs that distort the meaning of the head noun more are considered less acceptable.

We find that most models, with the exception of the F.ADD model, are able to approximate the plausibility judgments with respect to the density measure, however the behavior of this measure varies greatly based on the model. The

Measure	*Model*	*Df*	*logLik*	*Chisq*	*Pr (>Chisq)*
WORD-BASED		12	-77393		
vlength	ADD	14	-76683	1420.9	≈ 0 ***
	W.ADD	14	-76684	1418.5	≈ 0 ***
	MULT	14	-76771	1243.3	≈ 0 ***
	DL	14	-77083	620.01	≈ 0 ***
	F.ADD	14	-76660	1465.6	≈ 0 ***
	LFM	14	-77022	742.63	≈ 0 ***
cosine	ADD	14	-76683	1420.9	≈ 0 ***
	W.ADD	14	-76702	1381.5	≈ 0 ***
	MULT	14	-76684	1417.6	≈ 0 ***
	DL	14	-77005	775.75	≈ 0 ***
	F.ADD	14	-77270	246.42	≈ 0 ***
	LFM	14	-76521	1744.4	≈ 0 ***
density	ADD	14	-77266	253.24	≈ 0 ***
	W.ADD	14	-77287	212.75	≈ 0 ***
	MULT	14	-77242	301.59	≈ 0 ***
	DL	14	-77291	203.55	≈ 0 ***
	F.ADD	14	-77304	177.95	≈ 0 ***
	LFM	14	-77300	186.56	≈ 0 ***
entropy	ADD	14	-77299	187.39	≈ 0 ***
	W.ADD	14	-77297	192.28	≈ 0 ***
	MULT	14	-77093	599.1	≈ 0 ***
	DL	14	-77277	231.79	≈ 0 ***
	F.ADD	14	-77269	248.64	≈ 0 ***
	LFM	14	-77255	275.35	≈ 0 ***

Table 4.5: **Improvement on word-based measures.** Results of the logit mixed effects models run on the CrowdFlower data: model/measure improvement on word-based measures. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 .

Model	*VLENGTH*	*COSINE*	*DENSITY*	*ENTROPY*
ADD	***	***	***	
W.ADD	***	***	**	
MULT	***	***	***	***
DL	***	***	**	***
F.ADD	***	***		***
LFM	***	***	*	***

Table 4.6: **Distributional semantic measures.** Results of the logit mixed effects models run on the CrowdFlower data including distributional semantic measures only. The results in black imply that *high* scores for the measure yield *acceptable* judgments, while results in red imply that *high* scores for the measure yield *unacceptable* judgments. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 .

ADD model performs as predicted for this measure, mainly, AN vectors found in a denser neighborhood tend to be more acceptable. On the other hand, although they are able to approximate our data with the density measure, the W.ADD, DL and LFM models do so in a direction contrary to our hypothesis. The results show that AN vectors with dense neighborhoods in the semantic space correspond to unacceptable phrases. In a qualitative analysis of the nearest neighbors, we found that the neighbors for unacceptable ANs with high density are more often similar to the meaning of the component adjective than acceptable ANs with high density. The examples in (9) list the nearest neighbors in the semantic space for a set of ANs with high neighborhood density, based on the results from the LFM composition method (here and below, we use asterisks to mark ANs with low acceptability scores; see the next section for additional examples).

(9) a. *animal metal {*animal*, *domestic animal*, *animal group*}
b. *nuclear fox {*nuclear development*, *nuclear danger*, *nuclear technology*}
c. warm garlic {*warm salad*, *red sauce*, *fresh salmon*}
d. spectacular striker {*spectacular goal*, *superb goal*, *crucial goal*}

We see that the nearest neighbors for the high-density, semantically deviant ANs in (9-a,b) are more similar in meaning to the component adjectives than the neighbors of high-density, acceptable ANs in (9-c,d). Furthermore, we find that neighbors for acceptable ANs with high density are more often similar to the

meaning of the component noun, while neighbors for unacceptable ANs do not maintain any meaning of the component noun. This result suggests that the adjective takes over the meaning in unacceptable ANs, "pulling" the AN to a place where the adjective dictates the meaning of all the neighbors, making them all similar (i.e., a denser neighborhood) and losing the meaning of the noun. Acceptable ANs are able to maintain the 'integrity' of the component noun, which keeps the AN from being placed into a neighborhood overruled by the meaning of the adjective and yields a sparser neighborhood. The result is also likely to be affected by the fact that the semantic space contains more ANs per adjective than per noun[1], making the adjective (or AN-sharing-the-same-A) neighborhoods artificially denser. Thus, if the meaning of the adjective overpowers the meaning of the AN in deviant cases, the composed meaning will likely occupy an area within this artificially denser neighborhood.

Adjective densification (a measure insensitive to the composition model since we computed it over the corpus-extracted AN data) has a slightly significant effect on the ability to model the plausibility judgments. The results show that unattested ANs that contain an adjective with a high densification factor are judged to be less acceptable phrases. This follows our intuition that a high densification factor implies a stronger adjective, which therefore generates an AN whose meaning is pulled further away from the head noun and into a neighborhood that is dominated by the adjective. This result supports and sheds light onto our findings for the density measure, which were contrary to our initial predictions.

Finally, we note that, like in the results reported in Section 4.3.3, the ***entropy*** measure is a significant variable in most models, however the direction of its effect fluctuates depending on the composition model. In the case of LFM, this measure is in line with our intuition, namely that ANs vectors with more noise (higher entropy scores) will be more semantically deviant. However, we find that the MULT, DL and F.ADD models result in an effect contrary to our hypothesis: AN vectors with higher entropy scores result in the more acceptable AN. In Table 4.7, we explore the highest/lowest entropy scores for significant models for this measure. Indeed, we confirm that in the case of LFM, ANs with lower

[1]There is an average of about 162 ANs per adjective in the semantic space, while there are only circa 30 ANs per noun.

entropy seem more semantically acceptable, while those with higher scores tend to be deviant. On the other hand, we notice the exact opposite effect with the examples for the MULT, DL and F.ADD models.

	Highest		*Lowest*	
model		ENTROPY		ENTROPY
MULT	surprising comrade	2.56	*safe alphabet	0.00
	lucky gardener	2.50	*online crop	0.00
	silent fame	2.45	*technological nail	0.01
	southern local	2.43	*graphic marriage	0.00
	rough belt	2.38	*affordable nominee	0.01
DL	terrible neighbor	4.02	*guilty mortgage	1.49
	massive villager	3.98	*vocal debit	1.82
	stunning handful	3.97	*disabled integer	1.94
	prestigious pair	3.96	*final pepper	1.94
	naval rest	3.96	*soft inning	1.94
F.ADD	popular parameter	5.43	*digital sauce	4.65
	Australian precision	5.41	*sexual cheese	4.68
	subsequent trap	5.42	*social onion	4.73
	legendary query	5.41	*statutory species	4.74
	tiny subsection	5.41	*criminal liver	4.77
LFM	*direct sauce	5.59	adverse youth	4.77
	*obvious flour	5.58	terrorist exchange	4.84
	*constant cake	5.57	mature flora	4.85
	*considerable blue	5.57	Democratic province	4.88
	*brief cow	5.55	archaeological finance	4.89

Table 4.7: Examples of the highest/lowest scores of the ***entropy*** measure for the significant models: MULT, DL, F.ADD and LFM. ANs with a low general acceptability score (<0.5) are marked with an (*).

Qualitative analysis of nearest neighbors

In addition to the analysis described above, we performed a qualitative analysis of the neighborhoods of the model-generated vectors as represented in our semantic space. In Table 4.8, we provide examples of the top 3 nearest neighbors for a set of ANs in our test set. Each composition model behaves quite differently with respect to both the types of words/phrases in the neighborhood and the distinction between acceptable and unacceptable ANs. It is clear that the nearest neighbors of the MULT function are quite odd for both acceptable and deviant ANs. The W.ADD and F.ADD models were able to model the acceptability judgements quite well, but we find that the nearest neighbors they predict are strongly related to the component noun in all ANs. The LFM, on the other hand, gives more importance to the modifier. The meaning of the adjective seems to take over for deviant ANs when using the LFM model, however we can see that in acceptable cases the nearest neighbors do represent the intuitive, functional combination of the meanings of the modifier and the head noun. Both the LFM and F.ADD seem to be the only composition models capable of capturing this.

	W.ADD	MULT	F.ADD	LFM
	fungus	cellar	several species	empty field
**empty fungus*	spore	dark passage	low plant	empty shell
	nematode	underground passage	Australian specie	empty area
	biscuit	jigsaw	basic recipe	mathematical idea
**mathematical biscuit*	crisp	sudoku	whole meal	mathematical problem
	chocolate	free child	original recipe	mathematical
	sunlight	financial loss	emotional disturbance	mental activity
**mental sunlight*	bright sunlight	omission	psychological response	psychological state
	glow	written warning	psychological problem	mental state
	monkey	free entertainment	African elephant	monthly programme
**monthly monkey*	parrot	fair ride	small monkey	monthly visitor
	gorilla	other entertainment	female elephant	educational publication
	flour	square inch	fresh cheese	wide mix
**wide flour*	white flour	yarn	natural juice	new presence
	white sugar	estimated weight	mature cheese	successful centre
	uprising	separate brigade	British occupation	continuous struggle
continuous uprising	revolt	major command	major revolt	continuous war
	armed uprising	rear operation	armed confrontation	long war
	farmland	flora	diverse environment	diverse area
diverse farmland	rich meadow	rare flora	distinctive area	rich diversity
	rich mosaic	diverse habitat	rich diversity	diverse life
	coordinator	employability	educational role	active part
important coordinator	educational role	effective learner	active role	important contact
	active role	lifelong	active interest	important appointment
	province	professional midfielder	former province	legendary city
legendary province	former province	Swedish ancestry	official capital	legendary figure
	current territory	British format	current territory	ancient land
	likelihood	cost-effectiveness	likelihood	systematic difference
systematic likelihood	statistical significance	systematic review	individual risk	systematic bias
	relative risk	economic evaluation	great likelihood	systematic relationship

Table 4.8: **Examples of the nearest neighbors of model-generated AN vectors.** We report the top three nearest neighbors of the AN vectors – generated using W.ADD, MULT, F.ADD and LFM – in the semantic space. The asterisk (*) implies that the general acceptability score of the AN in the CF experiment (i.e., the number of times it was chosen as the more acceptable AN with respect to the number of times it was seen by participants) is less than 0.2. While the other ANs reported here have a general acceptability score greater than 0.8.

4.4.4 Discussion

The aim of this study is to provide a new challenge to the computational distributional semantics community, namely that of characterizing what happens, distributionally, when composition leads to semantically anomalous composite expressions. The results of this study provide evidence that we are able to significantly model human intuitions about the semantic acceptability of novel AN phrases using simple, unsupervised cues.

We find that baseline psycholinguistic measures, such as string length and family size, approximate human judgments significantly. However, we also find that all indices of semantic deviance that we propose significantly improve the goodness of fit in comparison to the baseline measures. Although all composition functions were able to model human intuition about the acceptability of novel AN phrases, we found that the W.ADD, DL and LFM functions were overall the most consistent and significant winners.

The measures and functions that model human intuition provide insight into the semantic processing and the acceptability of novel AN phrases. Above all, we find that the degree in which the head noun is modified, or distorted, from its original meaning, is the most significant indicator of deviance. This is indicated by both the cosine measure and our interpretation of the density results (supported in turn by the densification patterns). Therefore, composition functions that are able to model this effect are in fact able to approximate semantic acceptability.

As a natural follow-up of this study, we intend to take a more fine-grained look at the data, studying e.g. the effect of the various measures and composition functions on specific subclasses of adjectives and nouns, or how specific A-N relations such as redundancy (i.e., *wooden tree*) or oxymorons (i.e., *dry liquid*) affect acceptability. We are also interested in expanding our CF experiment to include a judgment of relatedness between the unattested AN and its nearest neighbors. In addition, we would like to use these methods to study metaphors, as well as detect word order restrictions in recursive cases of adjective modification. Finally, we also hope to use supervised learning to discover which are the most important features to determine the acceptability of adjective-noun phrases.

Chapter 5

Behavior of recursive adjective modification

5.1 Introduction

A prominent approach for representing the meaning of a word in Natural Language Processing (NLP) is to treat it as a numerical vector that codes the pattern of co-occurrence of that word with other expressions in a large corpus of language Sahlgren (2006); Turney & Pantel (2010). This approach to semantics (sometimes called *distributional semantics*) scales well to large lexicons and does not require words to be manually disambiguated Schütze (1997). Until recently, however, this method had been almost exclusively limited to the level of single content words (nouns, adjectives, verbs), and had not directly addressed the problem of *compositionality* Frege (1892); Montague (1970); Partee (2004), the crucial property of natural language which allows speakers to derive the meaning of a complex linguistic constituent from the meaning of its immediate syntactic subconstituents.

Several recent proposals have strived to extend distributional semantics with a component that also generates vectors for complex linguistic constituents, using compositional operations in the vector space Baroni & Zamparelli (2010); Grefenstette & Sadrzadeh (2011a); Guevara (2010); Mitchell & Lapata (2010); Socher *et al.* (2012). All of these approaches construct distributional representations for novel phrases starting from the corpus-derived vectors for their lexical

constituents and exploiting the geometric quality of the representation. Such methods are able to capture complex semantic information of adjective-noun (AN) phrases, such as characterizing modification Boleda *et al.* (2012, 2013), and can detect semantic deviance in novel phrases Vecchi *et al.* (2011). Furthermore, these methods are naturally recursive: they can derive a representation not only for, e.g., *red car*, but also for *new red car*, *fast new red car*, etc. This aspect is appealing since trying to extract meaningful representations for all recursive phrases directly from a corpus will result in a problem of sparsity, since most large phrases will never occur in any finite sample.

Once we start seriously looking into recursive modification, however, the issue of modifier ordering restrictions naturally arises. Such restrictions have often been discussed in the theoretical linguistic literature Crisma (1991); Scott (2002); Sproat & Shih (1990), and have become one of the key ingredients of the 'cartographic' approach to syntax Cinque (2002). In this paradigm, the ordering is derived by assigning semantically different classes of modifiers to the specifiers of distinct functional projections, whose sequence is hard-wired.

While it is accepted that in different languages movement can lead to a principled rearrangement of the linear order of the modifiers Cinque (2010); Steddy & Samek-Lodovici (2011), one key assumption of the cartographic literature is that exactly one intonationally unmarked order for stacked adjectives should be possible in languages like English. The possibility of alternative orders, when discussed at all, is attributed to the presence of idioms (*high American building*, but *American high officer*), to asyndetic conjunctive meanings (e.g. *new creative idea* parsed as *[new & creative] idea*, rather than *[new [creative idea]]*), or to semantic category ambiguity for any adjective which appears in different orders (see Cinque (2004) for discussion).

In this study, we show that the existence of both rigid and flexible order cases is robustly attested at least for adjectival modification, and that flexible ordering is unlikely to reduce to idioms, coordination or ambiguity. Moreover, we show that at least for some recursively constructed adjective-adjective-noun phrases (AANs) we can extract meaningful representations from the corpus, approximating them reasonably well by means of compositional distributional semantic models, and that the semantic information contained in these models characterizes which AA

will have rigid order (as with *rapid social change* vs. **social rapid change*), or flexible order (e.g. *total estimated population* vs. *estimated total population*). In the former case, we find that the same distributional semantic cues discriminate between correct and wrong orders. Given that the existence of rigid ordering of adjectives is attributed to the semantic classes of modifiers, a good semantic representation should be able to capture restrictions in ordering due to their semantics.

To achieve these goals, we consider various properties of the distributional representation of AANs (both corpus-extracted and compositionally-derived), and explore their correlation with restrictions in adjective ordering. We conclude that measures that quantify the degree to which the modifiers have an impact on the distributional meaning of the AAN can be good predictors of ordering restrictions in AANs.

The rest of this chapter is structured as follows. The methodology and evaluation materials are detailed in Section 5.3, whereas the experiments' results are presented and analyzed in Section 5.4. It concludes by summarizing and proposing future directions in Section 5.5.

5.2 The syntax of adjectives

In Cinque (1990, 1994), Cinque proposed a head movement analysis to describe the DP-internal word order difference between Romance and Germanic languages. In Cinque (2010), however, he re-examines this analysis in order to address "its inability to capture the pattern of interpretive differences between pre- and postnominal adjectives in the two language families".

Chapter 1 outlines a number of problems for N-movement in Romance languages. First of all, the author points out the existence of a restriction on the number of postnominal adjectives which occur before a complement (or adjunct) of the N, a restriction that raises a problem in an analysis in which postnominal adjectives result from the head N raising past them. Cinque also provides evidence that postnominal adjectives in Romance languages are ordered in a way that is the mirror image of the order of adjectives found prenominally in Germanic languages. He notes that this is an unexpected phenomenon that becomes

problematic for his original analysis since it considers postnominal adjectives in Romance languages to be a consequence of N movement. Finally, he discusses cases in which a non predicative, postnominal adjective is able to take scope over the pronominal adjective in Romance languages. This result is unexpected and unexplained by the head movement analysis.

Cinque points out that the most serious of problems with the original head movement approach is that it does not provide a unified analysis for the fact that prenominal and postnominal adjectives differ in their interpretation in terms of a number of semantic distinctions. Specifically, he focuses on a pattern which runs in opposite directions in the two language families: prenominal adjectives in Germanic languages are ambiguous with respect to a number of semantic distinctions, while in postnominal position they have only one semantic value, and vice-versa in Romance languages. While some claim that pre- and postnominal adjectives in Romance languages can never have the same interpretations, Cinque claims that there do exist cases in which adjectives in Romance languages retain the meaning they have prenominally when found in postnominal position. He states that this conclusion is therefore problematic for Bouchard's (2002) analysis that claims that a shared meaning in the two positions is not possible.

Chapter 2 provides evidence using 9 levels of semantic distinction to demonstrate a systematic pattern of oppositions in the readings of adjectives between Germanic and Romance language families. These semantic distinctions include: stage-level vs. individual-level readings, restrictive vs. nonrestrictive readings, implicit relative clause vs. modal readings, intersective vs. nonintersective readings, relative vs. absolute readings, comparative vs. absolute readings of superlatives, specificity vs. non-specificity inducing readings, evaluative vs. epistemic readings of *unknown*, and NP-dependent vs. discourse anaphoric readings of *different*. Using these various semantic distinctions, Cinque displays that in English the prenominal position is systematically ambiguous between the values of each property, while only one value is possible in postnominal position. On the other hand, he shows that in Italian, the adjective in postnominal position is systematically ambiguous in each property, while the adjective in prenominal position has only one reading, specifically, the opposite values of those found in prenominal position in English.

Cinque states that if and when the two readings available prenominally in English cooccur, they are seen to follow a strict order: with the leftmost adjective corresponding to the postnominal reading. The asymmetric distribution between the two language families is further supported by evidence that when the readings available postnominally in Italian cooccur, they are systematically ordered in the opposite way: with the leftmost adjective corresponding to the prenominal reading.

Cinque points out that this systematic ordering highlights another problem for the N-movement analysis previously proposed: it cannot derive the desired generalizations within a unified Merge structure for Germanic and Romance languages. Specifically, together with N movement, no single structure of Merge for Germanic and Romance is able to derive the different patterns of interpretation found in prenominal and postnominal adjectives in both language families. However, Cinque claims that an alternative analysis in which the movement is of phrases containing the NP, rather than of only the N, would be compatible with a unique structure of Merge for Germanic and Romance as well as provide an account for observed generalizations.

In Chapter 3, Cinque provides evidence to support the idea that adnominal adjectives (APs) have two separate sources: a direct adnominal modification source and a (reduced) relative clause source. As demonstrated in Chapters 1-2, each source is associated with a value for the semantic distinctions that is the opposite of the value associated with the other source, leading to different interpretive properties of the two sources.

An additional interpretive difference between the two sources introduced here is that only direct modification adjectives can give rise to idiomatic readings. Cinque states that this is likely a consequence of the nonintersective nature of direct modification versus the necessarily intersective nature of indirect modification, which is not compatible with the semantic non-compositionality of idioms.

Beyond these semantic distinctions, Cinque highlights a number of syntactic properties associated with each source. First, direct modification adjectives are closer to the noun than adjectives deriving from relative clauses, as seen with English prenominal and Italian postnominal adjectives. This property is a consequence of the different heights at which relative clauses and direct modification

adjectives are merged.

A second syntactic difference between the two sources is the word order: direct modification adjectives are rigidly ordered while adjectives deriving from relative clauses are not. Although English and Italian appear not to have an absolutely rigid order, and instead a "preferred" or unmarked order, Cinque points out that this unmarked order corresponds to the rigid order of languages which do. Cinque shows that even in English or Italian direct modification adjectives are in fact rigidly ordered with cases in which the adjectives have no independent predicative usage and can therefore only be direct modifiers of the NP, like "classificatory" and "adverbial" adjectives, as seen in (1) and (2).

(1) a. La ripresa economica americana vs. *la ripresa americana economica
b. the American economic recovery vs. the *economic American recovery

(2) a. He is an occasional hard worker
b. *He is a hard occasional worker

Cinque suggests that the apparent non-rigid ordering of adjectives may be explained in cases where the lower adjective, in direct modification, can also be used predicatively and can then access the higher reduced relative clause source. The apparent freedom of adjective ordering is also found in cases where all adjectives involved can have a reduced relative clause source, or in instances of asyndetic coordination, or "parallel modification", where each adjective belongs to a separate intonational phrase and modifies the NP independently of the others. Apparent freedom in adjective ordering is also found whenever the lower of the two adjectives is in the (definite) superlative form. This is seen in examples like in (3) and (4) where the unmarked order of shape and color adjectives is reversed if either adjective is in the definite superlative form.

(3) a. a long white plane
b. %a white long plane

(4) a. *?the long whitest plane (that I saw)
b. the whitest long place (that I saw)

Cinque claims that these cases of apparent free order and order reversals are not sufficient to conclude that no ordering exists among direct modification adjectives in English and Italian.

Cinque also provides cross-linguistic and acquisitional evidence for the dual source of adnominal adjectives. For example, in languages like Slave (Athapaskan) and Lango adjectives can be used as predicates (also within a relative clause), but not as adnominal (direct modification) attributes, while adjectives in languages such as Yoruba can appear only in adnominal position, not in predicate position. In addition, Cinque argues that the fact that stage-level adjectives systematically appear later than individual-level adjectives in both English and Italian is evidence that acquisition of indirect modification is delayed with respect to that of direct modification.

Cinque claims that only phrasal movement, or movement of phrases containing the NP, plays a role in the grammar of Romance and Germanic languages. As discussed in Chapters 1-2, an N-movement analysis of adjectives is unable to derive generalizations for the different patterns of interpretation found in prenominal and postnominal adjectives within a unified Merge structure for Germanic and Romance languages. Cinque also discards the possibility of a base generation analysis based primarily on the fact that cross-linguistically one finds prenominally only one order, while postnominally there are (at least) two; either the same as the prenominal order, or its exact opposite. This is the case also for the order of direct modification adjectives as seen in (5).

(5)

a.	$A_{size} > A_{color} > A_{nationality} > N$	(English, Chinese, ...)
b.	$*A_{nationality} > A_{color} > A_{size} > N$	0
c.	$N > A_{size} > A_{color} > A_{nationality}$	(Welsh, Irish, ...)
d.	$N > A_{nationality} > A_{color} > A_{size}$	(Indonesian, Yoruba, ...)

Since each of these orders would have to be generated independently of the others under a base generation analysis, an absolute principle, rather than just a tendency, would have to adopt an abstract, asymmetric, view in which there is only one order available for all languages, and any variation in this is a function of independently motivated types of movement. However, Cinque compares this

with the fact that languages vary with respect to whether or not they displace interrogative *wh*-phrases, and that the movement can affect just the phrase bearing the feature triggering the movement, or a larger phrase containing the phrase bearing the relevant feature, i.e. Pied Piping. Cinque argues that precisely these two independent parameters can account for the three attested orders found in (5) and for the principled absence of the fourth ((5-b) cannot be derived because the NP has not moved and the base structure has the modifiers in the wrong order). The author states that Anationality, Acolor or Asize cannot move by themselves just as phrases not bearing the *wh*-feature cannot move by themselves. This comparison supports the claim that a phrasal movement analysis is better equipped than either a N-movement or a base generation analysis.

5.3 Materials and methods

5.3.1 Expansion of semantic space

Our initial step was to construct a *semantic space* for our experiments, consisting of a matrix where each row represents the meaning of an adjective, noun, AN or AAN as a distributional vector, each column a semantic dimension of meaning. We first introduce the source corpus, then the vocabulary of words and phrases that we represent in the space, and finally the procedure adopted to build the vectors representing the vocabulary items from corpus statistics, and obtain the semantic space matrix. We work here with a traditional, window-based semantic space, since our focus is on the effect of different composition methods given a common semantic space. In addition, Blacoe & Lapata (2012) found that a vanilla space of this sort performed best in their composition experiments, when compared to a syntax-aware space and to neural language model vectors such as those used for composition by Socher *et al.* (2011).

Semantic space vocabulary. The words/phrases in the semantic space must of course include the items that we need for our experiments (adjectives, nouns, ANs and AANs used for model training, as input to composition and for evaluation). Therefore, we first populate our semantic space with a core vocabulary

containing the 8K most frequent nouns and the 4K most frequent adjectives from the corpus.

The ANs included in the semantic space are composed of adjectives with very high frequency in the corpus so that they are generally able to combine with many classes of nouns. They are composed of the 700 most frequent adjectives and 4K most frequent nouns in the corpus, which were manually controlled for problematic cases – excluding adjectives such as *above*, *less*, or *very*, and nouns such as *cant*, *mph*, or *yours* – often due to tagging errors. We generated the set of ANs by crossing the filtered 663 adjectives and 3,910 nouns. We include those ANs that occur at least 100 times in the corpus in our vocabulary, which amounted to a total of 128K ANs.

Finally, we created a set of AAN phrases composed of the adjectives and nouns used to generate the ANs. Additional preprocessing of the generated A_xA_yNs includes: (i) control that both A_xN and A_yN are attested in the corpus; (ii) discard any A_xA_yN in which A_xN or A_yN are among the top 200 most frequent ANs in the source corpus (as in this case, order will be affected by the fact that such phrases are almost certainly highly lexicalized); and (iii) discard AANs seen as part of a conjunction in the source corpus (i.e., where the two adjectives appear separated by comma, *and*, or *or*; this addresses the objection that a flexible order AAN might be a hidden A(&)A conjunction: we would expect that such a conjunction should also appear overtly elsewhere). The set of AANs thus generated is then divided into two types of adjective ordering:

1. **Flexible Order** (FO): phrases where *both* orders, A_xA_yN and A_yA_xN, are attested (f>10 in both orders).

2. **Rigid Order** (RO): phrases with *one* order, A_xA_yN, attested (20<f<200)[1] and A_yA_xN unattested.

All AANs that did not meet either condition were excluded from our semantic space vocabulary. The preserved set resulted in 1,438 AANs: 621 flexible order

[1]The upper threshold was included as an additional filter against potential multiword expressions. Of course, the boundary between phrases that are at least partially compositional and those that are fully lexicalized is not sharp, and we leave it to further work to explore the interplay between the semantic factors we study here and patterns of lexicalization.

and 817 rigid order. Note that there are almost as many flexible as rigid order cases; this speaks against the idea that free order is a marginal phenomenon, due to occasional ambiguities that reassign the adjective to a different semantic class. The existence of freely ordered stacked adjectives is a robust phenomenon, which needs to be addressed.

Semantic vector construction For each of the items in our vocabulary, we first build 10K-dimensional vectors by recording the item's sentence-internal co-occurrence with the top 10K most frequent content lemmas (nouns, adjectives, verbs or adverbs) in the corpus. We built a rank of these co-occurrence counts, and excluded as stop words from the dimensions any element of any POS whose rank was from 0 to 300. The raw co-occurrence counts were then transformed into (positive) Pointwise Mutual Information (pPMI) scores Church & Hanks (1990). Next, we reduce the full co-occurrence matrix to 300 dimensions applying the Non-negative Matrix Factorization (NMF) operation Lin (2007). We did not tune the semantic vector construction parameters, since we found them to work best in a number of independent earlier experiments.

Corpus-extracted vectors (CORP) were computed for the ANs and for the flexible order and attested rigid order AANs, and then mapped onto the 300-dimension NMF-reduced semantic space. As a sanity check, the first row of Table 2.2 reports the correlation between the AN phrase similarity ratings collected in Mitchell & Lapata (2010) and the cosines of corpus-extracted vectors in our space, for the same ANs. For the AAN vectors, which are sparser, we used human judgements to build a reliable subset to serve as our gold standard, as detailed in Section 5.3.4.

5.3.2 Recursive compositional distributional semantics

One the one hand, FS semantic representations in terms of logical formulas are able to represent and account for compositionality, however they are not well suited to modeling similarity quantitatively as they are based on discrete symbols. On the other hand, DSMs can easily measure similarity but they are not naturally compositional. As a result, current research in Computational Linguistics and Cognitive Science attempts to incorporate compositionality in DSMs (Baroni & Zamparelli, 2010; Erk & Padó, 2008; Guevara, 2010; Mitchell & Lapata, 2008). Following the insights gained from FS, the principle of compositionality, and current implementations of DSMs, this work aims at modeling the compositional phenomena in natural language semantics in a natural and linguistically relevant manner.

Semantic representations of single words can be represented as vectors in high-dimensional DSMs. By exploiting the geometric nature of these representations, given two independent vectors v_1 and v_2 in the space, we can then combine the independent vectors to produce a semantically compositional result v_3. Attempts in this task have explored a number of possible operations to combine these vectors, described in detail below. We can measure the success of such approaches in terms of their ability to model semantic properties of simple phrases, in tasks such as phrase similarity (Baroni & Zamparelli, 2010; Erk & Padó, 2008; Grefenstette & Sadrzadeh, 2011b; Mitchell & Lapata, 2010), textual entailment, semantic plausibility analysis (Vecchi *et al.*, 2011), and sentiment analysis (Socher *et al.*, 2011).

Compositional methods. We focus on four composition functions proposed in recent literature with high performance in a number of semantic tasks. We first consider methods proposed by Mitchell & Lapata (2010) in which the model-generated vectors are simply obtained through component-wise operations on the constituent vectors. Given input vectors $\vec{u}$ and $\vec{v}$, the multiplicative model (MULT) computes a composed vector by component-wise multiplication ($\odot$) of the constituent vectors, where the i-th component of the composed vector is given by $c_i = u_i v_i$. Given an $\mathrm{A}_x\mathrm{A}_y\mathrm{N}$ phrase, this model extends naturally to the recursive

setting of this experiment, as seen in Equation (5.1).

$$\vec{c} = \vec{a}_x \odot \vec{a}_y \odot \vec{n} \tag{5.1}$$

This composition method is order-insensitive, the formula above corresponding to the representation of both A_xA_yN and A_yA_xN.

In the weighted additive model (W.ADD), we obtain the composed vector as a weighted sum of the two component vectors: $\vec{c} = \alpha\vec{u} + \beta\vec{v}$, where α and β are scalars. Again, we can easily apply this function recursively, as in Equation (5.2).

$$\vec{c} = \alpha\vec{a}_x + \beta(\alpha\vec{a}_y + \beta\vec{n}) = \alpha\vec{a}_x + \alpha\beta\vec{a}_y + \beta^2\vec{n} \tag{5.2}$$

We also consider the full extension of the additive model (F.ADD), presented in Guevara (2010) and Zanzotto *et al.* (2010), such that the component vectors are pre-multiplied by weight matrices before being added: $\vec{c} = \mathbf{W}_1\vec{u} + \mathbf{W}_2\vec{v}$. Similarly to the W.ADD model, Equation (5.3) describes how we apply this function recursively.

$$\begin{aligned} \vec{c} &= \mathbf{W}_1\vec{a}_x + \mathbf{W}_2(\mathbf{W}_1\vec{a}_y + \mathbf{W}_2\vec{n}) \\ &= \mathbf{W}_1\vec{a}_x + \mathbf{W}_2\mathbf{W}_1\vec{a}_y + \mathbf{W}_2^2\vec{n} \end{aligned} \tag{5.3}$$

Finally, we consider the lexical function model (LFM), first introduced in Baroni & Zamparelli (2010), in which attributive adjectives are treated as functions from noun meanings to noun meanings. This is a standard approach in Montague semantics Thomason (1974), except noun meanings here are distributional vectors, not denotations, and adjectives are (linear) functions learned from a large corpus. In this model, predicted vectors are generated by multiplying a function matrix $\mathbf{U}$ with a component vector: $\vec{c} = \mathbf{U}\vec{v}$. Given a weight matrix, $\mathbf{A}$, for each adjective in the phrase, we apply the functions in sequence recursively as shown in Equation (5.4).

$$\vec{c} = \mathbf{A}_x(\mathbf{A}_y\vec{n}) \tag{5.4}$$

Composition model estimation Parameters for W.ADD, F.ADD and LFM were estimated following the strategy proposed by Guevara (2010) and Baroni & Zamparelli (2010), recently extended to all composition models by Dinu *et al.* (2013b). Specifically, we learn parameter values that optimize the mapping from the noun to the AN as seen in examples of corpus-extracted N-AN vector pairs, using least-squares methods, or Ridge Regression in the case of LFM. All parameter estimations and phrase compositions were implemented using the DISSECT toolkit[1] Dinu *et al.* (2013a), with a training set of 74,767 corpus-extracted N-AN vector pairs, ranging from 100 to over 1K items across the 663 adjectives. Importantly, while below we report experimental results on capturing various properties of recursive AAN constructions, no AAN was seen during training, which was based entirely on mapping from N to AN. Table 2.2 reports the results attained by our model implementations on the Mitchell and Lapata AN similarity data set.

5.3.3 Measures of adjective ordering

Our general goal is to determine which linguistically-motivated factors distinguish the two types of adjective ordering. We hypothesize that in cases of flexible order, the two adjectives will have a similarly strong effect on the noun, thus transforming the meaning of the noun equivalently in the direction of both adjectives and component ANs. For example, in the phrase *creative new idea*, the *idea* is both *new* and *creative*, so we would expect a similar impact of modification by both adjectives.

On the other hand, we predict that in rigid order cases, one adjective, the one closer to the noun, will dominate the meaning of the phrase, distorting the meaning of the noun by a significant amount. For example, the phrase *different architectural style* intuitively describes an *architectural style* that is *different*, rather than a *style* that is to the same extent *architectural* and *different*.

We consider a number of measures that could capture our intuitions and quantify this difference, exploring the distance relationship between the AAN vectors and each of the AAN subparts. First, we examine how the similarity

[1] `http://clic.cimec.unitn.it/composes/toolkit`

of an AAN to its component adjectives affects the ordering, using the cosine between the A_xA_yN vector and each of the component A vectors as an expression of similarity (we abbreviate this as $cosA_x$ and $cosA_y$ for the first and second adjective, respectively).[1] Our hypothesis predicts that flexible order AANs should remain similarly close to both component As, while rigid order AANs should remain systematically closer to their A_y than to their A_x.

Next, we consider the similarity between the A_xA_yN vector and its component N vector ($cosN$). This measure is aimed at verifying if the degree to which the meaning of the head noun is distorted could be a property that distinguishes the two types of adjective ordering. Again, vectors for flexible order AANs should remain closer to their component nouns in the semantic space, while rigid order AANs should distort the meaning of the head noun more notably.

We also inspect how the similarity of the AAN to its component AN vectors affects the type of adjective ordering ($cosA_xN$ and $cosA_yN$). Considering the examples above, we predict that the flexible order AAN *creative new idea* will share many properties with both *creative idea* and *new idea*, as represented in our semantic space, while rigid order AANs, like *different architectural style*, should remain quite similar to the A_yN, i.e., *architectural style*, and relatively distant from the A_xN, i.e., *different style*.

Finally, we consider a measure that does not exploit distributional semantic representations, namely the difference in PMI between A_xN and A_yN (ΔPMI). Based on our hypothesis described for the other measures, we expect the association in the corpus of A_yN to be much greater than A_xN for rigid order AANs, resulting in a large negative ΔPMI values. While flexible order AANs should have similar association strengths for both A_xN and A_yN, thus we expect ΔPMI to be closer to 0 than for rigid order AANs.

5.3.4 Gold standard

To our knowledge, this is the first study to use distributional representations of recursive modification; therefore we must first determine if the composed AAN

[1] In the case of LFM, we compare the similarity of the AAN with the AN centroids for each adjective, since the model does not make use of A vectors Baroni & Zamparelli (2010).

vector representations are semantically coherent objects. Thus, for vector analysis, a *gold standard* of 320 corpus-extracted AAN vectors were selected and their quality was established by inspecting their nearest neighbors. In order to create the gold standard, we ran a crowdsourcing experiment on CrowdFlower[1] Callison-Burch & Dredze (2010); Munro *et al.* (2010), as follows.

First, we gathered a randomly selected set of 600 corpus-extracted AANs, containing 300 flexible order and 300 attested rigid order AANs. We then extracted the top 3 nearest neighbors to the corpus-extracted AAN vectors as represented in the semantic space[2]. Each AAN was then presented with each of the nearest neighbors, and participants were asked to judge "how strongly related are the two phrases?" on a scale of 1-7. The rationale was that if we obtained a good distributional representation of the AAN, its nearest neighbors should be closely related words and phrases. Each pair was judged 10 times, and we calculated a *relatedness* score for the AAN by taking the average of the 30 judgments (10 for each of the three neighbors).

The final set for the gold standard contains the 320 AANs (152 flexible order and 168 attested rigid order) which had a relatedness score over the median-split (3.9). Table 5.1 shows examples of gold standard AANs and their nearest neighbors. As these example indicate, the gold standard AANs reside in semantic neighborhoods that are populated by intuitively strongly related expressions, which makes them a sensible target for the compositional models to approximate.

We also find that the neighbors for the AANs represent an interesting variety of types of semantic similarity. For example, the nearest neighbors to the corpus-extracted vectors for *medieval old town* and *rapid social change* include phrases which describe quite complex associations, cf. Table 5.1. In addition, we find that the nearest neighbors for flexible order AAN vectors are not necessarily the same for both adjective orders, as seen in the difference in neighbors of *national daily newspaper* and *daily national newspaper*. We can expect that the change in order, when acceptable and frequent, does not necessarily yield synonymous phrases, and that corpus-extracted vector representations capture subtle differences in

[1] `http://www.crowdflower.com`

[2] The top 3 neighbors included adjectives, nouns, ANs and AANs. The preference for ANs and AANs, as seen in Table 5.1, is likely a result of the dominance of those elements in the semantic space (c.f. Section 5.3.1).

medieval old town	*contemp. political issue*
fascinating town	cultural topic
impressive cathedral	contemporary debate
medieval street	contemporary politics
rural poor people	*British naval power*
poor rural people	naval war
rural infrastructure	British navy
rural people	naval power
friendly helpful staff	*last live performance*
near hotel	final gig
helpful staff	live dvd
quick service	live release
creative new idea	*rapid social change*
innovative effort	social conflict
creative design	social transition
dynamic part	cultural consequence
national daily newspaper	*new regional government*
national newspaper	regional government
major newspaper	local reform
daily newspaper	regional council
daily national newspaper	*fresh organic vegetable*
national daily newspaper	organic vegetable
well-known journalist	organic fruit
weekly column	organic product

Table 5.1: Examples of the nearest neighbors of the gold standard, both flexible order (left column) and rigid order (right column) AANs.

	Gold	*FO*	*RO*
W.ADD	0.565	0.572	0.558
F.ADD	0.618	0.622	0.614
MULT	0.424	0.468	0.384
LFM	0.655	0.675	0.637

Table 5.2: Mean cosine similarities between the corpus-extracted and model-generated gold AAN vectors. All pairwise differences between models are significant according to Bonferroni-corrected paired t-tests ($p<0.001$). For MULT and LFM, the difference between mean flexible order (FO) and rigid order (RO) cosines is also significant.

meaning.

5.4 Results

5.4.1 Quality of model-generated AAN vectors

Our nearest neighbor analysis suggests that the corpus-extracted AAN vectors in the gold standard are meaningful, semantically coherent objects. We can thus assess the quality of AANs recursively generated by composition models by how closely they approximate these vectors. We find that the performances of most composition models in approximating the vectors for the gold AANs is quite satisfactory (cf. Table 5.2). To put this evaluation into perspective, note that 99% of the simulated distribution of pairwise cosines of corpus-extracted AANs is below the mean cosine of the worst-performing model (MULT), that is, a cosine of 0.424 is very significantly above what is expected by chance for two random corpus-extracted AAN vectors. Also, observe that the two more parameter-rich models are better than W.ADD, and that LFM also significantly outperforms F.ADD.

Further, the results show that the models are able to approximate flexible order AAN vectors better than rigid order AANs, significantly so for LFM and MULT. This result is quite interesting because it suggests that flexible order AANs express a more literal (or intersective) modification by both adjectives, which is what we would expect to be better captured by compositional models. Clearly, a

more complex modification process is occurring in the case of rigid order AANs, as we predicted to be the case.

5.4.2 Distinguishing flexible vs. rigid order

In the results reported below, we test how both our baseline ΔPMI measure and the distance from the AAN and its component parts changes depending on the type of adjective ordering to which the AAN belongs. From this point forward, we only use gold standard items, where we are sure of the quality of the corpus-extracted vectors. The first block of Table 5.3 reports the *t*-normalized difference between flexible order and rigid order mean cosines for the corpus-extracted vectors.

These results show, in accordance with our considerations in Section 5.3.3 above: (i) flexible order A_xA_yNs are closer to A_xN and the component N than rigid order A_xA_yNs, and (ii) rigid order A_xA_yNs are closer to their A_y (flexible order AANs are also closer to A_x but the effect does not reach significance).[1] The results imply that the degree of modification of the A_y on the noun is a significant indicator of the type of ordering present.

In particular, rigid order A_xA_yNs are heavily modified by A_y, distorting the meaning of the head noun in the direction of the closest adjective quite drastically, and only undergoing a slight modification when the A_x is added. In other words, in rigid order phrases, for example *rapid social change*, the A_yN expresses a single concept (probably a "kind", in the terminology of formal semantics), strongly related to *social*, *social change*, which is then modified by the A_x. Thus, the *change* is not both *social* and *rapid*, rather, the *social change* is *rapid*. On the other hand, flexible order AANs maintain the semantic value of the head noun while being modified only slightly by both adjectives, almost equivalently. For example, in the phrase *friendly helpful staff*, one is saying that the *staff* is both *friendly* and *helpful*. Most importantly, the corpus-extracted distributional representations are able to model this phenomenon inherently and can significantly

[1] As an aside, the fact that mean cosines are significantly larger for the flexible order class in two cases but for the rigid order class in another addresses the concern, raised by a reviewer, that the words and phrases in one of the two classes might systematically inhabit denser regions of the space than those of the other class, thus distorting results based on comparing mean cosines.

	Measure	*t*	*sig.*	
CORP	$\cos A_x$	2.478		
	$\cos A_y$	-4.348	*	RO>FO
	cosN	4.656	*	FO>RO
	$\cos A_x N$	5.913	*	FO>RO
	$\cos A_y N$	1.970		
W.ADD	$\cos A_x$	4.805	*	FO>RO
	$\cos A_y$	-1.109		
	cosN	1.140		
	$\cos A_x N$	1.059		
	$\cos A_y N$	0.584		
F.ADD	$\cos A_x$	2.050		
	$\cos A_y$	-1.451		
	cosN	4.493	*	FO>RO
	$\cos A_x N$	-0.445		
	$\cos A_y N$	2.300		
MULT	$\cos A_x$	3.830	*	FO>RO
	$\cos A_y$	-0.503		
	cosN	5.090	*	FO>RO
	$\cos A_x N$	4.435	*	FO>RO
	$\cos A_y N$	3.900	*	FO>RO
LFM	$\cos A_x$	-1.649		
	$\cos A_y$	-1.272		
	cosN	5.539	*	FO>RO
	$\cos A_x N$	3.336	*	FO>RO
	$\cos A_y N$	4.215	*	FO>RO
ΔPMI		8.701	*	FO>RO

Table 5.3: **Flexible vs. Rigid Order AANs.** *t*-normalized differences between flexible order (FO) and rigid order (FO) mean cosines (or mean ΔPMI values) for corpus-extracted and model-generated vectors. For significant differences ($p<0.05$ after Bonferroni correction), the last column reports whether mean cosine (or ΔPMI) is larger for flexible order (FO) or rigid order (RO) class.

distinguish the two adjective orders.

The results of the composition models (cf. Table 5.3) show that for all models at least some properties do distinguish flexible and rigid order AANs, although only MULT and LFM capture the two properties that show the largest effect for the corpus-extracted vectors, namely the asymmetry in similarity to the noun and the A_xN (flexible order AANs being more similar to both).

It is worth remarking that MULT approximated the patterns observed in the corpus-extracted vectors quite well, despite producing order-insensitive representations of recursive structures. For flexible order AANs, order is indeed only slightly affecting the meaning, so it stands to reason that MULT has no problems modeling this class. For rigid order AANs, where we consider here the attested-order only, evidently the order-insensitive MULT representation is sufficient to capture their relations to their constituents.

Finally, we see that the ΔPMI measure is the best at distinguishing between the two classes of AAN ordering. This confirms our hypothesis that a lot has to do with how integrated A_y and N are. While it is somewhat disappointing that ΔPMI outperforms all distributional semantic cues, note that this measure conflates semantic and lexical factors, as the high PMI of A_yN in at least some rigid order AANs might be also a cue of the fact that the latter bigram is a lexicalized phrase (as discussed in footnote 1, it is unlikely that our filtering strategies sifted out all multiword expressions). Moreover, ΔPMI does not produce a semantic representation of the phrase (see how composed distributional vectors approximate of high quality AAN vectors in Table 5.2). Finally, this measure will not scale up to cases where the ANs are not attested, whereas measures based on composition only need corpus-harvested representations of adjectives and nouns.

5.4.3 Properties of the correct adjective order

Having shown that flexible order and rigid order AANs are significantly distinguished by various properties, we proceed now to test whether those same properties also allow us to distinguish between correct (corpus-attested) and wrong (unattested) adjective ordering in rigid AANs (recall that we are working with cases where the attested-order occurs more than 20 times in the corpus, and both

adjectives modify the nouns at least 10 times, so we are confident that there is a true asymmetry).

We expect that the fundamental property that distinguishes the orders is again found in the degree of modification of both component adjectives. We predict that the single concept created by the A_yN in attested-order rigid AANs, such as *legal status* in *formal legal status*, is an effect of the modification strength of the A_y on the head noun, and when seen in the incorrect ordering, i.e., ?*legal formal status*, the strong modification of *legal* will still dominate the meaning of the AAN. Composition models should be able to capture this effect based on the distance from both the component adjectives and ANs.

Clearly, we cannot run these analyses on corpus-extracted vectors since the unattested order, by definition, is not seen in our corpus, and therefore we cannot collect co-occurrence statistics for the AAN phrase. Thus, we test our measures of adjective ordering on the model-generated AAN vectors, for all gold rigid order AANs in both orders.

We also consider the ΔPMI measure which was so effective in distinguishing flexible vs. rigid order AANs. We expect that the greater association with A_yN for attested-order AANs will again lead to large, negative differences in PMI scores, while the expectation that unattested-order AANs will be highly associated with their A_xN will correspond to large, positive differences in PMI.

Across all composition models, we find that the distance between the model-generated AAN and its component adjectives, A_x and A_y, are significant indicators of attested vs. unattested adjective ordering (cf. Table 5.4). Specifically, we find that rigid order AANs in the correct order are closest to their A_y, while we can detect the unattested order when the rigid order AAN is closer to its A_x. This finding is quite interesting, since it shows that the order in which the composition functions are applied does not alter the fact that the modification of one adjective in rigid order AANs (the A_y in the case of attested-order rigid order AANs) is much stronger than the other. Unlike the measures that differentiated flexible and rigid order AANs, here we see that the distance from the component N is not an indicator of the correct adjective ordering (trivially so for MULT, where attested and unattested AANs are identical).

Next, we find that for W.ADD, F.ADD and LFM, the distance from the compo-

	Measure	*t*	*sig.*	
W.ADD	cosA_x	-7.840	*	U>A
	cosA_y	7.924	*	A>U
	cosN	2.394		
	cosA_xN	-5.462	*	U>A
	cosA_yN	3.627	*	A>U
F.ADD	cosA_x	-8.418	*	U>A
	cosA_y	6.534	*	A>U
	cosN	-1.927		
	cosA_xN	-3.583	*	U>A
	cosA_yN	-2.185		
MULT	cosA_x	-5.100	*	U>A
	cosA_y	5.100	*	A>U
	cosN	0.000		
	cosA_xN	-0.598		
	cosA_yN	0.598		
LFM	cosA_x	-7.498	*	U>A
	cosA_y	7.227	*	A>U
	cosN	-2.172		
	cosA_xN	-5.792	*	U>A
	cosA_yN	0.774		
ΔPMI		-11.448	*	U>A

Table 5.4: **Attested- vs. unattested-order rigid order AANs.** *t*-normalized mean paired cosine (or ΔPMI) differences between attested (A) and unattested (U) AANs with their components. For significant differences (paired *t*-test $p<0.05$ after Bonferroni correction), last column reports whether cosines (or ΔPMI) are on average larger for A or U.

nent A_xN is a strong indicator of attested- vs. unattested-order rigid order AANs. Specifically, attested-order AANs are further from their A_xN than unattested-order AANs. This finding is in line with our predictions and follows the findings of the impact of the distance from the component adjectives.

ΔPMI, as seen in the ability to distinguish flexible vs. rigid order AANs, is the strongest indicator of correct vs wrong adjective ordering. This measure confirms that the association of one adjective (the A_y in attested-order AANs) with the head noun is indeed the most significant factor distinguishing these two classes. However, as we mentioned before, this measure has its limitations and is likely not to be entirely sufficient for future steps in modeling recursive modification.

5.5 Discussion

While AN constructions have been extensively studied within the framework of compositional distributional semantics Baroni & Zamparelli (2010); Boleda *et al.* (2012, 2013); Guevara (2010); Mitchell & Lapata (2010); Turney (2012); Vecchi *et al.* (2011), for the first time, we extended the investigation to recursively built AAN phrases.

First, we showed that composition functions applied recursively can approximate corpus-extracted AAN vectors that we know to be of high semantic quality.

Next, we looked at some properties of the same high-quality corpus-extracted AAN vectors, finding that the distinction between "flexible" AANs, where the adjective order can be flipped, and "rigid" ones, where the order is fixed, is reflected in distributional cues. These results all derive from the intuition that the most embedded adjective in a rigid AAN has a very strong effect on the distributional semantic representation of the AAN. Most compositional models were able to capture at least some of the same cues that emerged in the analysis of the corpus-extracted vectors.

Finally, similar cues were also shown to distinguish (compositional) representations of rigid AANs in the "correct" (corpus-attested) and "wrong" (unattested) orders, again pointing to the degree to which the (attested-order) closest adjective affects the overall AAN meaning as an important factor.

Comparing the composition functions, we find that the linguistically moti-

vated LFM approach has the most consistent performance across all our tests. This model significantly outperformed all others in approximating high-quality corpus-extracted AAN vectors, it provided the closest approximation to the corpus-observed patterns when distinguishing flexible and rigid AANs, and it was one of the models with the strongest cues distinguishing attested and unattested orders of rigid AANs.

From an applied point of view, a natural next step would be to use the cues we proposed as features to train a classifier to predict the preferred order of adjectives, to be tested also in cases where neither order is found in the corpus, so direct corpus evidence cannot help. For a full account of adjectival ordering, non-semantic factors should also be taken into account. As shown by the effectiveness in our experiments of PMI, which is a classic measure used to harvest idioms and other multiword expressions Church & Hanks (1990), ordering is affected by arbitrary lexicalization patterns. Metrical effects are also likely to play a role, like they do in the well-studied case of "binomials" such as *salt and pepper* Benor & Levy (2006); Copestake & Herbelot (2011). In a pilot study, we found that indeed word length (roughly quantified by number of letters) is a significant factor in predicting adjective ordering (the shorter adjective being more likely to occur first), but its effect is not nearly as strong as that of the semantic measures we considered here. In our future work, we would like to develop an order model that exploits semantic, metrical and lexicalization features jointly for maximal classification accuracy.

Adjectival ordering information could be useful in parsing: in English, it could tell whether an AANN sequence should be parsed as A[[AN]N] or A[A[NN]]; in languages with pre- and post-N adjectives, like Italian or Spanish, it could tell whether ANA sequences should be parsed as A[NA] or [AN]A. See Lazaridou *et al.* (2013) for an initial study at using measures extracted from distributional representations of compound NPs to improve bracketing in parsing. The ability to detect ordering restrictions could also help Natural Language Generation tasks Malouf (2000); Mitchell *et al.* (2011); Shaw & Hatzivassiloglou (1999), especially for the generation of unattested combinations of As and Ns.

From a theoretical point of view, we would like to extend our analysis to adjective coordination (what's the difference between *new and creative idea* and

new creative idea?). Additionally, we could go more granular, looking at whether compositional models can help us to understand why certain classes of adjectives are more likely to precede or follow others (why is size more likely to take scope over color, so that *big red car* sounds more natural than *red big car*?) or studying the behaviour of specific adjectives (can our approach capture the fact that *strong alcoholic drink* is preferable to *alcoholic strong drink* because *strong* pertains to the alcoholic properties of the drink?).

In the meantime, we hope that the results we reported here provide convincing evidence of the usefulness of compositional distributional semantics in tackling topics, such as recursive adjectival modification, that have been of traditional interest to theoretical linguists from a new perspective.

Chapter 6

Conclusions

In this thesis, I study the ability of compositional distributional semantics to model adjective modification. I present three novel studies that provide insight into the behavior of these models in the setting of adjective-noun composition, as well as an understanding of the semantic properties that motivate a number of linguistic phenomena in modification phrases. This work provides strong support for compositional distributional semantics, as it is able to generalize and capture the complex semantic intuition of natural language speakers for adjective-noun phrases, even without being able to rely on co-occurrence relations between the constituents.

In a study that explored the ability of distributional models to distinguish degrees of adjective modification (c.f. Chapter 3), we found that the relationship between the phrase vector, either corpus-observed or model-generated, and its (corpus-observed) constituent vectors significantly distinguishes literal, or intersective, modification (e.g., *white shirt*) from non-literal, or subsective, modification (e.g., *white wine*). Moreover, this research provides strong evidence for treating adjectives as matrices or functions, rather than vectors, as in composition functions like LFM and F.ADD, although simple operations on vectors such as ADD and W.ADD (for their excellent approximation to observed vectors) still account for some aspects of adjectival modification.

Beyond the new data it offers regarding the comparative ability of the different composition functions to account for different kinds of adjectival modification, the study presented here underscores the complexity of modification as a semantic

phenomenon. The role of adjectival modifiers as restrictors of descriptive content is reflected differently in distributional data than is their role in providing information about whether or when a description applies to some individual. Formal semantic models, thanks to their abstractness, are able to handle these two roles with little difficulty, but also with limited insight. Distributional models, in contrast, offer the promise of greater insight into each of these roles, but face serious challenges in handling both of them in a unified manner.

In Chapter 4, I introduce a study that attempts to detect semantic deviance in never-before-seen (or unattested) adjective-noun phrases. The main aim of this study was to propose a new challenge to the computational distributional semantics community, namely that of characterizing what happens, distributionally, when composition leads to semantically anomalous composite expressions. The hope is, on the one hand, to bring further support to the distributional approach by showing that it can be both productive and constrained; and on the other, to provide a more general characterization of the somewhat elusive notion of semantic deviance – a notion that the field of formal semantics acknowledges but might lack the right tools to model.

The results of this study provide evidence that we are able to significantly model human intuitions about the semantic acceptability of novel AN phrases using simple, unsupervised cues. In fact, we find that all indices of semantic deviance we propose significantly improve the goodness of fit in comparison to baseline psycholinguistic measures, such as string length and family size. Although all composition functions were able to model human intuition about the acceptability of novel AN phrases, we found that the W.ADD, DL and LFM functions were overall the most consistent and significant winners.

The measures and functions that model human intuition provide insight into the semantic processing and the acceptability of novel AN phrases. Above all, we find that the degree in which the head noun is modified, or distorted, from its original meaning, is the most significant indicator of deviance. This is indicated by both the cosine measure and our interpretation of the density results (supported in turn by the densification patterns). Therefore, composition functions that are able to model this effect are in fact able to approximate semantic acceptability.

While adjective-noun constructions have been extensively studied within the

framework of compositional distributional semantics Baroni & Zamparelli (2010); Boleda *et al.* (2012, 2013); Guevara (2010); Mitchell & Lapata (2010); Turney (2012); Vecchi *et al.* (2011), for the first time, we extended the investigation to recursively built AAN phrases in Chapter 5. This study showed, firstly, that composition functions applied recursively can approximate corpus-extracted AAN vectors that we know to be of high semantic quality. Next, we found that distributional cues of the same high-quality corpus-extracted AAN vectors reflect the distinction between "flexible" AANs, where the adjective order can be flipped, and "rigid" ones, where the order is fixed. Finally, similar cues were also shown to distinguish (compositional) representations of rigid AANs in the "correct" (corpus-attested) and "wrong" (unattested) orders, again pointing to the degree to which the (attested-order) closest adjective affects the overall AAN meaning as an important factor.

A number of curiosities have been addressed with these studies, and the results provide a great deal of insight into both the behavior of compositional distributional models in the setting of adjective-noun composition and the semantic properties driving certain linguistic phenomena present in such phrases. However, there are still a number of paths that have yet to be explored in the interest of this thesis. First, a model of adjective-noun semantics should also be able to handle the distinction between attributive and predicative adjective modification. In a similar setting to the experiments described here, one should be able to exploit the inherent properties of the distributional representations of composed phrases to determine the difference between *red*, which can be seen in both settings, e.g. *the red car* and *the car is red*, and adjectives such as the intensional modifier *former*, which cannot be seen in the predicative position, e.g. *the former president* but not *the president is former*.

In line with this, one future goal should be to extend the cues of semantic deviance to obtain a larger toolbox of plausibility measures. The implications of being able to generalize unattested data are extremely appealing, and although they are simple, intuitive and cost-efficient, currently the range of cues was quite limited. The degree of semantic information residing in the distributional representation of phrases should be more extensive, and with more complex measures we can exploit that information as well as these simple cues. In addition, as in

all experiments, I would be interested in exploring the effect of the typology of adjective-nouns on the results. For example, we do not distinguish cases such as *parliamentary tomato*, where the adjective does not apply to the conceptual semantic type of the noun (or at least, where it is completely undetermined which relation could bridge the two objects), from oxymorons such as *dry water*, or vacuously redundant ANs (*liquid water*) and so on.

Finally, our results also show that LFM and F.ADD in general perform better than other models. Although the LFM is the best performing model, and is in directly line with linguistic intuition, the drawback is one must learn parameters for each adjective, and is dependent on a large sample of training examples. On the other hand, F.ADD remains very attractive in principle because it generalizes across adjectives and is thus more parsimonious. However, the linguistic literature and the present results suggest that it might be useful to try a compromise between LFM and F.ADD, training one matrix for each subclass of adjectives under analysis. For example, it would be beneficial to be able to generalize the function of *maroon*, which may have few and infrequent training examples, with respect to the learned functions of *brown*, *purple*, *dark*, etc. In this case, we would not only be able to generalize distributional representations for low-frequency, even unattested, phrases, we would be able to generalize the function of low-frequency constituents, specifically adjectives.

Appendix A

A.1 Access to datasets

The dataset used for the experiments on detecting the degree of adjective modification, Section 3, is available online for download: `www.vecchi.com/eva/resources/data-emnlp2012.zip`.

The dataset of acceptable and deviant unattested AN phrases, introduced in Section 4.3, is freely available to the public and can be downloaded at `www.vecchi.com/eva/resources/vbz2011_deviant_AN_testset.txt` and `www.vecchi.com/eva/resources/vbz2011_acceptable_AN_testset.txt`.

The dataset of acceptability judgments described in Section 4.4 is currently available and downloaded at `www.vecchi.com/eva/temporary/vecchi_etal_2013_cf_judgments.csv`.

A.2 Evaluation materials

The dataset of acceptability judgments collected in the CF experiment (see Section 4.4.1) and used in these experiments is publicly available and can be downloaded from `www.evavecchi.com`. In the figures below, we show the instructions for the CF experiment as presented to the contributors (Fig. A.2.1), as well as an example of the judgment task for a set of AN pairs (Fig. A.2.2).

Figure A.2.1: Screenshot of the instructions presented to the contributors of the CF task.

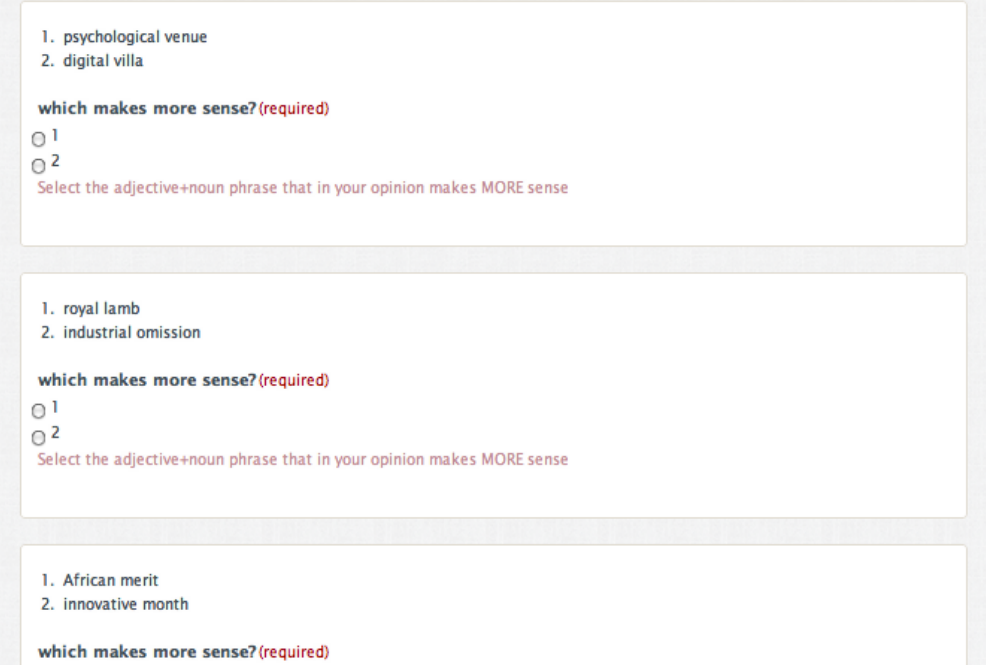

Figure A.2.2: Screenshot of a set of AN-AN pairs as presented to the contributors to be judged in the CF task.

References

Andrews, M., Vigliocco, G. & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, **116**, 463–498. 19

Andrews, S., Miller, B. & Rayner, K. (2004). Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology*, **16**, 285–311. 44

Asher, N. (2011). *Lexical Meaning in Context: A Web of Words*. Cambridge University Press. 41, 42

Baayen, R., Feldman, L. & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, **55**, 290–313. 57

Baayen, R., Davidson, D. & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, **59**, 390–412. 63, 64

Baroni, M. & Lenci, A. (2010). Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, **36**, 673–721. 4, 11, 12, 14, 26

Baroni, M. & 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 (EMNLP 2010)*, 1183–1193, Boston, MA. 6, 7, 9, 12, 16, 17, 19, 20, 23, 37, 39, 49, 52, 53, 75, 85, 86, 87, 88, 97, 102

Benor, S.B. & Levy, R. (2006). The chicken or the egg? A probabilistic analysis of english binomials. *Language*, 233–278. 98

Berlin, B. & Kay, P. (1969). Basic colour terms. *University of California Press*, **19**, 23. 24

Bertram, R. & Hyönä, J. (2003). The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long finnish compounds. *Journal of Memory and Language*, **48**, 615–634. 44, 57

Blacoe, W. & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*, 546–556, Jeju Island, Korea. 9, 37, 82

Blei, D., Ng, A. & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022. 5

Boleda, G., Vecchi, E.M., Cornudella, M. & McNally, L. (2012). First-order vs. higher-order modification in distributional semantics. In *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*, 1223–1233, Jeju Island, Korea. 12, 24, 25, 39, 76, 97, 102

Boleda, G., Baroni, M., McNally, L. & Pham, N. (2013). Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of IWCS*, 35–46, Potsdam, Germany. 33, 76, 97, 102

Bouchard, D. (2002). *Adjectives, Number and Interfaces: why languages vary*. Elsevier. 78

Bruni, E., Boleda, G., Baroni, M. & Tran, N.K. (2012). Distributional semantics in technicolor. In *Proceedings of the Association for Computational Linguistics (ACL 2012)*, 136–145, Association for Computational Linguistics. 13, 14, 24

Bullinaria, J. & Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, **39**, 510–526. 4

Callison-Burch, C. & Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 1–12, Los Angeles, CA. 54, 89

Chierchia, G. & McConnell-Ginet, S. (2000). *Meaning and grammar: An introduction to semantics*. The MIT Press. 2

Chomsky, N. (1957). *Syntactic Structures*. Mouton. 36

Church, K. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**, 22–29. 84, 98

Cinque, G. (1990). *Types of A′-Dependencies*, vol. 17 of *Linguistic Inquiry Monographs*. MIT Press, Cambridge, Mass. 77

Cinque, G. (1994). Partial N-movement in the Romance DP. In G. Cinque *et al.*, eds., *Paths Towards Universal Grammar*, 85–110, Georgetown University Press, Washington, D.C. 77

Cinque, G., ed. (2002). *Functional Structure in DP and IP - The Carthography of Syntactic Structures*, vol. 1. Oxford University Press. 76

Cinque, G. (2004). Issues in adverbial syntax. *Lingua*, **114**, 683–710. 76

Cinque, G. (2010). *The syntax of adjectives: a comparative study*. MIT Press. 76, 77

Collins, A. & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, **8**, 240–247. 2

Copestake, A. & Herbelot, A. (2011). Exciting and interesting: issues in the generation of binomials. In *Proceedings of the UCNLG+ Eval: Language Generation and Evaluation Workshop*, 45–53, Edinburgh, UK. 98

Crisma, P. (1991). Functional categories inside the noun phrase: A study on the distribution of nominal modifiers, "Tesi di Laurea", University of Venice. 76

De Jong, N., Feldman, L., Schreuder, R., Pastizzo, M. & Baayen, R. (2002). The processing and representation of dutch and english compounds: Peripheral morphological and central orthographic effects. *Brain and Language*, **81**, 555–567. 44, 57

Delfitto, D. & Zamparelli, R. (2009). *Le Strutture del Significato*. Il linguaggio umano, Il Mulino. 41

Denhière, G. & Lemaire, B. (2004). A computational model of children's semantic memory. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, 297–302, Mahway, NJ. 5

Devereux, B. & Costello, F. (2006). Modelling the interpretation and interpretation ease of noun-noun compounds using a relation space approach to compound meaning. 44

Dinu, G., Pham, N.T. & Baroni, M. (2013a). DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of the System Demonstrations of ACL 2013*, East Stroudsburg, PA. 17, 53, 87

Dinu, G., Pham, N.T. & Baroni, M. (2013b). General estimation and evaluation of compositional distributional semantic models. In *Proceedings of the ACL 2013 Workshop on Continuous Vector Space Models and their Compositionality (CVSC 2013)*, East Stroudsburg, PA. 17, 53, 87

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**, 61–74. 12

Erk, K. & Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, 897–906, Honolulu, HI, USA. 6, 7, 16, 19, 22, 85

Evert, S. (2005). *The Statistics of Word Cooccurrences*. Dissertation, Stuttgart University. 11

Fass, D. & Wilks, Y. (1983). Preference semantics, ill-formedness, and metaphor. *Computational Linguistics*, **9**, 178–187. 41

Fleiss, J.L., Cohen, J. & Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**, 323. 24

Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift fuer Philosophie un philosophische Kritik*, **100**. 5, 36, 75

Gagné, C. (2002). Lexical and relational influences on the processing of novel compounds. *Brain and Language*, **81**, 723–735. 44

Gagné, C. & Shoben, E. (2002). Priming relations in ambiguous noun-noun combinations. *Memory & cognition*, **30**, 637–646. 44

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press, Cambridge, MA. 22

Gardner, M., Rothkopf, E., Lapan, R. & Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory & cognition*, **15**, 24–28. 43

Garrette, D., Erk, K. & Mooney, R. (2011). Integrating logical representations with probabilistic information using markov logic. In *Proceedings of IWCS 2011*. 19

Giora, R. (2002). Literal vs. figurative language: Different or equal? *Journal of Pragmatics*, **34**, 487–506. 41

Golub, G., Heath, M. & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 215–223. 18

Gordon, B. (1983). Lexical access and lexical decision: Mechanisms of frequency sensitivity. *Journal of Verbal Learning and Verbal Behavior*, **22**, 24–44. 43

Grefenstette, E. & Sadrzadeh, M. (2011a). Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, Edinburgh, UK. 37, 75

Grefenstette, E. & Sadrzadeh, M. (2011b). Experimenting with transitive verbs in a DisCoCat. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. 7, 19, 22, 85

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer, Boston, MA. 4

Griffiths, T., Steyvers, M. & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review*, **114**, 211–244. 5

Guevara, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the the Association for Computational Linguistics (ACL 2010) GEMS Workshop*, 33–37, Uppsala, Sweden. 6, 16, 17, 23, 26, 37, 39, 53, 75, 85, 86, 87, 97, 102

Harris, Z.S. (1968). *Mathematical structures of language*. Wiley, New York. 4

Hasher, L. & Zacks, R. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist*, **39**, 1372. 43

Jaeger, T. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, **59**, 434–446. 63

Juhasz, B., Starr, M., Inhoff, A. & Placke, L. (2003). The effects of morphology on the processing of compound words: Evidence from naming, lexical decisions and eye fixations. *British Journal of Psychology*, **94**, 223–244. 44

Kamp, H. (1975). Two theories of adjectives. In *Formal Semantics of Natural Language*, 123–155, Cambridge University Press, Cambridge. 21

Kennedy, C. & McNally, L. (2010). Color, context, and compositionality. *Synthese*, **174**, 79–98. 22

Kintsch, W. (2001). Predication. *Cognitive Science*, **25**, 173–202. 16

Kuperman, V., Schreuder, R., Bertram, R. & Baayen, R. (2009). Reading polymorphemic dutch compounds: toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, **35**, 876. 44

Laham, D. (2000). *Automated content assessment of text using latent semantic analysis to simulate human cognition*. Dissertation, University of Colorado at Boulder. 5

Landauer, T. & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211–240. 4, 5, 12, 16

Landauer, T., Laham, D., Rehder, B. & Schreiner, M. (1997). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, 412–417. 6

Lapata, M., McDonald, S. & Keller, F. (1999). Determinants of adjective-noun plausibility. In *Proceedings of the ninth conference on EACL*, 30–36, Association for Computational Linguistics. 43

Lapata, M., Keller, F. & McDonald, S. (2001). Evaluating smoothing algorithms against plausibility judgements. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL 2001)*, 354–361, Association for Computational Linguistics. 43

Lazaridou, A., Vecchi, E.M. & Baroni, M. (2013). Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, 1908–1913, Seattle, WA. 47, 61, 98

Lee, D.D. & Seung, H.S. (2000). Algorithms for non-negative matrix factorization. In *In NIPS*, 556–562, MIT Press. 11

Lin, C.J. (2007). Projected gradient methods for Nonnegative Matrix Factorization. *Neural Computation*, **19**, 2756–2779. 84

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, **28**, 203–208. 4, 5

Malouf, R. (2000). The order of prenominal adjectives in natural language generation. In *Proceedings of the Association for Computational Linguistics (ACL 2000)*, 85–92, East Stroudsburg, PA. 98

McDonald, S. (2000). *Environmental determinants of lexical processing effort*. Dissertation, University of Edinburgh. 5

Mitchell, J. & Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of the Association for Computational Linguistics (ACL 2008)*, 236–244, Columbus, OH, USA. 6, 85

Mitchell, J. & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, **34**, 1388–1429. 3, 7, 10, 11, 13, 14, 15, 16, 17, 18, 19, 22, 37, 39, 75, 84, 85, 97, 102

Mitchell, M., Dunlop, A. & Roark, B. (2011). Semi-supervised modeling for prenominal modifier ordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011): Human Language Technologies*, 236–241, Association for Computational Linguistics, Portland, Oregon, USA. 98

Montague, R. (1970). Universal Grammar. *Theoria*, **36**, 373–398. 75

Montague, R. (1974). *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New York. 3, 5, 23

Mullaly, A., Gagne, C., Spalding, T. & Marchak, K. (2010). Examining ambiguous adjectives in adjective-noun phrases: Evidence for representation as a shared core-meaning with sense specialization. *The Mental Lexicon*, **5**, 87–114. 44

Munro, R., Bethard, S., Kuperman, V., Lai, V.T., Melnick, R., Potts, C., Schnoebelen, T. & Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 122–130, Los Angeles, CA. 54, 89

New, B., Ferrand, L., Pallier, C. & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the english lexicon project. *Psychonomic Bulletin & Review*, **13**, 45–52. 57

Nunberg, G., Sag, I. & Wasow, T. (1994). Idioms. *Language*, 491–538. 6, 37

Padó, S. & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, **33**, 161–199. 4, 12

Padó, S., Padó, U. & Erk, K. (2007). Flexible, corpus-based modelling of human plausibility judgements. In *EMNLP-CoNLL*, 400–409. 43

Parsons, T. (1970). Some problems concerning the logic of grammatical modifiers. *Synthese*, **21**, 320–324. 21

Partee, B. (1995). Lexical semantics and compositionality. *An invitation to cognitive science: Language*, **1**, 311–360. 5

Partee, B. (2004). Compositionality. In *Compositionality in Formal Semantics: Selected Papers by Barbara H. Partee*, Blackwell, Oxford. 5, 36, 75

Pollatsek, A., Hyönä, J. & Bertram, R. (2000). The role of morphological constituents in reading finnish compound words. *Journal of Experimental Psychology: Human Perception and Performance*, **26**, 820. 44

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the 9th MT Summit*, 315–322, New Orleans, LA, USA. 12

Rubenstein, H. & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, **8**, 627–633. 10, 13

Sahlgren, M. (2006). *The Word-Space Model*. Dissertation, Stockholm University. 36, 75

Sandra, D. (1990). On the representation and processing of compound words: Automatic access to constituent morphemes does not occur. *The Quarterly Journal of Experimental Psychology*, **42**, 529–567. 43

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL-SIGDAT Workshop*, Dublin, Ireland. 9

Schmidt, L., Kemp, C. & Tenenbaum, J. (2006). Nonsense and sensibility: Inferring unseen possibilities. In *Proceedings of the 27th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society*. 44

Schütze, H. (1997). *Ambiguity Resolution in Natural Language Learning*. CSLI, Stanford, CA. 4, 12, 36, 75

Scott, G.J. (2002). Stacked adjectival modification and the structure of nominal phrases. In G. Cinque, ed., *Functional Structure in DP and IP. The Carthography of Syntactic Structures*, vol. 1, Oxford University Press. 76

Shaw, J. & Hatzivassiloglou, V. (1999). Ordering among premodifiers. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, 135–143, Association for Computational Linguistics, College Park, Maryland, USA. 98

Smith, E. & Medin, D. (1981). *Categories and concepts*. Harvard University Press Cambridge, MA. 3

Socher, R., Huang, E., Pennington, J., Ng, A.Y. & Manning, C. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Advances in Neural Information Processing Systems*, **24**, 801–809. 7, 19, 82, 85

Socher, R., Huval, B., Manning, C.D. & Ng, A.Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*, 1201–1211, Edinburgh, UK. 37, 75

Sommers, F. (1971). Structural ontology. *Philosophia*, **1**, 21–42. 45

Sproat, R. & Shih, C. (1990). The cross-linguistics distribution of adjective ordering restrictions. In C. Georgopoulos & I. R., eds., *Interdisciplinary approaches to language: essays in honor of Yuki Kuroda*, 565–593, Kluver, Dordrecht. 76

Steddy, S. & Samek-Lodovici, V. (2011). On the ungrammaticality of remnant movement in the derivation of greenberg's universal 20. *Linguistic Inquiry*, **42**, 445–469. 76

Steyvers, M. & Tenenbaum, J. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, **29**, 41–78. 3

Thomason, R.H., ed. (1974). *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New York. 17, 86

Turney, P. (2012). Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, **44**, 533–585. 97, 102

Turney, P. & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, **37**, 141–188. 4, 12, 19, 36, 75

Vecchi, E.M., Baroni, M. & Zamparelli, R. (2011). (Linear) maps of the impossible: Capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, 1–9, Association for Computational Linguistics, Portland, Oregon, USA. 7, 39, 47, 55, 58, 60, 61, 76, 85, 97, 102

Vecchi, E.M., Marelli, M., Zamparelli, R. & Baroni, M. (2013a). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. In Review. 39

Vecchi, E.M., Zamparelli, R. & Baroni, M. (2013b). Studying the recursive behaviour of adjectival modification with compositional distributional semantics. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, 141–151, Seattle, WA. 12, 39

Zanzotto, F., Korkontzelos, I., Falucchi, F. & Manandhar, S. (2010). Estimating linear models for compositional distributional semantics. In *Proceedings of COLING*, 1263–1271, Beijing, China. 16, 86

Zhou, C.L., Yang, Y. & Huang, X.X. (2007). Computational mechanisms for metaphor in languages: a survey. *Journal of Computer Science and Technology*, **22**, 308–319. 41

Zwitserlood, P. (1994). The role of semantic transparency in the processing and representation of dutch compounds. *Language and Cognitive Processes*, **9**, 341–368. 43