Intensionality was only alleged: On adjective-noun composition in distributional semantics

Gemma Boleda
The University of Texas at Austin
gboleda@cs.utexas.edu

Marco Baroni
University of Trento
marco.baroni@unitn.it

Nghia The Pham
University of Trento
thenghia.pham@unitn.it

Louise McNally
Universitat Pompeu Fabra
louise.mcnally@upf.edu

Abstract

Distributional semantics has very successfully modeled semantic phenomena at the word level, and recently interest has grown in extending it to capture the meaning of phrases via semantic composition. We present experiments in adjective-noun composition which (1) show that adjectival modification can be successfully modeled with distributional semantics, (2) show that composition models inspired by the semantics of higher-order predication fare better than those that perform simple feature union or intersection, (3) contrary to what the theoretical literature might lead one to expect, do not yield a distinction between intensional and non-intensional modification, and (4) suggest that head noun polysemy and whether the adjective corresponds to a typical attribute of the noun are relevant factors in the distributional representation of adjective phrases.

1 Introduction

Distributional semantics (see Turney and Pantel, 2010, for an overview) has been very successful in modeling lexical semantic phenomena, from psycholinguistic facts such as semantic priming (McDonald and Brew, 2004) to tasks such as picking the right synonym on a TOEFL exercise (Landauer and Dumais, 1997). More recently, interest has increased in using distributional models to account not only for word meaning but also for phrase meaning, i.e. semantic composition (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Socher et al., 2012; Garrette et al., 2012).

Adjectival modification of nouns is a particularly useful and at the same time challenging testbed for different distributional models of composition, because syntactically it is very simple, while the semantic effect of the composition is very variable and potentially complex due to the frequent context dependence of the relation between the adjective and the noun (Asher, 2011, provides recent discussion). As a comparatively underexplored area of semantic theory, it is also an empirical domain where distributional models can give feedback to theoreticians about how adjectival modification works. In the formal semantic tradition, the analysis of adjectives has been largely motivated by the general entailment patterns in which they participate (Parsons, 1970; Kamp, 1975, and subsequent work). For example, if something is a *white towel*, then it is both white and a towel. This use of *white* is *intersective*: it yields an adjective-noun phrase (hereafter, AN phrase) whose denotation is the intersection of the denotations of the adjective and the noun. If someone is a *skillful surgeon*, then she is a surgeon but not necessarily skillful in general. Such adjectives are *subsective*: The denotation of the phrase is a subset of that of the noun. Finally, if someone is an *alleged murderer*, we cannot be sure that she is a murderer, and it is not even grammatical to say that she is "alleged". Intensional adjectives thus do not appear to describe attributes or relations; rather, they are almost universally modeled as higher-order properties, whereas intersective and subsective (hereafter, non-intensional) adjectives have been given both first-order and higher-order analyses.

Given these facts, we can expect that intensional adjectives will be more difficult to model computationally than non-intensional adjectives. Moreover, they raise specific issues for the increasingly popular distributional approaches to semantics. First, as intensional adjectives cannot be modeled as first-order properties, it is hard to predict what their representations might look like or what their semantic effect would be in standard distributional models of composition based on vector addition or multiplication. This is so because addition and multiplication correspond to feature combination (see Section 2 for discussion), and it is not obvious what set of distinctive distributional features an intensional adjective would contribute on a consistent basis.

In Boleda et al. (2012), we presented a first distributional semantic study of intensional adjectives. However, our study was limited in two ways. First, it compared intensional adjectives with a very narrow class of non-intensional adjectives, namely color terms; this raises doubts about the generality of our results. Second, the study had methodological weaknesses, as we did not separate training and test data, nor did we do any systematic parameter tuning prior to carrying out our experiments. This paper adresses these limitations by covering a wider variety of adjectives and using a better implementation of the composition functions, and performs several qualitative analyses on the results.

Our results confirm that high quality adjective composition is possible in distributional models: Meaningful vectors can be composed, if we take phrase vectors directly extracted from the corpus as a benchmark. In addition, we find (perhaps unsurprisingly) that models that replicate higher-order predication within a distributional approach, such as Baroni and Zamparelli (2010) and Guevara (2010), fare better than models based on vector addition or multiplication (Mitchell and Lapata, 2010). However, unlike our previous study, we find no difference in the relative success of the different composition models on intensional vs. non-intensional modification, nor in relevant aspects of the distributional representations of corpus-harvested phrases. Rather, two relevant effects involve the polysemy of the noun and the extent to which the adjective denotes a typical attribute of the entity described by the noun.

These results indicate that, in general, adjectival modification is more complex than simple feature intersection, even for adjectives like *white* or *ripe*. We therefore find tentative support for modeling adjectives as higher-order functors as a rule, despite the fact that entailment phenomena do not force such a conclusion and certain facts have even been used to argue against it (Larson, 1998, and others). The results also raise deeper and more general questions concerning the extent to which the entailment-based classification is cognitively salient, and point to the need for clarifying how polysemy and typicality intervene in the composition process and how they are to be reflected in semantic representations.

2 Composition functions in distributional semantics

Distributional semantic models represent words with vectors that record their patterns of co-occurrence with other words (or other linguistic contexts) in corpora. The raw counts are then typically transformed by reweighting and dimensionality selection or reduction operations (see Clark, 2012; Erk, 2012; Turney and Pantel, 2010, for recent surveys). Although there has always been interest in how these models could encode the meaning of phrases and larger constituents, the last few years have seen a huge increase in the number of studies devoted to *compositional* distributional semantics. We will now briefly review some of the composition methods that have been proposed and that we re-implemented here, focusing in particular on how they model AN phrases.

Mitchell and Lapata, in a set of very influential recent studies summarized in Mitchell and Lapata (2010), propose three simple and effective approaches to composition, showing that they outperform more complex models from the earlier literature. Their **weighted additive** model derives a phrase vector $\mathbf{p}$ by a weighted sum of its parts $\mathbf{u}$ and $\mathbf{v}$ (in our study, the $\mathbf{u}$ and $\mathbf{v}$ vectors to be composed will stand for adjectives and nouns, respectively):

$$\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v}$$

The **multiplicative** model proceeds by component-wise multiplication:

$$p_i = u_i v_i$$

Assuming that one of the words in the phrase acts as its "head", the **dilation** model performs composition by analyzing the head vector $\mathbf{v}$ in terms of components parallel and orthogonal to the modifier vector $\mathbf{u}$, and stretching only the parallel component by a factor λ:

$$\mathbf{p} = (\lambda - 1)(\mathbf{u} \cdot \mathbf{v})\mathbf{u} + (\mathbf{u} \cdot \mathbf{u})\mathbf{v}$$

The natural assumption, in our case, is that the noun acts as head ($\mathbf{v}$) and the adjective as modifier ($\mathbf{u}$). We experimented with the other direction as well, obtaining, unsurprisingly, worse results than those we report below for dilation with noun as head. Note that dilation can be seen as a special way to estimate the parameters of weighted addition on a phrase-by-phrase basis ($\alpha = (\lambda - 1)(\mathbf{u} \cdot \mathbf{v})$; $\beta = \mathbf{u} \cdot \mathbf{u}$).

If we interpret the components of distributional vectors as *features* characterizing the meaning of a target word, the Mitchell and Lapata models amount to essentially feature union or intersection, where the components of a phrase are those features that are active in either (union; additive model) or both (intersection; multiplicative model) the noun and/or adjective vectors. Thus, the result is "adjective-like" and/or "noun-like". Indeed, in our experiments below the nearest neighbors of phrase vectors built with these models are very often the adjective and noun components.[1] This makes intuitive sense: for example, as discussed in Boleda et al. (2012), for *white dress* feature combination makes the phrase more similar to *wedding* than to *funeral*, through the association between *white* and *wedding*. However, as formal semanticists have long observed, adjective-noun composition is often *not* a feature combination operation. Most obviously in the case of intensional adjectives, it is not correct to think of an *alleged murderer* as somebody who possesses an intersection (or union, for that matter) of features of *murderers* and features of *alleged* things.

Guevara (2010) explores the **full additive** model, an extension of the additive model where, before summing, the two n-dimensional input vectors are multiplied by two $n \times n$ weight matrices:

$$\mathbf{p} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v}$$

Unlike weighted addition and dilation, the full additive method derives the value in each component of the output vector by a weighted combination of *all* components of both input vectors, providing more flexibility. Still, a single weight matrix is used for all adjectives, which fails to capture the intuition that adjectives can modify nouns in very different ways (again, compare *white* to *alleged*).

Baroni and Zamparelli (2010) go one step further, taking the classic Fregean view of composition as function application, where certain words act as functions that take other words as input to return the semantic representation of the phrase they compose. Given that matrices encode linear functions, their **lexical function** model treats composition as the product of a matrix $\mathbf{U}$ representing the word acting as the functor and a vector $\mathbf{v}$ representing the argument word (essentially the same idea is put forth also by Coecke et al., 2010):

$$\mathbf{p} = \mathbf{U}\mathbf{v}$$

In our case, adjectives are functors and nouns arguments. Each adjective is represented by a separate matrix, thus allowing maximum flexibility in the way in which adjectives produce phrases, with the goal of capturing relevant adjectival modification phenomena beyond union and intersection.

As mentioned in the Introduction, "matrix-based" models such as the full additive and lexical function models are more similar to higher-order modification in formal semantics than feature combination models are. Thus, we expected them to perform better in modeling intensional modification, while it could be the case that for non-intensional modification feature combination models work just as well. As will be shown in Section 4, what we find is that the matrix-based model perform best across the board, and that no model finds intensional modification more difficult.

[1] Nearest neighbors are the semantic space elements having the highest cosines with the phrase of interest. These can be any of the 42K elements presented in Section 3.3: adjectives, nouns, or AN phrases.

3 Experimental setup

3.1 Semantic space

A distributional semantic space is a matrix whose rows represent target elements in terms of (functions of) their patterns of co-occurrence with contexts (columns or dimensions). Several parameters must be manually fixed or tuned to instantiate the space.

Our source corpus is given by the concatenation of the ukWaC corpus, a mid-2009 dump of the English Wikipedia and the British National Corpus,[2] for a total of about 2.8 billion tokens. The corpora have been dependency-parsed with the MALT parser (Hall, 2006), so it is straightforward to extract all cases of adjective-noun modification. We use part-of-speech-aware lemmas as our representations both for target elements and dimensions. (e.g., we distinguish between noun and verb forms of *can*).

The target elements in our semantic space are the 4K most frequent adjectives, the 8K most frequent nouns, and approximately 30K AN phrases. The phrases were composed only of adjectives and nouns in the semantic space, and were chosen as follows: a) all the phrases for the dataset that we evaluate on (see Section 3.3 below), and b) the top 10K most frequent phrases, excluding the 1,000 most frequent ones to avoid highly collocational / non-compositional phrases. The phrases were used for training purposes, and also entered in the computation of the nearest neighbors.

The dimensions of our semantic space are the top 10K most frequent content words in the corpus (nouns, adjectives, verbs and adverbs). We use a bag-of-words representation: Each target word or phrase is represented in terms of its co-occurrences with content words within the same sentence. Note that this also applies to the AN phrases: We build vectors for phrases in the same way we do for adjectives and nouns, by collecting co-occurrence counts with the dimensions of the space (Baroni and Zamparelli, 2010; Guevara, 2010). This way, we have the same type of representation for, say, *hard*, *rock*, and *hard rock*. We will call the vectors directly extracted from the corpus (as opposed to derived compositionally) **observed vectors**.

We optimized the remaining parameters of our semantic space construction on the independent task of maximizing correlation with human semantic relatedness ratings on the MEN benchmark[3] (see the references on distributional semantics at the beginning of Section 2 above for an explanation of the parameters). We found that the best model on this task was one where all dimensions where used (as opposed to removing the 50 or 300 most frequent dimensions), the co-occurrence matrix was weighted by Pointwise Mutual Information (as opposed to: no weighting, logarithm transform, Local Mutual Information), dimensionality reduction was performed by Nonnegative Matrix Factorization[4] (as opposed to: no reduction, Singular Value Decomposition), and the dimensionality of the reduced space was 350 (among values from 50 to 350 in steps of 50). The best performing model achieved very high 0.78 (Pearson) and 0.76 (Spearman) correlation scores with the MEN dataset, suggesting that we are using a high-quality semantic space.

3.2 Parameters of composition models

Except for the multiplication method, all composition models have parameters to be tuned. Following Guevara (2010) and Baroni and Zamparelli (2010), we optimize the parameters of the models by minimizing (with standard least squares regression methods) the average distance of compositionally derived vectors representing a phrase to the corresponding observed vectors extracted from the corpus (e.g., minimize the distance between the *hard rock* vector constructed by a model and the corresponding *hard rock* vector directly extracted from the corpus). There is independent evidence that such observed phrase vectors are semantically meaningful and provide a good optimization criterion. Baroni et al. (2013) report an experiment in which subjects consistently prefer the nearest neighbors of observed phrase vectors

[2] http://wacky.sslmit.unibo.it/; http://en.wikipedia.org; http://www.natcorp.ox.ac.uk/

[3] http://clic.cimec.unitn.it/~elia.bruni/MEN

[4] Unlike the more commonly used Singular Value Decomposition method, Nonnegative Matrix Factorization produces reduced dimensions that have no negative values, and are not fully dense.

I	alleged	former	future	hypothetical	impossible	likely	mere	mock
N	loose	wide	white	naive	severe	hard	intelligent	ripe
I	necessary	past	possible	potential	presumed	probable	putative	theoretical
N	modern	black	free	safe	vile	nasty	meagre	stable

Table 1: Evaluated adjectives. Intensional (I) and non-intensional (N) adjectives are paired by frequency.

over challenging foils. Turney (2012) shows how the observed vectors outperform any compositionally-derived model in a paraphrasing task. Grefenstette et al. (2013) reach state-of-the-art performance on widely used sentence similarity test sets with composition functions optimized on the observed vectors (see also Baroni et al., 2012; Baroni and Zamparelli, 2010; Boleda et al., 2012).

Since we use the same criterion to evaluate the quality of the models, we are careful to separate training phrases from those used for evaluation (we introduce the test set in the next section). The weighted additive, dilation and full-additive models require one single set of parameters for all adjectives, and we thus use the top 10K most frequent phrases in our semantic space (excluding test items) for training. For the lexical function model, we need to train a separate weight matrix for each adjective. We do this by using as training data, for each adjective, all phrase vectors in our semantic space that contain the adjective and are not in the test set. These range between 52 (*ripe*) and 1,789 (*free*). For weighted additive, we find that the best weights are $\alpha = 0.48$, $\beta = 0.61$, giving only marginally more weight to the noun. For dilation, $\lambda = 1.69$.

3.3 Evaluation set

We evaluate the models on a set of 16 intensional adjectives and a set of 16 non-intensional adjectives, paired according to frequency (see Table 1). The intensional adjectives were chosen starting from the candidate list elaborated for Boleda et al. (2012), with two modifications. First, the frequency criteria were altered, allowing the addition of seven more adjectives (e.g., *alleged* and *putative*). Second, we removed adjectives that can be used predicatively with the same intensional interpretation despite having been claimed to meet the entailment test for intensionality; this excludes, e.g., *false* (cp. *This sentence is false*). Adjectives that have a non-intensional predicative use alongside a non-predicative intensional one, e.g., *possible* (cp. *The possible winner* vs. ??*The winner was possible*, but *Peace was possible*) were left in, despite the potential for introducing some noise. The non-intensional adjectives were chosen by generating, for each intensional adjective, a list of the 20 adjectives closest in frequency and taking from that list the closest match in frequency that was morphologically simple (excluding, e.g., *unexpected* or *photographic*) and unambiguously an adjective (excluding, e.g., *super* and *many*).

We used all the AN phrases in the corpus with a frequency of at least 20 for all adjectives except the underrepresented ones (*nasty*, *mock*, *probable*, *hypothetical*, *impossible*, *naive*, *presumed*, *putative*, *vile*, *meagre*, *ripe*), for which we selected at most 200 phrases, taking phrases down to a frequency of 5 if needed. For each adjective, we randomly sampled 50 phrases for testing (total: 1,600).[5] The rest were used for training, as described above. The results and analyses in sections 4 and 5 concern the test data only.

4 Results

4.1 Overall results

Table 2 (first column) shows the results of the main evaluation: Average cosine of phrase vectors produced by composition models (henceforth, **predicted vectors**) with the corresponding observed vectors. As a baseline (last row in the table), we take doing no composition at all, that is, taking as the predicted vector simply the noun vector. This is a hard baseline: Since AN phrases in general denote a set closely related to the noun, noun-phrase similarities are relatively high.

[5] The dataset is available from the first author's webpage.

Model	**Global**	**Intensional**	**Non-intensional**	**NN=A**	**NN=N**
observed	-	-	-	*8.2*	*3.3*
lexical function	**0.60**±0.11	**0.60**±0.10	**0.60**±0.10	0.9	0.6
full additive	0.52±0.13	0.52±0.13	0.51±0.12	10.0	4.8
weighted additive	0.48±0.14	0.48±0.14	0.48±0.14	23.2	13.3
dilation	0.42±0.18	0.42±0.17	0.42±0.17	31.0	11.6
multiplicative	0.32±0.21	0.32±0.20	0.32±0.20	29.9	16.6
noun only	*0.40±0.18*	*0.40±0.17*	*0.40±0.17*	-	-

Table 2: Predicted-to-observed vector cosines for each model (mean ± standard deviation), globally and by adjective type. The last two columns show the average % of the 50 nearest neighbors that are adjectives (NN=A) and nouns (NN=N), as opposed to AN phrases.

The global results show that the matrix-based models (lexical function and full additive) clearly outperform the models based on a simple combination of the component vectors, and the lexical function model ranks best, with a high cosine score of 0.6.[6] It is also robust, as it exhibits the lowest standard deviation (0.11). The models that are based on some form of weighted addition[7] score in the middle, above the baseline but clearly below matrix-based models. Contrary to Mitchell and Lapata's results, where often multiplicative is the best performing model, multiplication in our experiments performs worst, and actually below the noun-only baseline. Moreover, the multiplicative model has the highest standard deviation (0.21), so it is the least robust model. This matches informal qualitative analysis of the nearest neighbors: The multiplicative model does very well on some phrases, and very poorly on others. Given the aggressive feature intersection that multiplication performs (zeroing out dimensions with no shared counts, inflating the values of shared dimensions), our results suggest that it is in general better to perform a "smoothed" union as in weighted addition. We leave it to further work to compare our results and task with Mitchell and Lapata's.

The table (columns *Intensional*, *Non-intensional*) also shows that, contrary to expectation, no model finds intensional modification more difficult, or indeed any difference between the two types of modification: The mean predicted-to-observed cosines for the two types of phrases are the same. This holds for both matrix-based and feature-combination-based models. For further discussion, see Section 5.

The last two columns of Table 2 show the average percentage of adjectives and nouns, respectively, among the 50 nearest neighbors of the phrase vectors. Observed phrases have few such single word neighbors (8.2% and 1.6% on average). We observe the same pattern as with the global evaluation: Matrix-based models also have low proportions of single word neighbors, thus corresponding more closely to the observed data,[8] while the other models exhibit a relatively high proportion of such neighbors. Single word neighbors are not always bad (e.g., the weighted additive model proposes *dolphin* for *white whale*), but their high proportion suggests that feature combination models often produce more general and therefore less related nearest neighbors. This was confirmed in a small qualitative analysis of nearest neighbors for the weighted additive model.

To sum up, the superior results of matrix-based models across the board suggest that adjectival modification is not about switching features on and off, but rather about a more complex type of transformation. Indeed, our results suggest that this is so not only for intensional adjectives, which have traditionally already been treated as higher-order predicates, but also for adjectives like *white*, *hard*, or *ripe*, whose analysis has been more controversial. If this is so, then it is not so surprising that in general the models do not find intensional adjectives any more difficult to model.

[6] Despite the large standard deviations, even the smallest difference between the models is highly significant, as is the smallest difference in the table: dilation vs. baseline (noun only), paired t-test, t = 38.2, df = 1599, p < 2.2e-16, mean of differences = 0.02.

[7] That dilation is essentially another way to estimate weighted addition, as discussed in section 2, is empirically confirmed by the fact that the correlation between the predicted-to-observed cosines for weighted additive and dilation is 0.9.

[8] In fact, the lexical function model is a bit extreme, producing almost no adjective and noun nearest neighbors.

Indeed, once an adjective is composed with a noun, the result is something that is not merely the sum of its parts. We associate with *black voter* something much more specific than merely *a voter that is black*, for instance, in the US, strong connotations of likely political inclinations. In this respect, an adjective does not just help to pick out a subset of the noun's denotation; it enriches the description contributed by the noun. This is in line with observations in the cognitive science literature on concept combination, essentially a counterpart of semantic composition. Murphy (2002, 453-453) discusses the case of *dog magazine* (with a noun modifier, but the same point holds for adjectives), arguing that its meaning is not just *magazine about dogs*: People "can infer other properties of this concept. A dog magazine probably is directed toward dog owners and breeders;. . . unlike many other magazines, it probably does not contain holiday recipes, weight-loss plans. . . Importantly, these kinds of properties. . . are not themselves properties of the concepts of dog or magazine but arise through the interaction of the two."

4.2 Comparing the quality of predicted and observed vectors

We have used observed data for phrases both to train and tune our models and to evaluate the results. If we can work with the observed data, what do we need composition for? Due to Zipf's Law, there is only a limited amount of phrases for which we can have enough data to build a meaningful representation. Perfectly plausible modifiers of nouns may never be observed in actual corpora. Thus, we need a way to combine semantic representations for words, and this is partly what drives the research on composition in distributional semantics. It is natural to hypothesize that, for rare phrases, predicted vectors will actually be more useful than observed vectors. We carried out a pilot study that supports this hypothesis.

A native speaker of English and linguist evaluated the quality of the nearest neighbors of frequent versus (relatively) rare phrases, comparing the lexical function model and the observed data. As frequent phrases, we took the top 100 most frequent phrases in the semantic space. As rare phrases, the 95 phrases with corpus frequency 20-21. The task of the judge was to choose, for a given target phrase, which of two randomly ordered nearest neighbors was more semantically related to it (we found, in earlier studies, that this type of choice is easier than assigning absolute scores to separate items). For instance, the judge had to choose whether *modern study* or *general introduction* was a semantically closer neighbor to *modern textbook*. The items were two nearest neighbors with the same rank, where the rank was randomly picked from 2-10 (the top nearest neighbor was excluded because it is trivially always the target phrase for observed vectors). The judge obviously did not know which model generated which nearest neighbor.

The results indicate that observed vectors yield better nearest neighbors for frequent phrases, as they were chosen 60% of the times (but note that the lexical function also fared well, since its nearest neighbors were preferred in 40% of the cases). However, for rare phrases we find the inverse pattern: The lexical function neighbor is preferred in 59% of the cases. For instance, the lexical function produces *nasty cold* for *nasty cough*, which was preferred to the observed nearest neighbor *medical attention*. This suggests that the composed vectors offer a better representation of rare phrases, and in tasks that depend on such phrases, they should yield better results than the observed ones.

5 Analysis

As mentioned in the Introduction, in Boleda et al. (2012) we found differences between intensional adjectives and color adjectives. We attributed these differences to the type of modification, intensional or not. We failed to to replicate these results here, with a wider range of adjectives.

Figure 5 shows the cosine distribution of the measures used in our previous work (compare to Figure 1 in Boleda et al., 2012), namely the cosines between the observed vectors for adjectives, nouns, and the corresponding phrase vectors for each AN phrase.[9] The figure shows that, contrary to expectation

[9] Each boxplot represents the distribution of cosine values across the relevant vector pair comparisons. The horizontal lines in the rectangles mark the first quartile, median, and third quartile, respectively. Larger rectangles correspond to a more widely spread distribution, and their (a)symmetry mirrors the (a)symmetry of the distribution. The lines above and below each 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.

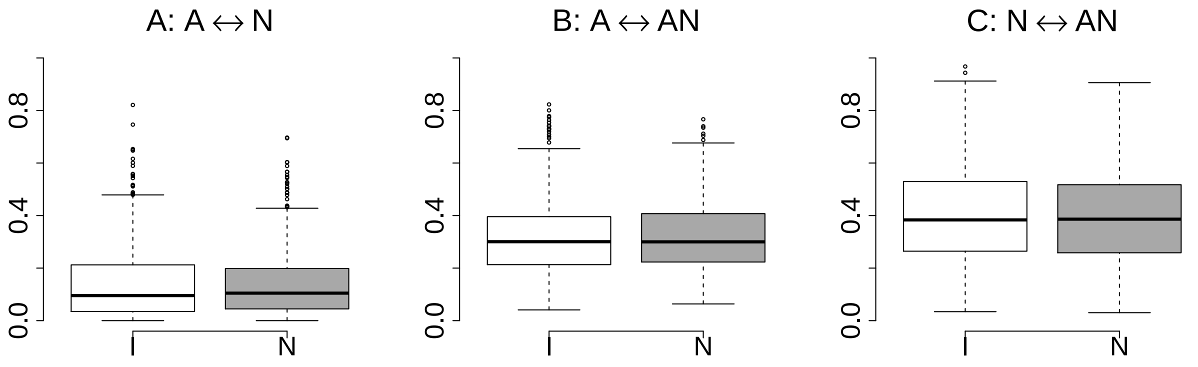

Figure 1: Distribution of cosines for observed vectors, by adjective type (intensional, I, or non-intensional, N). From left to right, adjective vs. noun, adjective vs. phrase, and noun vs. phrase cosines.

	Monosemous	**Polysemous**
I	*alleged accomplice*, *former surname*, *necessary competence*	*mock charge*, *putative point*, *past range*
N	*modern aircraft*, *severe hypertension*, *wide disparity*	*nasty review*, *ripe shock*, *meagre part*
	Typical	**Nontypical**
I	*former mayor*, *likely threat*, *alleged killer*	*former retreat*, *likely base*, *alleged fact*
N	*severe pain*, *free download*, *wide perspective*	*severe budget*, *free attention*, *wide detail*

Table 3: Examples of adjective-noun phrases for the two factors analyzed (polysemy of the head noun, typicality of the attribute) by adjective type: I(ntensional), N(on-intensional). See text for details.

and the previous results, in the observed data there is absolutely no difference in these measures between intensional and non-intensional modification: The distributions overlap completely. In a preliminary study, we paired phrases on the basis of the noun (e.g. *former bassist-male bassist*) instead of on the basis of the adjective as in the present experiments. With that design, too, we obtained no difference between the two types of phrases. We therefore take this to be a robust negative result, which suggests that the differences observed in our previous work were due to our having chosen a very narrow set of adjectives (color terms) for comparison to the intensional adjectives.

This result is surprising insofar as intensional and non-intensional adjectives have often been assumed to denote very different types of properties. One possibility is that the tools we are using are not the right ones: Perhaps using bags-of-words as the dimensions cannot capture the differences, or perhaps these differences are not apparent in the cosines between phrase and adjective/noun vectors. However, these results may also mean that all kinds of adjectival modification share properties that have gone unappreciated.

If the type of modification does not explain the differences in the observed data, what does? An analysis reveals two relevant factors. The first one is the polysemy of the head noun. We find that, the more polysemous a noun is, the less similar its vector is to the corresponding phrase vector. It is plausible that modifying a noun has a larger impact when the noun is polysemous, as the adjective narrows down the meaning of the noun; indeed, adjectives have been independently shown to be powerful word sense disambiguators of nouns (Justeson and Katz, 1995). In distributional terms, the adjective notably “shifts” the vector of polysemous nouns, but for monosemous nouns there is just not much shifting room.

This is reasonable but unsurprising; what is more worthy of attention is that this effect is invariant to adjective type. Both non-intensional and intensional adjectives have meaning modulating power, as shown in Table 3. For example, *ripe* selects for the sense of *shock* that has to do with a pile of sheaves of grain or corn. Similarly, *past* is incompatible with physical senses of *range* such as that referring to mountains or a cooking appliance.

The second effect that we find is that, the more typical the attribute described by an adjective is for the sort of thing the noun denotes, the closer the phrase vector is to both its adjective and its noun vector components. This can be explained along similar lines as the first factor: A ripe raspberry is probably more like other raspberries than, say, a humongous raspberry is. Similarly, a ripe raspberry is more like most other ripe things than a ripe condition is. Therefore, the effect of the adjective on the noun is larger if it does not describe a typical attribute of whatever the noun describes. The difference is mirrored in the contexts in which the phrases appear, which leads to larger differences in their vector representations.[10]

Interestingly, we find that typicality is also invariant across adjective type, as the examples in Table 3 show. Intensional adjectives do seem to describe typical attributes of some nouns. For example, nouns like *mayor* arguably have a temporal component to their semantics (see, e.g., Musan, 1995), the meaning of *threat* involves future intention and it is thus inherently modal, and it is culturally highly relevant whether a description like *killer* holds of a particular individual or not. Note also that typicality is not a matter of the specific adjective, but of the combination of the adjective and the noun, as illustrated by the fact that the same adjectives appear in both columns of the table: *Wide* arguably corresponds to a typical attribute of perspectives, but not of details.

The interpretation just presented is supported by a statistical analysis of the data. We estimated polysemy using the number of synsets in which a given noun appears in WordNet,[11] and typicality using an association measure, Local Mutual Information (Evert, 2005).[12] When fitting a mixed-effects model to the observed data with adjective as random effect, we find that intensionality plays no significant role in predicting the cosines between observed vectors (neither adjective vs. phrase nor noun vs. phrase cosines). Polysemy has a strong negative effect on noun vs. phrase cosines (and no effect on adjective vs. phrase cosines). Typicality has a strong positive effect on both adjective-phrase and noun-phrase cosines. We also find that these factors (but not intensionality) play a role in the difficulty of modeling a given AN phrase, since they are also highly significant (in the same directions) in predicting observed-to-predicted cosines for the lexical function model.

To sum up, in this section we have shown that there are semantic effects that are potentially relevant to adjectival modification and cut across the intensionality range, and that distributional representations of words and phrases capture such semantic effects. Thus, the analysis also provides support for the use of distributional representations for phrases.

6 Conclusion

In this paper we have tackled the computational modeling of adjective-noun composition. We have shown that adjective modification can be successfully modeled with distributional semantics, both in terms of approximating the actual distribution of phrases in corpora and in terms of the quality of the nearest neighbors they produce. We have also shown that composition models inspired in higher-order predication fare better than those that essentially intersect or combine features. Finally, contrary to what the theoretical linguistics literature might lead one to expect, we did not find a difference between intensional and non-intensional modifers in the distributional representation of phrases, nor did we find that composition functions have a harder time with intensional modification. Together, these results suggest that adjective-noun composition rarely corresponds to a simple combination of attributes of the noun and

[10] A similar explanation is provided in Boleda et al. (2012) to explain the difference between intersective and subsective uses of color terms. Here we generalize it.

[11] `http://wordnet.princeton.edu/`

[12] An association measure is not all there is to typicality; for instance, multi-word expressions like *black hole* will score high on LMI despite *black* not describing a typical attribute of holes. However, we find it a reasonable approximation because typical attributes can be expected to score higher than nontypical ones, an expectation that receives support from qualitative exploration of the data. We leave it to future work to identify alternative sources of information about typicality, such as the WordNet-based adjectival attributes in Hartung and Frank (2011).

the modifier (in line with research in cognitive science), but rather that adjectives denote functions that operate on nouns to yield something that is more than the sum of its parts. Thus, at least when used as modifiers, they denote properties of properties, rather than properties of entities.

The results of our study also indicate that intensional adjectives share a significant number of properties with non-intensional adjectives. We are of course not claiming that there are no differences between the two: For instance, there are clearly relevant semantic differences that are mirrored in the syntax. Rather, we claim that the almost exclusive focus on entailment relations in the formal semantic tradition has obscured factors that are potentially relevant, and that cut across the intensionality parameter. These are related to graded phenomena such as the polysemy of the head noun or the typicality of the attribute contributed by the adjective. We hope that our results promote closer scrutiny of these factors by theoretical semanticists, and ultimately a more complete understanding of the semantics of modification.

Acknowledgements

We thank Miquel Cornudella for help in constructing the dataset. We acknowledge the support of Spanish MICINN grant FFI2010-09464-E (McNally, Boleda), the ICREA Foundation (McNally), Catalan AGAUR grant 2010BP-A00070, MICINN grant TIN2009-14715-C04-04, EU grant PASCAL2, FP7-ICT-216886, the DARPA DEFT program under AFRL grant FA8750-13-2-0026 (Boleda) and the ERC under the 2011 Starting Independent Research Grant 283554 to the COMPOSES project (Baroni, Pham). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of DARPA, AFRL or the US government.

References

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

Baroni, M., R. Bernardi, N.-Q. Do, and C.-C. Shan (2012). Entailment above the word level in distributional semantics. In *Proceedings of EACL*, Avignon, France, pp. 23–32.

Baroni, M., R. Bernardi, and R. Zamparelli (2013). Frege in space: A program for compositional distributional semantics. Submitted, draft at `http://clic.cimec.unitn.it/composes`.

Baroni, M. and R. Zamparelli (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, Boston, MA, pp. 1183–1193.

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

Clark, S. (2012). Vector space models of lexical meaning. In S. Lappin and C. Fox (Eds.), *Handbook of Contemporary Semantics, 2nd edition*. Malden, MA: Blackwell. In press.

Coecke, B., M. Sadrzadeh, and S. Clark (2010). Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis 36*, 345–384.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*. In press.

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

Garrette, D., K. Erk, and R. Mooney (2012). A formal approach to linking logical form and vector-space lexical semantics. In H. Bunt, J. Bos, and S. Pulman (Eds.), *Computing Meaning, Vol. 4*. In press.

Grefenstette, E., G. Dinu, Y.-Z. Zhang, M. Sadrzadeh, and M. Baroni (2013). Multi-step regression learning for compositional distributional semantics. In *Proceedings of IWCS*, Potsdam, Germany. In press.

Guevara, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the ACL GEMS Workshop*, Uppsala, Sweden, pp. 33–37.

Hall, J. (2006). *MaltParser: An Architecture for Labeled Inductive Dependency Parsing*. Licentiate thesis, Växjö University, Växjö, Sweden.

Hartung, M. and A. Frank (2011). Exploring supervised lda models for assigning attributes to adjective-noun phrases. In *Proceedings of EMNLP 2011*, Stroudsburg, PA, USA, pp. 540–551.

Justeson, J. S. and S. M. Katz (1995). Principled disambiguation: Discriminating adjective senses with modified nouns. *Computational Linguistics 21*(1), 1–27.

Kamp, J. A. W. (1975). Two theories about adjectives. In E. L. Keenan (Ed.), *Formal Semantics of Natural Language*, pp. 123–155. Cambridge: Cambridge University Press.

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

Larson, R. K. (1998). Events and modification in nominals. In *Proceedings of SALT*, Ithaca, NY.

McDonald, S. and C. Brew (2004). A distributional model of semantic context effects in lexical processing. In *Proceedings of ACL*, pp. 17–24.

Mitchell, J. and M. Lapata (2010). Composition in distributional models of semantics. *Cognitive Science 34*(8), 1388–1429.

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge, MA (etc.): The MIT Press.

Musan, R. (1995). *On the temporal interpretation of noun phrases*. Ph. D. thesis, MIT.

Parsons, T. (1970). Some problems concerning the logic of grammatical modifiers. *Synthese 21*(3-4), 320–324.

Socher, R., B. Huval, C. Manning, and A. Ng (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP*, Jeju Island, Korea, pp. 1201–1211.

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

Turney, P. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research 37*, 141–188.