Studying the recursive behaviour of adjectival modification with compositional distributional semantics

Eva Maria Vecchi and **Roberto Zamparelli** and **Marco Baroni**
Center for Mind/Brain Sciences (University of Trento, Italy)
(evamaria.vecchi|roberto.zamparelli|marco.baroni)@unitn.it

Abstract

In this study, we use compositional distributional semantic methods to investigate restrictions in adjective ordering. Specifically, we focus on properties distinguishing Adjective-Adjective-Noun phrases in which there is flexibility in the adjective ordering from those bound to a rigid order. We explore a number of measures extracted from the distributional representation of AAN phrases which may indicate a word order restriction. We find that we are able to distinguish the relevant classes and the correct order based primarily on the degree of modification of the adjectives. Our results offer fresh insight into the semantic properties that determine adjective ordering, building a bridge between syntax and distributional semantics.

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 and 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 and Zamparelli, 2010; Guevara, 2010; Mitchell and Lapata, 2010; Grefenstette and Sadrzadeh, 2011; 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; Boleda et al., 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 (Sproat and Shih, 1990; Crisma, 1991; Scott, 2002), and have become one of the key in-

gredients 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 and 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, approximate 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 the correct and wrong orders.

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.

2 Materials and methods

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, 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 and 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).

Source corpus We use as our source corpus the concatenation of the Web-derived ukWaC corpus, a mid-2009 dump of the English Wikipedia and the British National Corpus[1]. The corpus has been tokenized, POS-tagged and lemmatized with the TreeTagger (Schmid, 1995), and it contains about 2.8 billion tokens. We extract all statistics at the lemma level, meaning that we consider only the canonical form of each word ignoring inflectional information, such as pluralization and verb inflection.

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

[1] `http://wacky.sslmit.unibo.it`, `http://en.wikipedia.org`, `http://www.natcorp.ox.ac.uk`

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 ($20<f<200$)[2] 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 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.

[2]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.

Model	ρ	*M&L*
CORP	0.41	0.43
W.ADD	0.41	0.44
F.ADD	0.40	–
MULT	0.33	0.46
LFM	0.40	–

Table 1: Correlation scores (Spearman's ρ, all significant at $p<0.001$) between cosines of corpus-extracted or model-generated AN vectors and phrase similarity ratings collected in Mitchell and Lapata (2010), as well as best reported results from Mitchell & Lapata (M&L).

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 and 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-order rigid order AANs, and then mapped onto the 300-dimension NMF-reduced semantic space. As a sanity check, the first row of Table 1 reports the correlation between the AN phrase similarity ratings collected in Mitchell and 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 2.4.

2.2 Composition models

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 and 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 $p_i = u_i v_i$.[3] Given an $A_x A_y N$ phrase, this model extends naturally to the recursive setting of this experiment, as seen in Equation (1).

$$\vec{p} = \vec{a}_x \odot \vec{a}_y \odot \vec{n} \quad (1)$$

This composition method is order-insensitive, the formula above corresponding to the representation of both $A_x A_y N$ and $A_y A_x N$.

In the weighted additive model (W.ADD), we obtain the composed vector as a weighted sum of the two component vectors: $\vec{p} = \alpha\vec{u} + \beta\vec{v}$, where α and β are scalars. Again, we can easily apply this function recursively, as in Equation (2).

$$\vec{p} = \alpha\vec{a}_x + \beta(\alpha\vec{a}_y + \beta\vec{n}) = \alpha\vec{a}_x + \alpha\beta\vec{a}_y + \beta^2\vec{n} \quad (2)$$

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: $\vec{p} = \mathbf{W}_1\vec{u} + \mathbf{W}_2\vec{v}$. Similarly to the W.ADD model, Equation (3) describes how we apply this function recursively.

$$\begin{aligned}\vec{p} &= \mathbf{W}_1\vec{a}_x + \mathbf{W}_2(\mathbf{W}_1\vec{a}_y + \mathbf{W}_2\vec{n}) \quad (3)\\ &= \mathbf{W}_1\vec{a}_x + \mathbf{W}_2\mathbf{W}_1\vec{a}_y + \mathbf{W}_2^2\vec{n}\end{aligned}$$

Finally, we consider the lexical function model (LFM), first introduced in Baroni and 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: $\vec{p} = \mathbf{U}\vec{v}$. Given a weight matrix, $\mathbf{A}$, for each adjective in the phrase, we apply the functions in sequence recursively as shown in Equation (4).

$$\vec{p} = \mathbf{A}_x(\mathbf{A}_y\vec{n}) \quad (4)$$

Composition model estimation Parameters for W.ADD, F.ADD and LFM were estimated following the strategy proposed by Guevara (2010) and Baroni and 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. All parameter estimations and phrase compositions were implemented using the DISSECT toolkit[4] (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 1 reports the results attained by our model implementations on the Mitchell and Lapata AN similarity data set.

2.3 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 *dif-*

[3] We conjecture that the different performance of our multiplicative model and M&L's (cf. Table 1) is due to the fact that we use log-transformed pPMI scores, making their multiplicative model more akin to our additive approach.

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

ferent, 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 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).[5] 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] 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 and Zamparelli, 2010).

2.4 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 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[6] (Callison-Burch and 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[7]. 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 2 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

[6] http://www.crowdflower.com

[7] The top 3 neighbors included adjectives, nouns, ANs and AANs. The preference for ANs and AANs, as seen in Table 2, is likely a result of the dominance of those elements in the semantic space (c.f. Section 2.1).

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

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 2. 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 meaning.

3 Results

3.1 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 3). 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.

	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 3: 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.

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.

3.2 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 4 reports the t-normalized difference between flexible order and rigid order mean cosines for the corpus-extracted vectors.

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

Table 4: **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.

These results show, in accordance with our considerations in Section 2.3 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).[8] 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.

[8]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.

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 distinguish the two adjective orders.

The results of the composition models (cf. Table 4) 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 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 2, 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 3). 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.

3.3 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.

	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: **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.

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). 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 component 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.

4 Conclusion

While AN constructions have been extensively studied within the framework of compositional distributional semantics (Baroni and Zamparelli, 2010; Boleda et al., 2012; Boleda et al., 2013; Guevara, 2010; Mitchell and 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 motivated 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 and 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 and Levy, 2006; Copestake and 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. The ability to detect ordering restrictions could also help Natural Language Generation tasks (Malouf, 2000), 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.

Acknowledgments

We would like to thank the anonymous reviewers, Fabio Massimo Zanzotto, Yao-Zhong Zhang and the members of the COMPOSES team. This research was supported by the ERC 2011 Starting Independent Research Grant n. 283554 (COMPOSES).

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, pages 1183–1193, Boston, MA.

Sarah Bunin Benor and Roger Levy. 2006. The chicken or the egg? A probabilistic analysis of english binomials. *Language*, pages 233–278.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*, pages 546–556, Jeju Island, Korea.

Gemma Boleda, Eva Maria Vecchi, Miquel Cornudella, and Louise McNally. 2012. First-order vs. higher-order modification in distributional semantics. In *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*, pages 1223–1233, Jeju Island, Korea.

Gemma Boleda, Marco Baroni, Louise McNally, and Nghia Pham. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of IWCS*, pages 35–46, Potsdam, Germany.

Chris Callison-Burch and Mark Dredze. 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*, pages 1–12, Los Angeles, CA.

Kenneth Church and Peter Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Guglielmo Cinque, editor. 2002. *Functional Structure in DP and IP - The Carthography of Syntactic Structures*, volume 1. Oxford University Press.

Guglielmo Cinque. 2004. Issues in adverbial syntax. *Lingua*, 114:683–710.

Guglielmo Cinque. 2010. *The syntax of adjectives: a comparative study*. MIT Press.

Ann Copestake and Aurélie Herbelot. 2011. Exciting and interesting: issues in the generation of binomials. In *Proceedings of the UCNLG+ Eval: Language Generation and Evaluation Workshop*, pages 45–53, Edinburgh, UK.

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

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013a. DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of the System Demonstrations of ACL 2013*, East Stroudsburg, PA.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 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.

Gottlob Frege. 1892. Über sinn und bedeutung. *Zeitschrift fuer Philosophie un philosophische Kritik*, 100.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of EMNLP*, Edinburgh, UK.

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

Chih-Jen Lin. 2007. Projected gradient methods for Nonnegative Matrix Factorization. *Neural Computation*, 19(10):2756–2779.

Robert Malouf. 2000. The order of prenominal adjectives in natural language generation. In *Proceedings of ACL*, pages 85–92, East Stroudsburg, PA.

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

Richard Montague. 1970. Universal Grammar. *Theoria*, 36:373–398.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 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*, pages 122–130, Los Angeles, CA.

Barbara Partee. 2004. Compositionality. In *Compositionality in Formal Semantics: Selected Papers by Barbara H. Partee*. Blackwell, Oxford.

Magnus Sahlgren. 2006. *The Word-Space Model*. Dissertation, Stockholm University.

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

Hinrich Schütze. 1997. *Ambiguity Resolution in Natural Language Learning*. CSLI, Stanford, CA.

Gary-John Scott. 2002. Stacked adjectival modification and the structure of nominal phrases. In Guglielmo Cinque, editor, *Functional Structure in DP and IP. The Carthography of Syntactic Structures*, volume 1. Oxford University Press.

Richard Socher, E.H. Huang, J. Pennington, Andrew Y. Ng, and C.D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Advances in Neural Information Processing Systems*, 24:801–809.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP*, pages 1201–1211, Edinburgh, UK.

Richard Sproat and Chilin Shih. 1990. The cross-linguistics distribution of adjective ordering restrictions. In C. Georgopoulos and Ishihara R., editors, *Interdisciplinary approaches to language: essays in honor of Yuki Kuroda*, pages 565–593. Kluver, Dordrecht.

Sam Steddy and Vieri Samek-Lodovici. 2011. On the ungrammaticality of remnant movement in the derivation of greenberg's universal 20. *Linguistic Inquiry*, 42(3):445–469.

Richmond H. Thomason, editor. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New York.

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

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

Eva Maria Vecchi, Marco Baroni, and Roberto Zamparelli. 2011. (Linear) maps of the impossible: Capturing semantic anomalies in distributional space. In *Proceedings of the ACL Workshop on Distributional Semantics and Compositionality*, pages 1–9, Portland, OR.

Fabio Zanzotto, Ioannis Korkontzelos, Francesca Falucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of COLING*, pages 1263–1271, Beijing, China.