Determining the Compositionality of Noun-Adjective Pairs with Lexical Variants and Distributional Semantics

Marco S. G. Senaldi *
Scuola Normale Superiore di Pisa

Gianluca E. Lebani **
Università di Pisa

Alessandro Lenci †
Università di Pisa

In this work we tested whether a series of compositionality indices that compute the distributional similarity between the vector of a given expression and the vectors of its lexical variants can effectively tell apart idiomatic and more compositional expressions in a set of 13 idiomatic and 13 non-idiomatic Italian target noun-adjective constructions. The lexical variants were obtained by replacing the components of the original expressions with semantically related words automatically extracted from Distributional Semantic Models or manually derived from Italian MultiWordnet. Indices based on the Mean or the Centroid cosine similarity between the target and the variant vectors performed comparably or better than the addition-based measure traditionally reported in the distributional literature on compositionality.

1. Introduction

When an adjective combines with a noun, the semantics of the resulting expression does not always reflect a straightforward combination of their meanings. While a sentence like *John has bought a white car* entails *John has bought something white* and *John has bought a car*, saying that *John is a skilled optician* for sure entails he's an *optician*, but not necessarily that he's *skilled* in general as a person. Moving further on, if John is an *alleged murderer*, we are not even sure he's a murderer at all and it's not even grammatical to say **John is alleged*. Finally, if one were to utter *I thought John was the murderer, but actually he was just a red herring*, they wouldn't be claiming John is either *red* or a *herring*, but they would be just figuratively asserting that John has taken their attention away from the real murderer. All these different entailment patterns exhibited by *white car*, *skilled optician*, *alleged murderer* and *red herring* show the complexity and variability of the compositionality of adjective-noun (AN henceforth) pairs, i.e. the extent to which the meaning of a phrase as a whole is a function of the meanings of its components and of the syntactic relationship that links them (Partee 1995).

* Laboratorio di Linguistica "G. Nencioni", Scuola Normale Superiore - Piazza dei Cavalieri 7, I-56126 Pisa, Italy. E-mail: marco.senaldi@sns.it
** CoLing Lab, Department of Philology, Literature and Linguistics - Via S. Maria 36, I-56126 Pisa, Italy. E-mail: gianluca.lebani@for.unipi.it
† CoLing Lab, Department of Philology, Literature and Linguistics - Via S. Maria 36, I-56126 Pisa, Italy. E-mail: alessandro.lenci@unipi.it

In formal semantic terms (Montague 1970; Kamp 1975), while the denotation of *white car* is said to be represented by the *intersection* of the denotations of *white* and *car* and the meaning of *skilled optician* is conceived as a *subset* of the denotation of optician, the *intensional* adjective *alleged* in *alleged murderer* is treated as a higher-order property that manipulates the modal parameter that is relevant for the interpretation of *murderer* (Chierchia and McConnell-Ginet 1990). Lastly and more interestingly for the study at hand, *red herring* classifies as an *idiom*, i.e. a semantically non-compositional multiword expression (MWE) characterized by figurativity, proverbiality and, in most cases, a certain emotional connotation (Cacciari and Glucksberg 1991; Nunberg, Sag, and Wasow 1994; Sag et al. 2002; Cacciari 2014). Furthermore, lack of compositionality entails lack of *salva-veritate*-interchangeability and systematicity (Fodor and Lepore 2002). On the one hand, idioms exhibit greater lexical and morphosyntactic frozenness with respect to literal phrases: while the component of a compositional combination can be replaced with a synonym or a semantically related word without considerably affecting the meaning of the whole expression (e.g., from *white car* to *white automobile*, from *skilled optician* to *skilled optometrist* and from *alleged murderer* to *alleged killer*), performing the same operation on an idiomatic expression (e.g., transforming *red herring* into *red fish*) hinders a possible figurative reading most of the time. With respect to systematicity, if we can understand the meaning of *white car* and *red herring* used in the literal sense, we can also understand what *white herring* and *red car* mean, but the same reasoning does not apply to *red herring* taken as an idiom.

AN compositionality therefore presents itself as a multifaceted and gradient phenomenon, whereby the interaction between the semantics of the adjective and the semantics of the noun leads to very different results in terms of the opacity of the output phrase. While previous computational literature on AN compositionality has been mainly concerned with the first three cases presented above (i.e. intersective, subsective and intensional), existing computational research on idiomaticity has mainly investigated verb-noun structures. In the present work we then decided to focus on the most opaque end of the AN compositionality continuum, by applying to Italian AN pairs a series of compositionality indices we already devised and tested on Italian idiomatic and non-idiomatic verbal constructions.

In developing the compositionality measures in Senaldi, Lebani, and Lenci (2016), we were mainly inspired by two groups of previous computational works. On the one hand, Lin (1999) and Fazly, Cook, and Stevenson (2009) label a given word combination as idiomatic if the Pointwise Mutual Information (PMI) (Church and Hanks 1991) between its component words is higher than the PMIs between the components of a set of lexical variants of this combination. These variants are obtained by replacing the component words of the original expressions with thesaurus-extracted synonyms. On the other hand, some researches have exploited Distributional Semantic Models (DSMs) (Sahlgren 2008; Turney and Pantel 2010), comparing the vector of a given phrase with the single vectors of its subparts (Baldwin et al. 2003; Venkatapathy and Joshi 2005; Fazly and Stevenson 2008) or comparing the vector of a phrase with the vector deriving from the sum or the products of their component vectors (Mitchell and Lapata 2010; Krčmář, Ježek, and Pecina 2013).

In our previous work (Senaldi, Lebani, and Lenci 2016), we started from a set of Italian verbal idiomatic (e.g. *dare i numeri* 'to lose one's marbles', lit. 'to give the numbers') and non-idiomatic (e.g. *leggere un libro* 'to read a book') phrases (henceforth *targets*) and generated lexical variants (simply *variants* henceforth) by replacing their components with semantic neighbors extracted from a window-based DSM and Italian MultiWordNet (Pianta, Bentivogli, and Girardi 2002). Examples of DSM-generated

lexical variants for *dare i numeri* are *offrire i numeri* 'to offer the numbers', *dare le unità* 'to give the units' and *offrire le unità* 'to offer the units', while examples of variants for *leggere un libro* are *sfogliare un libro* 'to leaf through a book', *leggere uno scritto* 'to read a work' and *sfogliare uno scritto* 'to leaf through a work'. Then, instead of measuring the associational scores between their subparts like in Lin (1999) and Fazly, Cook, and Stevenson (2009), we exploited Distributional Semantics to observe how different the context vectors of our targets were from the vectors of their variants, expecting, say, the vector of idiomatic *dare i numeri* to be less similar to the vectors of its variants *offrire i numeri* and *dare le unità* with respect to the similarity between the vector of non-idiomatic *leggere un libro* and the vectors of its variants *sfogliare un libro* and *leggere uno scritto*. Our proposal stemmed from the consideration that a high PMI value does not necessarily imply the idiomatic or multiword status of an expression, but just that its components co-occur more frequently than expected by chance, as in the case of *read* and *book* or *solve* and *problem*, which are all instances of compositional pairings. By contrast, what watertightly distinguishes an idiomatic from a collocation-like yet still compositional expression is their context of use. Comparing the distributional contexts of the original expressions and their alternatives should therefore represent a more precise refinement of the PMI-based procedure. Actually, idiomatic expressions vectors were found to be less similar to their variants vectors with respect to compositional expressions vectors. In some of our models, we also kept track of the variants that were not attested in our corpus by representing them as orthogonal vectors to the vector of the original expression, still achieving considerable results.

In the present contribution, we propose to extend the method in Senaldi, Lebani, and Lenci (2016) to a set of Italian AN pairs, since this kind of structure has been usually left aside in the idiom literature. The performance of our indices is also compared with that of addition-based and multiplication-based measures, which are taken as a reference point in the distributional literature on compositionality modeling (Mitchell and Lapata 2010; Krčmář, Ježek, and Pecina 2013).

2. Related work

The present work inserts itself in a longstanding tradition of studies addressing the computational modeling of compositionality. In the following section, we will first review previous research on the compositionality of AN structures in general. We will then switch our focus on how computational studies have so far tackled idiomatic expressions, which we said to represent the most opaque end of the compositionality continuum. As will become clear, our investigation on AN idioms combined insights from both these research strands.

As for the first group of works, Distributional Semantics (Harris 1954; Lenci 2008; Turney and Pantel 2010) has been extensively applied as a computational model of compositionality. DSMs encode target lexical items as vectors in a high-dimensionality space that register their co-occurrence statistics with some contextual features, like documents in a corpus or words occurring in the same contextual window. Vector similarity measures are then applied to model the semantic relatedness or similarity between the words represented by the distributional vectors. In recent years, this approach has been extended from representing the meaning of single words to modeling the semantics of complex phrases. Mitchell and Lapata (2010) propose three methods for combining vector representations that are regarded as a reference point in the distributional literature on compositionality. The *weighted additive* model derives the vector of a complex phrase p from the weighted sum of the vectors of its components u and v (which in our case

stand for the adjective and the noun respectively) and roughly corresponds to feature union:

$$p = \alpha u + \beta v \tag{1}$$

The *pointwise multiplicative* model, which corresponds to feature intersection, multiplies each corresponding pair of dimensions of the u and v vectors to derive the corresponding dimension of the p vector. In this way, mutually exclusive features are reduced to zero in the final vector:

$$p = u_i v_i \tag{2}$$

Finally, the *dilation* model decomposes a head vector v (the noun vector in AN strings) into a parallel and an orthogonal component with respect to the modifier vector u (the adjective vector) and stretches the parallel component by a factor λ:

$$p = (\lambda - 1)(u \cdot v)u + (u \cdot u)v \tag{3}$$

So as to balance the way adjectives and nouns contribute to the meaning of the whole phrase, Guevara (2010) proposes a *full additive* model, in which the two n-dimensional component vectors are multiplied by two n x n weight matrices before being summed:

$$p = Au + Bv \tag{4}$$

The two A and B matrices are estimated by means of partial least squares regression, using u and v as predictors and the corresponding observed AN pair vector as dependent variable. The problem of estimating the A and B matrices in a full additive model is also investigated by Zanzotto et al. (2010), who come up with a linear equation system that is solved by resorting to Moore-Penrose pseudo-inverse matrices (Penrose 1955). To take account of the fact that each adjective can interact differently with the semantics of the noun it modifies, Baroni and Zamparelli (2010) propose a *lexical function model* that draws on the Fregean conception of compositionality as function application and learns adjective-specific functions by predicting the dimensions of the observed AN pair vectors from the dimensions of the component noun vectors. The estimated matrix U is then multiplied by the noun vector v:

$$p = Uv \tag{5}$$

The adjective is then conceived of as a function that takes the meaning of the noun as an argument and returns the meaning of the modified noun. Boleda et al. (2013) compare the performance of all the aforementioned compositionality models on intensional vs. non-intensional adjectival modification. Their hypothesis is that the full additive and the lexical function models should achieve better scores in modeling intensional modification, since they should represent an attempt to transpose formal semantics higher-order modification into distributional semantic terms. On the contrary, non-intensional adjectival modification should be modeled equally well by the weighted additive and the pointwise multiplicative ones, which are supposed to reflect feature combination. Their findings anyway show an overall better performance of matrix-based methods, irrespectively of the kind of modification at play (intensional vs. non-intensional).

To obviate the limitation of the lexical function model in treating rare adjectives, Bride, Van de Cruys, and Asher (2015) come up with tensor for adjectival composition $\mathscr{A}$ which replaces adjective-specific matrices and is multiplied by the adjective vector u with a tensor dot product. The resulting matrix X is then multiplied with the noun vector. Hartung et al. (2017) apply all the compositional operations listed so far on CBOW word embeddings (Mikolov et al. 2013) of adjectives and nouns, registering superior performances with respect to count-based models in attribute selection and phrase similarity prediction tasks. Finally, Asher et al. (2017) resort to Latent Vector Weighting and tensor factorization to implement the Type Composition Logic (Asher 2011) conception of adjective-noun composition as a combination of two properties respectively representing the contextual contribution of the noun on the adjectival meaning and vice versa.

As regards previous computational studies on idiomaticity, two complementary issues have been mainly addressed so far: automatically separating potentially idiomatic strings like *spill the beans* from strings that can only receive a literal reading like *read a book* (*idiom type identification*) and automatically telling apart idiomatic vs. literal usages of a given string in context (*idiom token identification*; e.g., *John finally kicked the bucket after being ill for more than a decade* vs. *John kicked the bucket after he had accidentally stumbled on it*). Since in the present work we will carry out a task of the first kind, we will just review the existing literature on idiom type detection. While some scholars like Graliński (2012) try to rely just on shallow features like metalinguistic markers (e.g. *proverbially* or *literally*) and quotation marks to spot the presence of idioms in running text, most research has exploited those linguistic properties that typically distinguish idioms from literals, namely compositionality and lexicosyntactic fixedness. Tapanainen, Piitulainen, and Järvinen (1998) compare the frequency of a target noun as object with the number of verbs that appear with that object, assuming that objects of idiomatic constructions occur with just one or a few verbs at most. McCarthy, Keller, and Carroll (2003) focus on the compositionality of phrasal verbs (e.g. *eat up*, *blow up*, etc.), finding a strong correlation between human compositionality judgments and thesaurus-based measures of the overlap between the neighbors of a phrasal verb (e.g. *eat up*) and those of its simplex verb (e.g. *eat*). Evert, Heid, and Spranger (2004) and Ritz and Heid (2006) use frequency information to determine the preferred morphosyntactic features of idiomatic expressions which distinguish them from compositional constructions, while Widdows and Dorow (2005) extract asymmetric lexicosyntactic patterns such as *A and/or B* which never occur in the reversed order *B and/or A* in their corpus. Such a fixed linear order emerges as a clue of various kinds of relationships between the lexemes pairs, among which idiomatic ones. Bannard (2007) studies syntactic variability of VP idioms, in the form of determiner variability, internal modification via adjectives and passivization. Conditional PMI is used to calculate how the syntactic variation of the pair differs from what would be expected considering the variation of the single lexemes. Muzny and Zettlemoyer (2013) propose a supervised technique for identifying idioms among the Wiktionary lexical entries with lexical and graph-based features extracted from Wiktionary and WordNet.

A group of studies employ distributional methods that compare the vector of a phrase with the vectors of its components (Baldwin et al. 2003; Venkatapathy and Joshi 2005; Fazly and Stevenson 2008) or with the vector deriving from the sum or the products of their components (Mitchell and Lapata 2010; Krčmář, Ježek, and Pecina 2013). Among their vector-based indices, Fazly and Stevenson (2008) also include the cosine distance between the vector of a MWE as a whole and the vector of a verb that is morphologically related to the multiword noun, e.g. between *make a decision* and

decide. Finally, in a similar fashion to our proposal, some scholars have more precisely addressed lexical fixedness as a clue of idiomaticity. Lin (1999) classifies a phrase as non-compositional if the PMI between its components is significantly different from the mutual information of its variants. Each of these alternative forms is obtained by replacing one word in the original phrase with a semantic neighbour. For example, *red tape*, which means 'bureaucracy', receives a high non-compositionality score, because the PMI between *red* and *tape* is far higher than the PMI between *yellow* and *tape* or *black* and *tape*, *yellow tape* and *black tape* being the thesaurus-generated variants of *red tape*. On the other hand, *economic impact* is labeled as compositional, since its PMI is very similar to the PMIs of its variants like *financial impact* and *economic consequence*. Fazly, Cook, and Stevenson (2009) elaborate on Lin's idea, labeling a given combination as idiomatic if the PMI between its constituents is significantly different from the average PMI between the components of its variants. In Senaldi, Lebani, and Lenci (2016) we built on the aforementioned distributional and variant-based approaches by investigating how similar the vectors of a set of target Italian verbal idiomatic and non-idiomatic constructions were to the vectors of lexical variants of these targets that were generated from a window-based DSM and the Italian section of MultiWordNet (Pianta, Bentivogli, and Girardi 2002). All in all, as we expected, idiom vectors appeared to be less similar to their variants vectors with respect to literal phrase vectors.

3. Our proposal

In this study, we firstly aim at extending the variant-based method tested in Senaldi, Lebani, and Lenci (2016) on verbal idioms to noun-adjective expressions, which are mostly neglected in the idiom literature, to observe whether our indices can perform effectively in separating idiomatic vs. non-idiomatic AN constructions as well. Secondly, differently from our former work, we compare the variant-based method against conventional additive and multiplicative compositionality indices proposed in the distributional literature (Mitchell and Lapata 2010; Krčmář, Ježek, and Pecina 2013). Finally, beside using a window-based DSM and Italian MultiWordNet (Pianta, Bentivogli, and Girardi 2002) to extract our variants, we also experimented with a syntactic-based DSM (Padó and Lapata 2007; Baroni and Lenci 2010) that keeps track of the dependency relations between a given target and its contexts, to see whether variants generated with a different kind of distributional information can lead to improved performances.

The rest of the paper is organized as follows: in Section 4 we describe the dataset of target AN constructions we started from and the generation and extraction of the lexical variants for our targets from a Linear DSM, a Structured DSM and the Italian section of MultiWordNet (Pianta, Bentivogli, and Girardi 2002); in Section 5 we describe the collection of human-elicited idiomaticity judgments we used as a gold standard to assess the performance of our measures; in Section 6 we present the compositionality indices we tested on our dataset; both quantitative results and a qualitative error analysis will be provided in Section 7; finally, in Section 8 we provide some concluding remarks and point out possible future research directions.

4. Data extraction

4.1 Extracting the target expressions

All in all, our dataset is composed of 26 types of Italian noun-adjective and adjective-noun combinations, including 13 Italian idioms selected from an idiom dictionary

(Quartu 1993) and then extracted from the itWaC corpus (Baroni et al. 2009), which totalizes about 1,909M tokens. The frequency of these targets vary from 21 (*alte sfere* 'high places', lit. 'high spheres') to 194 (*punto debole* 'weak point, weak spot'). The remaining 13 items are compositional pairs of comparable frequencies (e.g., *nuova legge* 'new law' or *scrittore famoso* 'famous writer').

4.2 Extracting lexical variants

As in Senaldi, Lebani, and Lenci (2016), we adopted two different procedures for deriving lexical variants out of our targets, since we were interested in observing whether a fully automatic extraction method like the DSM-based one performed comparably with a more careful but time-consuming manual selection carried out on Italian MultiWordNet (Pianta, Bentivogli, and Girardi 2002).

Linear DSM variants. For both the noun and the adjective of each target, we extracted its top cosine neighbors in a window-based DSM created from the La Repubblica corpus (Baroni et al. 2004) (about 331M tokens). In Senaldi, Lebani, and Lenci (2016) we experimented with different thresholds of selected top neighbors (3, 4, 5 and 6). Since the number of top neighbors that were extracted for each constituent of the target did not significantly affect our performances, we decided to use the maximum number (i.e., 6) for the present study. In the DSM, all the content words occurring more than 100 times were represented as target vectors, ending up with 26,432 vectors, while the top 30,000 content words were used as dimensions. The co-occurrence counts were collected with a context window of $\pm$ 2 content words from each target word. The co-occurrence matrix was then weighted by Positive Pointwise Mutual Information (PPMI) (Evert 2008), which calculates whether the co-occurrence of two words x and y is more frequent than expected by chance and sets to zero all the negative results:

$$PPMI(x, y) = max(0, log\frac{P(x, y)}{P(x)P(y)}) \quad (6)$$

Finally, Singular Value Decomposition (SVD) (Deerwester et al. 1990) to 300 latent dimensions was run on our initial 30,000-dimension matrix. The variants were finally obtained by combining the adjective with each of the noun's top 6 neighbors (e.g., *punto debole* $\rightarrow$ *vantaggio debole* 'weak advantage', *termine debole* 'weak end', etc.), the noun with all the top 6 neighbors of the adjective (e.g., *punto debole* $\rightarrow$ *punto fragile* 'fragile point', *punto incerto* 'uncertain point', etc.) and finally all the top 6 neighbors of the adjective and the noun with each other (e.g., *punto debole* $\rightarrow$ *vantaggio fragile* 'fragile advantage', *termine incerto* 'uncertain end', etc.), ending up with 48 potential Linear DSM variants per target.

Structured DSM variants. While unstructured (i.e., window-based) DSMs just record the words that linearly precede or follow a target lemma when collecting co-occurrence counts, structured DSMs conceive co-occurrences as $< w_1, r, w_2 >$ triples, where *r* represents the lexico-syntactic pattern or, like in our case, the parser-extracted dependency relation between w_1 and w_2 (Padó and Lapata 2007; Baroni and Lenci 2010). The grounding assumption of such models is that the syntactic relation linking the two words should act as a cue of their semantic relation (Grefenstette 1994; Turney 2006; Padó and Lapata 2007). Actually, structured DSMs have been shown to perform competitively or better than linear DSMs in a variety of semantic tasks, like modeling similarity judgments or selectional preferences or detecting synonyms (Baroni and Lenci 2010).

Since we wanted to exploit different kinds of distributional information to generate our variants, following the method described in Baroni and Lenci (2010) we created a structured DSM from La Repubblica (Baroni et al. 2004), where all the content words occurring more than 100 times were kept as targets and the co-occurrence matrix was once again weighted via PPMI and reduced to 300 latent dimensions. For each target, we generated 48 virtual lexical variants with the same procedure described for the window-based DSM variants.

iMWN variants. For each noun, we extracted the words occurring in the same synsets and its co-hyponyms from Italian MultiWordNet (iMWN) (Pianta, Bentivogli, and Girardi 2002). As for the adjectives, we experimented with two different approaches, extracting just their synonyms in the first case (iMWN$_{syn}$ variants) and adding also the antonyms in the second case (iMWN$_{ant}$ variants). Since antonyms were not available in the Italian section of MultiWordNet, we had to translate them from the English WordNet (Fellbaum 1998). For each noun and adjective, we kept its top 6 iMWN neighbors in terms of cosine similarity in the same DSM used to acquire the linear DSM variants. Once again, this method provided us with up to 48 potential iMWN$_{syn}$ and 48 potential iMWN$_{ant}$ variants per target.

5. Gold standard idiomaticity judgments

To validate our computational indices, we presented 9 linguistics students with our 26 targets and asked them to rate how idiomatic each expression was on a 1-7 scale, with 1 standing for "totally compositional" and 7 for "totally idiomatic". The 26 targets were presented together in three different randomized lists. Each list was rated by three subjects. The mean score given to our idioms was 6.10 (SD = 0.77), while the mean score given to compositional expressions was 2.03 (SD = 1.24). This difference was proven by a t-test to be statistically significant (t = 10.05, $p < 0.001$). Inter-coder reliability, measured via Krippendorff's α (Krippendorff 2012), was 0.76. Following established practice, we took such value as a proof of reliability for the elicited ratings (Artstein and Poesio 2008).

6. Calculating compositionality indices

Two kinds of compositionality indices were computed for our 26 idiomatic and non-idiomatic AN targets. The former, described in Subsection 6.1, comprehends the variant-based measures we previously tested on verbal idioms (Senaldi, Lebani, and Lenci 2016). The latter, presented in Subsection 6.2, comprehends addition-based and multiplication-based measures that have been previously proposed in the distributional literature (Mitchell and Lapata 2010; Krčmář, Ježek, and Pecina 2013).

6.1 Variant-based indices

The variant generation procedure explained in Section 4.2 provided us with just automatically generated and potential lexical alternatives for our initial constructions, but of course it could not assure us that they were actually attested in our corpus. For each of our 26 targets, we extracted from itWaC all the occurrences we could find of their respective 48 linear DSM, structured DSM, iMWN$_{syn}$ and iMWN$_{ant}$ variants. For every variant type (linear DSM, structured DSM, iMWN$_{syn}$ and iMWN$_{ant}$) we built a DSM from itWaC representing the 26 targets and their variants as vectors. While the dimension of the La Repubblica corpus seemed to be enough for the variants extraction

procedure, we resorted to five-times bigger itWaC to represent the variants as vectors and compute the compositionality scores to avoid data sparseness and have a considerable number of variants frequently attested in our corpus. Using two different corpora has the additional advantage of showing the variants method to be generalizable to corpora of different text genres and size. Co-occurrence statistics recorded how many times each target or variant construction occurred in the same sentence with each of the 30,000 top content words in the corpus. The matrices were then weighted with PPMI and reduced to 150 dimensions via SVD. We finally calculated four different indices:

Mean. The mean of the cosine similarities between the vector of a target construction and the vectors of its variants.

Max. The maximum value among the cosine similarities between a target vector and its variants vectors.

Min. The minimum value among the cosine similarities between a target vector and its variants vectors.

Centroid. The cosine similarity between a target vector and the centroid of its variants vectors.

In our predictions, ranking our 26 targets in ascending order according to each of the four compositionality indices should result in idioms being placed at the top of the ranking and non-idioms at the bottom of it, since idioms are expected to be the least similar ones with respect to their lexical alternatives.

As we said before, the variant creation method presented in Section 4.2 consists in the generation of a list of potential variants for each target construction, but most of these were not actually found in itWaC (Baroni et al. 2009). Let's consider the Linear DSM variants of the idiom *testa calda* 'hothead'. While 11 of the Linear DSM-generated variants were actually retrieved in itWaC, like *mano fredda* 'cold hand' (12 tokens), *mano calda* 'warm hand' (6 tokens) and *piede freddo* 'cold foot' (2 tokens), the other 37 variants, like *mano torrida* 'torrid hand', *gamba fresca* 'cool leg' and *spalla umida* 'moist shoulder', were not attested at all. Since some of our targets had many variants that were not found in itWaC, each measure was computed twice: in the first case we simply did not consider the non-occurring variants; in the second case, we conceived them as vectors that were orthogonal to their target vector. For the sake of clarity, the first kind of models will henceforth be referred to as *no models*, while the latter will be labeled *orth models*. The rationale behind taking this negative evidence into account was that, if lexical alternatives for a given AN pair could not even be traced in the corpus, this should be taken as an additional and stronger clue of its formal idiosyncrasy and idiomatic status. Non-occurring variants were encoded as orthogonal vectors since, given a vector x, the cosine similarity between x and a vector y which is perpendicular to x is equal to 0.0, so that y is the most distant vector from x. Such an implementation reflects the consideration that a non-existing variant is *de facto* conceivable as an expression whose meaning is the farthest possible from the meaning of the respective target expression and which contributes in tilting its compositionality score towards 0.0. In practical terms, for the Mean, Max and Min indices, this meant to automatically set to 0.0 the cosine similarity between the target vector and the vector of the non-occurring variant at hand. For the Centroid measure, we first computed the cosine similarity between the target vector and the centroid of its attested variants (cs_a). From this initial cosine

Table 1
Number of non-attested variants for each of the four DSM spaces built from Linear DSM, Structured DSM, iMWN$_{syn}$ and iMWN$_{ant}$ variants respectively.

Space	Total zero variants	Zero variants per target	
		Mean	SD
Linear DSM	810	31.15	11.21
Structured DSM	1002	38.54	10.03
iMWN$_{syn}$	717	27.58	14.52
iMWN$_{ant}$	703	27.04	13.56

value we then subtracted the product between the number of non-attested variants (n), cs_a and a costant factor k. This factor k, which was set to 0.01 in previous investigations, represented the contribution of each zero variant in reducing the target-variants similarity towards 0.0. k was multiplied by the original cosine since we hypothesized that zero variants contributed differently in lowering the target-variants similarity, depending on the construction under consideration:

$$Centroid = cs_a - (cs_a \cdot k \cdot n) \tag{7}$$

Table 1 reports how many variants were not attested in our corpus for each of the four spaces we built (Linear DSM, Structured DSM, iMWN$_{syn}$ and iMWN$_{ant}$). As we can see, the issue of non-attested variants was an across-the-board phenomenon which involved each of the four space types and was only slightly smaller in iMWN-derived spaces.

6.2 Addition-based and multiplication-based indices

The indices in Section 6.1 were compared against two of the measures by Mitchell and Lapata (2010) and Krčmář, Ježek, and Pecina (2013). We trained a DSM on itWaC that represented all the content words with token frequency > 300 and our 26 targets as row-vectors and the top 30,000 content words as contexts. The co-occurrence window was still the entire sentence and the weighting was still the PPMI. SVD was carried out to 300 final dimensions. Please note that the context vectors of a given word did not include the co-occurrences of a target idiom that was composed of that word (e.g. the vector for *punto* did not include the contexts of *punto debole*), so as to make sure that the vector of the idiom as a whole was compared with vectors that actually represented the literal meaning of its constituents. We then computed the following measures:

Additive. The cosine similarity between a target vector and the vector resulting from the component-wise sum of the noun vector and adjective vector. This roughly corresponded to performing a weighted addition as explained in Section 2 with both weights set to 1, since in our assumption both component vectors equally contributed to the representation of the whole phrase.

Multiplicative. The cosine similarity between a target vector and the vector resulting from the component-wise product of the noun vector and adjective vector.

7. Results and discussion

Our 26 targets were sorted in ascending order for each compositionality score. In each ranking, we predicted idioms (our positives) to be placed at the top and compositional phrases (our negatives) to be placed at the bottom, since we expected idiom vectors to be less similar to the vectors of their variants. First and foremost, we must report that three idioms for every type of variants (Window-based DSM, Structured DSM and iMWN) obtained a 0.0 score for all the variant-based indices since no variants were found in itWaC. Nevertheless, we kept this information in our ranking as an immediate proof of the multiwordness and idiomaticity of such expressions. These were *punto debole*, *passo falso* 'false step' and *colpo basso* 'cheap shot' for the Structured DSM spaces, *punto debole*, *pecora nera* 'black sheep' and *faccia tosta* 'cheek' for the iMWN spaces and *punto debole*, *passo falso* and *zoccolo duro* 'hard core' for the Window-based DSM spaces.

Table 2 reports the 5 best models for Interpolated Average Precision (IAP), the F-measure at the median and Spearman's ρ correlation with our gold standard idiomaticity judgments respectively. Coherently with Fazly, Cook, and Stevenson (2009), IAP was computed as the average of the interpolated precisions at recall levels of 20%, 50% and 80%. Interestingly, while Additive was the model that best ranked idioms before non-idioms (IAP), closely followed by our variant-based measures, and figured among those with the best precision-recall trade-off (F-measure), Multiplicative performed comparably to the Random baseline. Although at odds with Mitchell and Lapata (2010)'s results, such a scarce performance of the Multiplicative model is in line with the findings by Baroni and Zamparelli (2010) and Boleda et al. (2013), who both found Addition to be more effective in modeling AN compositionality. While Baroni and Zamparelli (2010) hypothesize that product-based indices could be disadvantaged by SVD, which can output negative dimensions and therefore lead to counterintuitive component-wise product results, Boleda et al. (2013) suggest that the feature union performed by addition could more accurately represent the semantics of AN structures with respect to the massive feature intersection provoked by multiplication, whereby shared dimensions are inflated and mutually exclusive ones are canceled out.

As for ρ correlation with the human ratings, the best score was achieved by one of our variant-based measures, namely Structured DSM Mean$_{orth}$ (-0.68). Additive did not belong to the 5 models with top correlation, but still achieved a high significant ρ score (-0.62). It's worth noting that, as we wished, all these correlational indices are negative: the more the subjects perceived a target to be idiomatic, the less its vector was similar to its variants. Max and Min never appeared among the best performing measures, with all top models using Mean and Centroid. Moreover, the DSM models that worked the best for IAP and F-measure both used dependency-related distributional information, with window-based DSM models not reaching the top 5 ranks. This difference was nonetheless ironed out when looking at the Top ρ models. In Senaldi, Lebani, and Lenci (2016), models encoding zero variants as orthogonal vectors ranked better than the other ones only when predicting speakers' judgments, while *no* models scored better as for IAP and F-measure. In this work, the majority of best IAP and F-measure models, and *de facto* all top ρ models, are *orth* models, thus showing that considering negative evidence about lexical variants is fruitful for AN compositionality estimation. In light of the overall results, generating variants from DSMs emerges as the best method, since these models had comparable performances with MultiWordNet-based models, but were fully automatic and did not require an intensive and time-consuming manual selection of the variants. Finally, the presence of antonymy-related information for iMWN models did not appear to influence the performances considerably.

Table 2
Best models ranked by IAP (top), F-measure at the median (middle) and Spearman's ρ correlation with the speakers' judgments (bottom) against the multiplicative model and the random baseline (** = $p < 0.01$, *** = $p < 0.001$).

Top IAP Models	**IAP**	**F**	ρ
Additive	0.85	0.77	-0.62***
Structured DSM Mean$_{orth}$	0.84	0.85	-0.68***
iMWN$_{syn}$ Centroid$_{orth}$	0.83	0.85	-0.57**
iMWN$_{ant}$ Centroid$_{orth}$	0.83	0.77	-0.52**
iMWN$_{ant}$ Mean$_{orth}$	0.83	0.69	-0.64***

Top F-measure Models	**IAP**	**F**	ρ
Structured DSM Mean$_{orth}$	0.84	0.85	-0.68***
iMWN$_{syn}$ Centroid$_{orth}$	0.83	0.85	-0.57**
Additive	0.85	0.77	-0.62***
iMWN$_{ant}$ Centroid$_{orth}$	0.83	0.77	-0.52**
iMWN$_{syn}$ Centroid$_{no}$	0.82	0.77	-0.57**

Top ρ Models	**IAP**	**F**	ρ
Structured DSM Mean$_{orth}$	0.84	0.85	-0.68***
Window-based DSM Mean$_{orth}$	0.75	0.69	-0.66***
iMWN$_{syn}$ Mean$_{orth}$	0.77	0.77	-0.65***
iMWN$_{syn}$ Mean$_{no}$	0.70	0.69	-0.65***
iMWN$_{ant}$ Mean$_{orth}$	0.83	0.69	-0.64***

Baselines	**IAP**	**F**	ρ
Multiplicative	0.58	0.46	0.03
Random	0.50	0.31	0.05

7.1 Error analysis

In order to understand whether specific items in our dataset could be particularly troublesome for our algorithms, we carried out a qualitative analysis of the most common *false positives* (FPs henceforth, i.e. literals wrongly labeled as idioms) and *false negatives* (FNs henceforth, i.e. idioms wrongly classified as compositional expressions). One of the most common FPs was the ambiguous expression *pesce grosso* 'big fish', which is sometimes used in newspapers to denote an influential person inside a criminal organization. Maybe, given the nature of the corpora we selected, it would have been

better to treat this AN pair as an idiom in the first place. As it happened in Senaldi, Lebani, and Lenci (2016), other frequent FPs were compositional but collocation-like combinations like *gruppo numeroso* 'large group' or *crescita rapida* 'rapid growth', which contain nouns and adjectives that co-occur very often. On the other hand, while FNs in our previous study mostly consisted in strongly ambiguous expressions liable to both a figurative and a literal reading according to the context, in this case they were evident idioms, like *testa calda* or *punto fermo* 'fundamental point' (lit. 'still point'). To discover why our algorithms ended up classifying them as literals, we had a look at the lexical variants that were generated and were available for each of them. For *testa calda*, only 1 Structured DSM variant occurring just 1 time and 2 iMWN variants occurring 1 time were found in itWaC and this led to a skewed and not reliable compositionality assessment. As regards *punto fermo*, the variants that were generated, like *punto solido* 'solid point' or *passaggio chiaro* 'clear step' seem to refer to the same semantic field of the original expression and exhibited quite predictably a similar distribution.

8. Conclusions

AN compositionality constitutes a highly multifaceted and complex phenomenon, whereby the interaction of the semantics of the two constituents can lead to different results, from fully transparent to fully opaque word combinations. Since AN structures are usually neglected in the computational literature on idioms, we decided to focus on the most opaque end of the AN compositionality continuum, applying to AN constructions the same variant-based distributional measures we had previously proposed and tested on verbal idioms (Senaldi, Lebani, and Lenci 2016). Once again, effective performances were obtained, therefore confirming that comparing the vector of a given phrase with the vectors of its lexical variants is a reliable way to estimate its compositionality. More specifically, models computing the mean cosine similarity between the target vector and the variant vectors or the cosine similarity between the target vector and the centroid of the variant vectors stood out as the best performing ones, as did models that kept track of the variants that were not found for a given target in the form of orthogonal vectors. Interestingly, our measures performed comparably to or even better than the Additive method proposed in the distributional literature (Krčmář, Ježek, and Pecina 2013), while the Multiplicative one performed considerably worse than all our models, together with the Random baseline. This finding mirrored results from previous studies (Baroni and Zamparelli 2010; Boleda et al. 2013) and suggested that the feature intersection carried out by component-wise vector product is not a viable approach for modeling the semantics of AN combinations.

Future work will concern testing whether these variant-based measures can be successfully exploited to predict psycholinguistic data about the processing of idiom compositionality and flexibility, together with other corpus-based indices of idiomaticity. Moreover, we plan to extend the comparison of the variant-based approach to matrix and tensor-based models of AN composition.

References

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Asher, Nicholas. 2011. *Lexical meaning in context: A web of words*. Cambridge University Press, Cambridge, UK.

Asher, Nicholas, Tim Van de Cruys, Antoine Bride, and Márta Abrusán. 2017. Integrating type theory and distributional semantics: A case study on adjective–noun compositions.

Computational Linguistics, 42(4):703–725.
Baldwin, Timothy, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan, July 12, 2003.
Bannard, Colin. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 1–8, Prague, Czech Republic, June 28, 2007.
Baroni, Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1771–1774, Lisbon, Portugal, May 26-28, 2004.
Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.
Baroni, Marco and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.
Baroni, Marco and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA, October 9-11, 2010.
Boleda, Gemma, Marco Baroni, Louise McNally, and Nghia Pham. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 35–46, Potsdam, Germany, March 19-22, 2013.
Bride, Antoine, Tim Van de Cruys, and Nicholas Asher. 2015. A Generalisation of Lexical Functions for Composition in Distributional Semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 281–291, Beijing, China, July 26-31, 2015.
Cacciari, Cristina. 2014. Processing multiword idiomatic strings: Many words in one? *The Mental Lexicon*, 9(2):267–293.
Cacciari, Cristina and Sam Glucksberg. 1991. Understanding idiomatic expressions: The contribution of word meanings. *Advances in Psychology*, 77:217–240.
Chierchia, Gennaro and Sally McConnell-Ginet. 1990. *Meaning and Grammar: An Introduction to Semantics*. MIT Press, Cambridge, MA.
Church, Kenneth W. and Patrick Hanks. 1991. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391.
Evert, Stefan. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, volume 2. Mouton de Gruyter, Berlin & New York, pages 1212–1248.
Evert, Stefan, Ulrich Heid, and Kristina Spranger. 2004. Identifying morphosyntactic preferences in collocations. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 907–910, Lisbon, Portugal, May 26-28, 2004.
Fazly, Afsaneh, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 1(35):61–103.
Fazly, Afsaneh and Suzanne Stevenson. 2008. A distributional account of the semantics of multiword expressions. *Italian Journal of Linguistics*, 1(20):157–179.
Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.
Fodor, Jerry A. and Ernest Lepore. 2002. *The Compositionality Papers*. Oxford University Press, Oxford, UK.
Graliński, Filip. 2012. Mining the web for idiomatic expressions using metalinguistic markers. In *Proceedings of Text, Speech and Dialogue: 15th International Conference*, pages 112–118, Brno, Czech Republic, September 3-7, 2012.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Guevara, Emiliano. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37, Uppsala, Sweden, July 16, 2010.

Harris, Zellig S. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Hartung, Matthias, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano. 2017. Learning Compositionality Functions on Word Embeddings for Modelling Attribute Meaning in Adjective-Noun Phrases. In *Proceedings of the 15th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 54–64, Valencia, Spain, April 3-7, 2017.

Kamp, Hans. 1975. Two theories about adjectives. In Edward L. Keenan, editor, *Formal Semantics of Natural Language*. Cambridge University Press, Cambridge, UK, pages 123–155.

Krippendorff, Klaus. 2012. *Content analysis: An introduction to its methodology*. Sage, Los Angeles, London, New Delhi & Singapore.

Krčmář, Lubomír, Karel Ježek, and Pavel Pecina. 2013. Determining Compositionality of Expresssions Using Various Word Space Models and Measures. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 64–73, Sofia, Bulgaria, August 9, 2013.

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31.

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324, College Park, Maryland, June 20-26, 1999.

McCarthy, Diana, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan, July 12, 2003.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26t th International Conference on Neural Information Processing System*, pages 3111–3119, Lake Tahoe, Nevada, December 5-10, 2013.

Mitchell, Jeff and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429.

Montague, Richard. 1970. Universal grammar. *Theoria*, 36(3):373–398.

Muzny, Grace and Luke S. Zettlemoyer. 2013. Automatic Idiom Identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, WA, October 19-21, 2013.

Nunberg, Geoffrey, Ivan Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Padó, Sebastian and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Partee, Barbara. 1995. Lexical semantics and compositionality. In Lila R. Gleitman, Daniel N. Osherson, and Mark Liberman, editors, *An invitation to cognitive science: Language*, volume 1. MIT Press, Cambridge, MA, pages 311–360.

Penrose, Roger. 1955. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413.

Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing and Aligned Multilingual Database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India, January 21-25, 2002.

Quartu, Monica B. 1993. *Dizionario dei modi di dire della lingua italiana*. RCS Libri, Milan, Italy.

Ritz, Julia and Ulrich Heid. 2006. Extraction tools for collocations and their morphosyntactic specificities. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1925–1930, Genoa, Italy, May 24-26, 2006.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, Mexico City, Mexico, February 17-23, 2002.

Sahlgren, Magnus. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

Senaldi, Marco S. G., Gianluca E. Lebani, and Alessandro Lenci. 2016. Lexical variability and compositionality: Investigating idiomaticity with distributional semantic models. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 21–31, Berlin, Germany, August

11, 2016.
Tapanainen, Pasi, Jussi Piitulainen, and Timo Järvinen. 1998. Idiomatic object usage and support verbs. In *Proceedings of the 17th international conference on Computational Linguistics*, pages 1289–1293, Montreal, Quebec, Canada, August 10-14, 1998.
Turney, Peter D. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.
Turney, Peter D. and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.
Venkatapathy, Sriram and Aravid Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 899–906, Vancouver, British Columbia, Canada, October 06-08, 2005.
Widdows, Dominic and Beate Dorow. 2005. Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 48–56, Ann Arbor, Michigan, June 30, 2005.
Zanzotto, Fabio Massimo, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating Linear Models for Compositional Distributional Semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1263–1271, Beijing, China, August 23-27, 2010.