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

Marco S. G. Senaldi
Laboratorio di Linguistica
Scuola Normale Superiore
Pisa, Italy
marco.senaldi@sns.it

Gianluca E. Lebani, Alessandro Lenci
Computational Linguistics Laboratory
Department of Philology, Literature, and Linguistics
University of Pisa, Italy
gianluca.lebani@for.unipi.it
alessandro.lenci@unipi.it

Abstract

English. In this work we employed a set of 26 Italian noun-adjective expressions to test compositionality indices that compare the distributional vector of an expression with the vectors of its lexical variants. These were obtained by replacing the components of the original expression with semantically related words. Our indices performed comparably or better than other compositionality measures reported in the distributional literature.

Italiano. *In questo lavoro si è utilizzato un set di 26 espressioni italiane nome-aggettivo per testare degli indici di composizionalità che confrontano il vettore distribuzionale di un'espressione con i vettori delle sue varianti lessicali. Queste sono state ottenute sostituendo i componenti dell'espressione di partenza con parole semanticamente correlate. La performance dei nostri indici si è dimostrata comparabile o superiore a quella di altri indici di composizionalità riportati nella letteratura distribuzionale.*

1 Introduction and previous research

While a *white car* is *white* and is a *car*, a *red herring* in a sentence like *I thought he was the culprit, but he was a red herring* is neither *red* nor a *herring*, but indicates something that distracts someone from a relevant issue. The former expression is compositional, since its meaning derives from the composition of the meanings of its subparts (Werning et al., 2012). The latter, by contrast, is an *idiom*, a non-compositional, figurative and proverbial word combination belonging to the wider class of *Multiword Expressions* (Nunberg et al., 1994; Cacciari, 2014). The compositionality of a given expression entails *salva-veritate*-interchangeability and systematicity (Fodor and Lepore, 2002). First of all, if we replace the constituents of a compositional expression with synonyms or similar words (e.g., from *white car* to *white automobile*), the whole meaning is not altered. Secondly, 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. Both these properties are not valid for idioms, which always exhibit lexical fixedness to some extent: variants of idiomatic *red herring* like *red fish* or *white herring* can just have a literal reading.

Computational studies to date have proposed several techniques to automatically measure idiomaticity. Of note, Lin (1999) and Fazly et al. (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 semantically related words. Other 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 components (Mitchell and Lapata, 2010; Krčmář et al., 2013).

In a previous contribution (Senaldi et al., 2016),

we started from a set of Italian verbal idiomatic and non-idiomatic phrases (henceforth our *targets*) and generated lexical variants (simply *variants* henceforth) by replacing their components with semantic neighbours extracted from a linear DSM and Italian MultiWordNet (Pianta et al., 2002). Then, instead of measuring the associational scores between their subparts like in Lin (1999) and Fazly et al. (2009), we exploited Distributional Semantics to observe how different the context vectors of our targets were from the vectors of their variants. 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. Noteworthily, most researches conducted so far have focused on verbal idioms, while the analysis of NP idioms like *red herring* or *second thoughts* has been usually left aside.

2 Applying variant-based distributional measures to noun-adjective pairs

In the present study, we firstly aimed to extend the variant-based method tested in Senaldi et al. (2016) on verbal idioms to noun-adjective expressions, which are mostly neglected in the idiom literature. In the second place, our former work lacked a comparison against conventional additive and multiplicative compositionality indices proposed in the distributional literature (Mitchell and Lapata, 2010; Krčmář et al., 2013). Finally, beside using a linear DSM and Italian MultiWordNet (Pianta et al., 2002) to extract our variants, we also experimented with a DSM (Padó and Lapata, 2007; Baroni and Lenci, 2010) which kept track of the syntactic dependency relations between a given target and its contexts.

3 Data extraction

3.1 Extracting the target expressions

All in all, our dataset was composed of 26 types of Italian noun-adjective and adjective-noun combinations. Of these, 13 were Italian idioms extracted from the itWaC corpus (Baroni et al., 2009), which totalizes about 1,909M tokens. The frequency of these targets varied from 21 (*alte sfere* 'high places', lit. 'high spheres') to 194 (*punto debole* 'weak point'). The remaining 13 items were compositional pairs of comparable frequencies (e.g., *nuova legge* 'new law').

3.2 Extracting lexical variants

Linear DSM variants. For both the noun and the adjective of our targets, we extracted its top cosine neighbors in a linear DSM created from the La Repubblica corpus (Baroni et al., 2004) (about 331M tokens). In Senaldi et al. (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 did not significantly affect our performances, for the present study we decided to use the maximum number (i.e., 6). 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 obtained matrix was then weighted by Positive Pointwise Mutual Information (PPMI) (Evert, 2008) and reduced to 300 latent dimensions via Singular Value Decomposition (SVD) (Deerwester et al., 1990). The variants were finally obtained by combining the adjective with each of the noun's top 6 neighbors, the noun with all the top 6 neighbors of the adjective and finally all the top 6 neighbors of the adjective and the noun with each other, ending up with 48 Linear DSM variants per target.

Structured DSM variants. While unstructured 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* rep-

resents the dependency relation between w_1 and w_2 (Padó and Lapata, 2007; Baroni and Lenci, 2010). Since we wanted to experiment with 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 lexical variants with the same procedure described for the linear DSM variants.

iMWN variants. For each noun, we extracted the words occurring in the same synsets and its co-hyponyms in Italian MultiWordNet (iMWN) (Pianta et al., 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). The antonyms were translated 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 48 iMWN variants per target.

4 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 targets were presented in three different randomized orders, with three raters per order. 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).

5 Calculating compositionality indices

For each of our 26 targets, we extracted from itWaC all the attested occurrences of the 48 linear DSM, structured DSM, iMWN$_{syn}$ and iMWN$_{ant}$ variants. We then computed two kinds of vector-based compositionality indices:

5.1 Variant-based indices

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. We also thought that using two different corpora had the additional advantage of showing the variants method to be generalizable to corpora of different text genres. 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 cosine similarity 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.

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 (*no* models); in the second case, we conceived them as orthogonal vectors to the target vector (*orth* models). For the Mean, Max and Min indices, this meant to automatically set to 0.0 the target-variant cosine similarity. 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 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:

$$Centroid = cs_a - (cs_a \cdot k \cdot n)$$

5.2 Addition-based and multiplication-based indices

The indices in Section 5.1 were compared against two of the measures described in Krčmář et al. (2013). We trained a DSM on itWaC that represented all the content words with $tokenfrequency > 300$ 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 vector of a given word did not include the co-occurrences of a target idiom or target compositional expression that was composed of that word (e.g. the vector for *punto* did not include the contexts of *punto debole*). We then computed the following measures:

Additive. The cosine similarity between a target vector and the vector resulting from the sum of the vectors of its components.

Multiplicative. The cosine similarity between a target vector and the vector resulting from the product of the vectors of its components.

6 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 notice that three idioms for every type of variants (Linear 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 idiomaticity of such expressions. These were *punto debole* 'weak point', *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 Linear DSM spaces.

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***
Linear 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***
Multiplicative	0.58	0.46	0.03
Random	0.50	0.31	0.05

Table 1: 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 (** = $p < 0.01$, *** = $p < 0.001$).

Table 1 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 et al. (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. The best correlation with idiomaticity judgments was instead achieved by one of our variant-based measures (-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 all these correlational indices are negative: the more the subjects perceived a target to be id-

iomatic, 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 linear DSM models not reaching the top 5 ranks. This difference was nonetheless ironed out when looking at the Top ρ models. Differently from what we observed for verbal idioms (Senaldi et al., 2016), the majority of our best models, and *de facto* all the Top ρ models, encoded zero variants as orthogonal vectors (*orth* models). Finally, the presence of antonymy-related information for iMWN models did not appear to influence the performances considerably.

7 Conclusions

In this contribution we applied to adjective-noun constructions the variant-based distributional measures we had previously tested on verbal idioms (Senaldi et al., 2016), obtaining effective performances. Interestingly, our measures performed comparably to or even better than the Additive method proposed in the distributional literature (Krčmář et al., 2013), while the Multiplicative one performed considerably worst than all our models, together with the Random baseline.

Future work will concern testing whether these variant-based measures can be succesfully exploited to predict psycholinguistic data about the processing of idiom compositionality and flexibility, together with other corpus-based indices of idiomaticity.

References

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

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni, 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.

Marco Baroni, 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.

Cristina Cacciari. 2014. Processing multiword idiomatic strings: Many words in one? *The Mental Lexicon*, 9(2):267–293.

Kenneth W. Church and Patrick Hanks. 1991. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Scott Deerwester, 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.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, volume 2, pages 1212–1248. Mouton de Gruyter.

Afsaneh Fazly and Suzanne Stevenson. 2008. A distributional account of the semantics of multiword expressions. *Italian Journal of Linguistics*, 1(20):157–179.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 1(35):61–103.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Jerry A. Fodor and Ernest Lepore. 2002. *The compositionality papers*. Oxford University Press.

Klaus Krippendorff. 2012. *Content analysis: An introduction to its methodology*. Sage.

Lubomír Krčmář, 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.

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

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

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

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

Emanuele Pianta, 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.

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

Marco S. G. Senaldi, 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.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Sriram Venkatapathy 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.

Markus Werning, Wolfram Hinzen, and Edouard Machery, editors. 2012. *The Oxford Handbook of Compositionality*. Oxford University Press.