Lexical Variability and Compositionality: Investigating Idiomaticity with Distributional Semantic Models

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

Gianluca E. Lebani and **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

In this work we carried out an *idiom type identification* task on a set of 90 Italian V-NP and V-PP constructions comprising both idioms and non-idioms. Lexical variants were generated from these expressions by replacing their components with semantically related words extracted distributionally and from the Italian section of MultiWordNet. Idiomatic phrases turned out to be less similar to their lexical variants with respect to non-idiomatic ones in distributional semantic spaces. Different variant-based distributional measures of idiomaticity were tested. Our indices proved reliable in identifying also those idioms whose lexical variants are poorly or not at all attested in our corpus.

1 Introduction

Extensive corpus studies have provided support to Sinclair (1991)'s claim that speakers tend to favor an *idiom principle* over an *open-choice principle* in linguistic production, resorting, where possible, to (semi-)preconstructed phrases rather than using compositional combinatorial expressions. These *multiword expressions* (MWEs) and *idioms* in particular (Nunberg et al., 1994; Sag et al., 2002; Cacciari, 2014; Siyanova-Chanturia and Martinez, 2014) exhibit an idiosyncratic behavior that makes their account troublesome for most grammar models (Chomsky, 1980; Jackendoff, 1997; Hoffmann and Trousdale, 2013), including restricted semantic compositionality and transparency, low morphosyntactic versatility and, crucially for the study at hand, a considerable degree of lexical fixedness. The existence of such prefabricated patterns ties in well with the basic tenets of constructionist approaches (Goldberg, 1995; Hoffmann and Trousdale, 2013), that view the lexicon and the grammar as a network of form-meaning correspondences spanning from abstract and complex syntactic schemata to single words and morphemes.

Idioms show a gradient behavior according to the lexicosyntactic variation that each of them can undergo. It has indeed been traditionally argued that while replacing the constituents of a literal combination like *to watch a movie* with synonymous or semantically related words (e.g. *to watch a film*) does not result in a significant change in meaning, modifying an idiomatic string like *to spill the beans* into something like *to spill the peas* entails the loss of the figurative interpretation (Cacciari and Glucksberg, 1991; Sag et al., 2002; Fazly and Stevenson, 2008). Actually, psycholinguistic studies investigating the comprehension of idiom lexical variants have found such alternative forms to be more acceptable when the idiom parts independently contribute to the idiomatic meaning (e.g. *burst the ice* from *break the ice*) than when they don't (e.g. *boot the bucket* from *kick the bucket*) (Gibbs et al., 1989) or when the idioms are more familiar to the speakers (McGlone et al., 1994). Anyway, while contributions of this kind are useful to assess whether potentially occurring variants can be understood by speakers or not, it is looking at corpus analyses that we can gain an insight into the actual occurrence of such lexical alternatives in real text. Moon (1998) and Duffley (2013) have found all kinds of idioms to be used sometimes in an altered form with the idiomatic reading preserved (e.g. *kick the pail* and *kick the can* for *kick the bucket*), with Moon (1998) positing the existence of *idiom schemas* that subsume alternative lexical realizations of idiomatic strings (e.g. *shake/quake/quiver in one's shoes/boots*). Nonetheless, this kind of lexical flexibility does not turn out to be so widespread, systematic and predictable as in literal constructions.

Proceedings of the 12th Workshop on Multiword Expressions, pages 21–31,
Berlin, Germany, August 7-12, 2016.

As we will briefly outline in Section 2, previous computational researches took advantage of the restricted formal variability exhibited by idioms to devise indices that automatically separate them from more literal combinations. Some of them have accomplished it by comparing the different collocational association between the canonical form of an expression and the lexical variants of that construction obtained by replacing its parts with semantically related words (Lin, 1999; Fazly et al., 2009). Others exploited the difference in cosine similarity between an entire phrase and its components that is observed in idioms and non-idioms in Distributional Semantics Models (DSMs) (Baldwin et al., 2003; Venkatapathy and Joshi, 2005; Fazly and Stevenson, 2008). Here, we combined insights from both the aforementioned approaches, using the generation of lexical variants as the departure point for a distributional semantic analysis. Compositional expressions exhibit systematicity (Fodor and Lepore, 2002) in that if a speaker can comprehend *spill the beans* as taken literally and *drop the peas*, he/she will also be able to understand *spill the peas* and *drop the beans*, but this does not happen if we read *spill the beans* as an idiom. The restricted lexical substitutability of a given construction could thus be regarded as a clue of its semantic non-compositionality and idiomatic status. To implement this idea, we generated a series of lexical variants from a set of target Italian V-NP and V-PP constructions, including both idioms and literals, but instead of measuring differences in the association scores between a given target and its variants, we computed the cosine similarities between them. Idiomatic expressions are expected to result less similar to their lexical variants with respect to literal ones.

2 Related work

Existing computational research on idiomaticity mainly splits into studies aimed at *idiom type identification* (i.e. separating potentially idiomatic constructions like *spill the beans* from only literal ones like *write a book*) and studies aimed at *idiom token identification* (i.e. distinguishing the idiomatic vs. literal usage of a given expression in context, e.g. *The interrogated man finally spilled the beans* vs. *The cook spilled the beans all over the kitchen floor*). Since in this paper we focus on the former issue, we only review related researches on idiom type identification.

Various techniques have been employed to separate idioms and non-idioms. McCarthy et al. (2003), for instance, focus on verb-particles constructions and find that thesaurus-based measures of the overlap between the neighbors of a phrasal verb and those of its simplex verb strongly correlate with human-elicited compositionality judgments given to the same expressions. Fixedness in the word order is exploited by Widdows and Dorow (2005), who observe that asymmetric lexicosyntactic patterns such as 'A and/or B' which never occur in the reversed order 'B and/or A' very often appear to represent idiomatic combinations. Bannard (2007) devises measures of determiner variability, adjectival modification and passivization to distinguish idiomatic and non-idiomatic VPs, resorting to conditional Pointwise Mutual Information (Church and Hanks, 1991) to calculate how the syntactic variation of a given V-N pair differs from what would be expected considering the variation of the single lexemes. In a similar way, Fazly et al. (2009) devise a syntactic flexibility index to single out V-NP idiomatic pairs that compares the behavior of a given pair to that of a typical V-N schema as regards the definiteness and the number of the noun and verbal voice. 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, while Graliński (2012) bases on metalinguistic markers such as *proverbially* or *literally* to retrieve idioms from the Web. Crucially for the present experiment, a series of studies have more precisely focused on lexical flexibility to identify non-compositional constructions. Among them, Lin (1999) classifies a phrase as non-compositional if the PMI between its components is significantly different from the PMI between the components of all its lexical variants. These variant forms are obtained by replacing the words in the original phrase with semantic neighbours. Fazly and Stevenson (2008) and Fazly et al. (2009) further elaborate on Lin's formula, regarding a certain V-N combination as lexically fixed and more likely to be idiomatic if its PMI highly differs from the mean PMI of its variants. Other contributions have employed distributional measures to determine the similarity between a given phrase and its components, observing that

idiomatic phrase vectors appear to be less similar to their component vectors than literal phrase vectors (Baldwin et al., 2003; Venkatapathy and Joshi, 2005; Fazly and Stevenson, 2008).

3 Measuring compositionality with variant-based distributional similarity

In the present work we propose a method for idiom type classification that starts from a set of V-NP and V-PP constructions, generates a series of lexical variants for each target by replacing the verb and the argument with semantically related words and then compares the semantic similarity between the initial constructions and their respective variants. For the sake of clarity, henceforth we will refer to the initial idiomatic and non-idiomatic expressions as *target* expressions, while the lexical alternatives that were generated for each target will be simply called *variants*. Since idiomatic expressions are supposed to exhibit a greater degree of non-compositionality and lexical fixedness than literal ones, with the substitution of their component words resulting in the impossibility of an idiomatic reading (e.g. *spill the beans* vs. *spill the peas*), we expected them to be less similar to their variants with respect to literal constructions. Starting from the assumption that we can study the semantics of a given word or expression by inspecting the linguistic contexts in which it occurs (Harris, 1954; Firth, 1957; Sahlgren, 2008), Distributional Semantic Models (DSMs) provide a viable solution for representing the content of our target and variant constructions with vectors recording their distributional association with linguistic contexts (Turney and Pantel, 2010). The semantic similarity between a given target and its variants is therefore implemented as the cosine similarity between them. Similarly to Lin (1999) and Fazly et al. (2009), we used lexical variants for each target expression, but instead of contrasting their associational scores, we used vector-based measures to grasp their degree of semantic compositionality.

3.1 Extraction of the target and variant constructions

45 Italian V-NP and V-PP idioms were selected from an Italian idiom dictionary (Quartu, 1993) and extracted from the itWaC corpus (Baroni et al., 2009), which consists of about 1,909M tokens. Their corpus frequency spanned from 364 (*ingannare il tempo* 'to while away the time') to 8294 (*andare in giro* 'to get about'). A set of 45 non-idioms (e.g. *leggere un libro* 'to read a book', *uscire da una stanza* 'to get out of a room') of comparable frequencies were then extracted from the corpus, ending up with 90 target constructions. Two different methods were explored for generating lexical variants from our targets:

DSM variants. For both the verb and argument component of each target construction, we extracted its 10 nearest neighbours (NNs) in terms of cosine similarity in a DSM created from the La Repubblica corpus (Baroni et al., 2004) (about 331M tokens); this space used all the content words (nouns, verbs, adjectives and adverbs) with token frequency > 100 as target vectors and the top 10,000 content words as contexts; the co-occurrence matrix, generated from a context window of $\pm$ 2 content words from each target word, was weighted by Positive Pointwise Mutual Information (PPMI) (Evert, 2008), a statistical association measure that assesses whether two elements x and y co-occur more frequently than expected by chance and sets to zero all the negative values:

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

The matrix was reduced to 300 latent dimensions via Singular Value Decomposition (SVD) (Deerwester et al., 1990). The variants were finally obtained by combining the verb with each of the 10 NNs of the argument, the argument with each of the 10 NNs of the verb and every NN of the verb with every NN of the argument. This resulted in 120 potential variants for each target expression, which were then extracted from itWaC.

iMWN variants. For both the verb and argument component of each target construction, the words occurring in same synsets and its co-hyponyms were extracted from the Italian section of MultiWordNet (iMWN) (Pianta et al., 2002). For each verbal head, we extracted 5.9 synonyms/co-hyponyms on average (SD = 5.41), while for the noun arguments we extracted 25.18 synonyms/co-hyponyms on average (SD = 27.45). The variants of the targets were then generated with the same procedure described for the distributionally derived variants and extracted from itWaC.

3.2 Collecting idiomaticity judgments

To provide the variant-based distributional measures with a gold standard, we collected idiomatic-

ity judgments for our 90 target expressions from Linguistics students. Nine undergraduate and graduate students were presented with a list of our targets and asked to evaluate how idiomatic each expression was on a 1-7 Likert scale. More specifically, we split our initial list into three sublists of 30 targets, each one being compiled by three subjects. Intercoder agreement, computed via Krippendorff's α (Krippendorff, 2012), was 0.83 for the first sublist and 0.75 for the other two. Following common practice, we interpreted these values as an evidence of reliability for the collected judgments (Artstein and Poesio, 2008).

4 Experiment 1

In the first experiment, we wanted to verify our predictions on a subset of our 90 target constructions that had a considerable number of variants represented in the corpus, so as to create reliable vector representations for them. We therefore selected those constructions that had at least 5 DSM and 5 iMWN variants occurring more than 100 times in itWaC. This selection resulted in a final set of 26 targets (13 idioms + 13 non-idioms).

4.1 Data extraction and method

Two DSMs were then built on the itWaC corpus, the first one representing the 26 targets and their DSM variants with token frequency > 100 as vectors, and the second one representing as vectors the 26 targets and their iMWN variants with token frequency > 100. Co-occurrences were recorded by counting 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 two matrices were weighted with PPMI and reduced to 300 dimensions via SVD.

Four different measures were tested to compute how much the vector representations of the targets differed from those of their respective variants:

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 the vector of a target construction and the vectors of its variants.

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

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

In both the DSMs, each of these four measures was computed for each of our 26 targets. We then sorted the targets in ascending order for each of the four scores, creating a ranking in which we expected idioms (our positives) to be placed at the top and non-idioms (our negatives) to be placed at the bottom, since idioms are expected to be less similar to the vectors of their lexical variants.

4.2 Results and discussion

The main goal of this study was to assess whether our variant-based method was suitable for identifying idiom types. Hence we evaluated the goodness of our four measures (Mean, Max, Min and Centroid) in placing idioms before non-idioms in the rankings generated by our idiomaticity indices.

Figures 1 and 2 plot the Interpolated Precision-Recall curves for the four measures in the two trained DSMs plus a random baseline. In the DSM variants model, Max, Mean and Centroid performed better than Min and the baseline. Max showed high precision at low levels of recall ($<$ 40%), but it dropped as far as higher recall levels were reached, while Mean and Centroid kept higher precision at higher levels of recall. Min initially performed comparably to Mean, but it drastically dropped after 50% of recall.

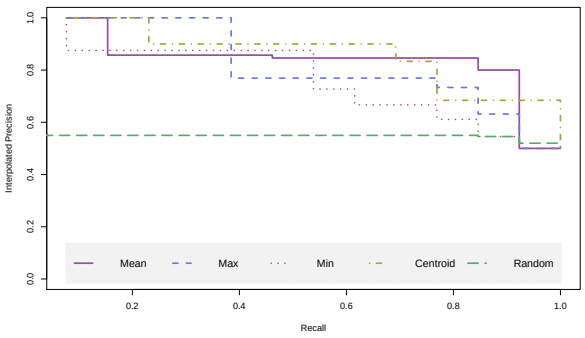

Figure 1: Interpolated Precision-Recall curve for Mean, Max, Min, Centroid and the baseline in the DSM variants space with 26 targets.

In the iMWN variants space both Mean and Centroid performed better than the other measures, with the baseline being the worst one. Both Max and Min exhibited the same pattern, with high precision at low recall levels and a subsequent drop in performance around 50% of recall.

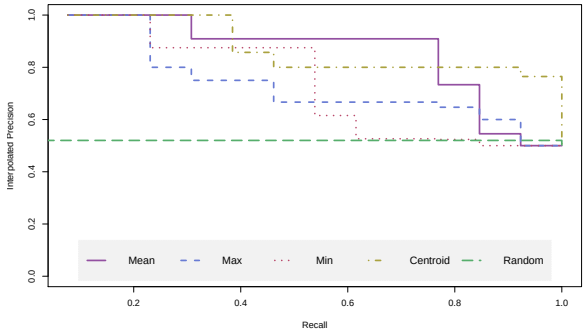

Figure 2: Interpolated Precision-Recall curve for Mean, Max, Min, Centroid and the baseline in the iMWN variants space with 26 targets.

The first two columns of Table 1 show the Interpolated Average Precision (IAP) and the F-measure of all the models employed in this first experiment. Interpolated Average Precision consists in the average of the interpolated precisions at recall levels of 20%, 50% and 80%, while F-measure is computed for the median. Both iMWN and DSM Mean and Centroid, together with DSM Max, had the highest IAPs, therefore standing out as the models the suceeded the most in placing idioms before non-idioms in the obtained rankings and exhibited the best trade-off between precision and recall, as shown by the F-measure values. The third column in Table 1 shows Spearman's ρ correlation between our models and the speaker-elicited idiomaticity judgments we described in Section 3.2. The Mean and the Centroid similarity in both the DSM and the iMWN variants spaces and the Max similarity in the DSM variants spaces showed a significant strong negative correlation with the speaker-collected ratings: the less the vector of a given expression resulted similar to the vectors of its lexical variants, the more the subjects perceived the expression as idiomatic. iMWN Min, DSM Min and iMWN Max exhibited a weak, non-significant, negative correlation, while the baseline showed a non-significant weak positive correlation score. All in all, Centroid and Mean turned out as the best measures in separating idioms from non-idioms, while there was no clear advantage of one variant type (DSM or iMWN) over the other.

5 Experiment 2

The first experiment proved our variant-based distributional measures to be suitable for telling apart idioms and non-idioms that had a fair number of lexical variants occurring in our corpus with considerable frequency, with the Mean and the Centroid measures performing the best. The research question at the root of the following experiment was whether such measures could be extended to all the 90 target constructions in our dataset (45 idioms + 45 non-idioms), including expressions whose lexical variants were poorly represented or not at all found in itWaC. Such negative evidence, in our reasoning, should be taken into account as an additional clue of the restricted lexical transformability of the expressions at hand and, consequently, of their idiosyncratic and idiomatic status.

Model	IAP	F	ρ
DSM Centroid	.83	.77	**-.66*****
iMWN Centroid	**.87**	.77	-.59**
DSM Mean	.80	**.85**	-.63***
iMWN Mean	.80	.77	-.58**
DSM Max	.74	.77	-.60**
iMWN Max	.68	.62	-.30
DSM Min	.69	.62	-.37
iMWN Min	.65	.62	-.28
Random	.53	.46	.30

Table 1: Interpolated Average Precision, F-measure at the median and Spearman's ρ correlation with the speaker judgments for the models with 26 targets ($^{**} = p < .01$, $^{***} = p < .001$).

5.1 Data extraction and method

As in the first experiment, two kinds of DSMs were built from itWaC, the former comprising the 90 initial idiomatic and non-idiomatic expressions and their DSM variants as target vectors and the latter considering the 90 expressions and their iMWN variants as target vectors. The parameters of these vector spaces are identical to those used in Experiment 1. The vectors of the targets were compared to the vectors of their variants by means of the four measures described in Section 4.1 (Mean, Max, Min, Centroid). Aside from the method chosen to extract the variants (DSM vs. iMWN), the parameter space explored in constructing the DSMs for the second experiment further comprised the following options:

Number of variants per target. For both the variants that were extracted distributionally and those that were chosen from iMWN, we built different DSMs, each time setting a fixed number of alternative forms for each target expression. As for

the DSM-generated variants, we kept the alternative expressions that were generated by combining the top 3, 4, 5 and 6 cosine neighbours of each verb and argument component of the initial 90 targets. As a result, we obtained 4 types of spaces, in which each target had respectively 15, 24, 35 and 48 variants represented as vectors. As for the spaces built with the iMWN variants, we experimented with eight types of DSMs. In the first four, we kept the variants that were created by combining the top 3, 4, 5 and 6 synonyms and co-hyponyms of each component of the initial 90 targets in terms of cosine similarity. These cosine similarities were extracted from a DSM trained on the La Repubblica corpus that had the same parameters as the space used to extract the DSM variants and described in Section 3.1. In the other four, we used the top 3, 4, 5 and 6 synonyms and co-hyponyms that were most frequent in itWaC.

Encoding of non-occurring variants. In each of the DSMs obtained above, every target was associated with a fixed number of lexical variants, some of them not occurring in our corpus. We experimented with two different ways of addressing this problem. In the first case, we simply did not take them into account, thus focusing only on the positive evidence in our corpus. In the second case, we represented them as orthogonal vectors to the vectors of their target. For the Mean, Max and Min measures, this merely consisted in automatically setting to 0.0 the cosine similarity between a target and a non-attested variant. For the Centroid measure, we first computed the cosine similarity between the vector of a target expression and the centroid of its attested variants and then hypothesized that each zero variant contributed by a costant factor k in tilting this centroid similarity towards 0.0. Preliminary investigations have proved a k-value of 0.01 to give reliable results. We leave to future contributions the tuning of this parameter, limiting ourselves to propose and test this centroid-based measure for the present work. Concretely, from the centroid similarity computed with the attested variants (cs_a), we subtracted the product of k and cs_a multiplied by the number of non-attested variants (n) for the construction under consideration, obtaining a final centroid similarity that also includes non-attested variants:

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

Crucially, the rationale behind multiplying k by the original centroid similarity lies in the fact that non-attested variants were not expected to contribute in modifying the original cosine value towards zero always in the same way, but depending on the specific target construction at hand and on the positive evidence available for it.

Table 2 summarizes the parameters explored in building the DSMs for the second experiment. In each model resulting from the combination of these parameters, we ranked our 90 targets in ascending order according to the idiomaticity scores given by the four variant-based distributional measures (Mean, Max, Min, and Centroid).

Parameter	**Values**
Variants source	DSM, iMWN
Variants filter	cosine (DSM, iMWN), raw frequency (iMWN)
Variants per target	15, 24, 35, 48
Non-attested variants	not considered (*no*), orthogonal vectors (*orth*)
Measures	Mean, Max, Min, Centroid

Table 2: Parameters explored in creating the DSMs for Experiment 2.

5.2 Results and discussion

All the 96 models obtained by combining the parameters in Table 2 had higher IAP and F-measure scores than the random baseline, with the exception of two models displaying lower (iMWN$_{cos}$ 35_{var} Centroid$_{orth}$) or comparable (iMWN$_{freq}$ 15_{var} Centroid$_{orth}$) F scores. All the models had significant correlational scores with the human-elicited ratings save 7 non significant models.

Table 3 reports the 5 best models for IAP, F-measure at the median and Spearman's ρ correlation with our gold standard idiomaticity judgments respectively. All the best models predictably employed the Centroid measure, which already turned out to perform better than the other indices in the first part of our study. The best performance in placing idioms before non-idioms (IAP) and the best trade-off between precision and recall (F-measure) were exhibited both by models that considered (*orth*) and not considered (*no*) non-attested variants, with a prevalence of the latter models. Moreover, the top IAP and top F-measure models used both DSM and iMWN variants. On the other hand, the models correlating

the best with the judgments all took non-occurring variants into account as orthogonal vectors and all made use of iMWN variants. There seemed not to be an effect of the number of variants per target across all the three evaluation measures.

Top IAP Models	**IAP**	**F**	ρ
iMWN$_{cos}$ 15$_{var}$ Centroid$_{no}$	.91	.80	-.58***
iMWN$_{cos}$ 24$_{var}$ Centroid$_{no}$	.91	.78	-.62***
iMWN$_{cos}$ 35$_{var}$ Centroid$_{no}$	.91	.82	-.60***
DSM 48$_{var}$ Centroid$_{no}$	.89	.82	-.64***
DSM 48$_{var}$ Centroid$_{orth}$	.89	.82	-.60***
Top F-measure Models	**IAP**	**F**	ρ
iMWN$_{cos}$ 35$_{var}$ Centroid$_{no}$	.91	.82	-.60***
DSM 48$_{var}$ Centroid$_{no}$	.89	.82	-.64***
DSM 48$_{var}$ Centroid$_{orth}$	.89	.82	-.60***
iMWN$_{cos}$ 15$_{var}$ Centroid$_{no}$	.91	.80	-.58***
DSM 24$_{var}$ Centroid$_{no}$	.89	.80	-.60***
Top ρ Models	**IAP**	**F**	ρ
iMWN$_{cos}$ 48$_{var}$ Centroid$_{orth}$	.86	.80	-.67***
iMWN$_{cos}$ 35$_{var}$ Centroid$_{orth}$	.72	.44	-.66***
iMWN$_{cos}$ 24$_{var}$ Centroid$_{orth}$	.85	.78	-.66***
iMWN$_{cos}$ 15$_{var}$ Centroid$_{orth}$	.88	.80	-.65***
iMWN$_{freq}$ 15$_{var}$ Centroid$_{orth}$	.66	.51	-.65***
Random	.55	.51	.05

Table 3: Best 5 models with 90 targets for IAP (top), F-measure at the median (middle) and Spearman's ρ correlation with the speaker judgments (bottom) against the random baseline (*** = $p < .001$).

After listing the best overall models for each evaluation measure, we resorted to linear regression to assess the influence of the parameter settings on the performance of our models, following the methodology proposed by Lapesa and Evert (2014). As for the IAP and correlation with human judgments, our linear models achieved adjusted R^2 of 0.90 and 0.94 respectively, therefore explaining the influence of our parameters and their interactions on these two evaluation measures very well. In predicting F-measure, our linear model reported an adjusted R^2 of 0.52. Figure 3 depicts the rankings of our parameters according to their importance in a feature ablation setting. The ΔR^2 values can be understood as a measure of the importance of a parameter, and it is calculated as the difference in fit that is registered by removing the target parameter together with all the pairwise interactions involving it from our full models.

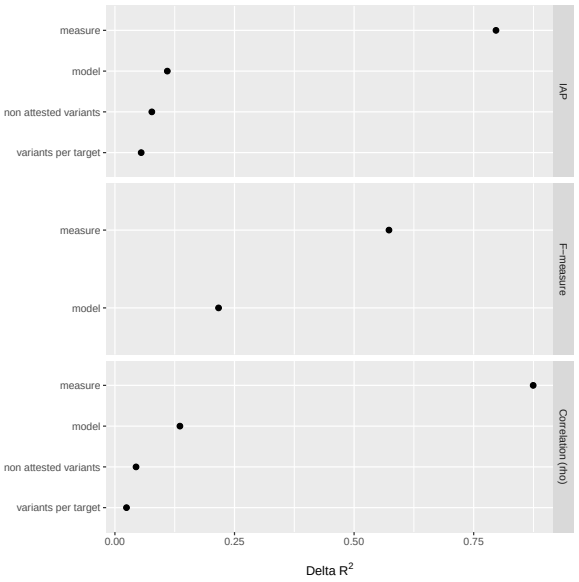

Figure 3: Parameters and feature ablation for IAP, F-measure and correlation with the human ratings.

The parameters we refer to are the same listed in Table 2, with the exception of the parameter *model*, which merges the *variants source* and the *variants filter* parameters. For all our three evaluation measures, *measure* (i.e. Mean, Max, Min vs. Centroid) turned out to be the most influential parameter, followed by *model* (i.e. DSM, iMWN$_{cos}$ vs. iMWN$_{freq}$). As for the *measure* parameter, both in the IAP and in the ρ models the best performing setting is Centroid, followed by Mean, Max and Min, all being significantly different from each other. In the F-measure model, only Min, i.e. the worst performing model, was significantly different from the other settings. As for *model*, the iMWN$_{freq}$ setting was significantly worse than DSM and iMWN$_{cos}$ in the IAP and in the ρ models, but not in the F-measure one.

Table 4 reports all the significant pariwise interactions and their ΔR^2. In line with results reported in Figure 3, almost all the interactions involved the *model* parameter.

Interaction	ΔR^2		
	IAP	**F**	ρ
model:measure	.03	.13	.08
model:non-attested var	.01	$n.s.$	.02
non-attested var:measure	.02	$n.s.$	.01
model:variants per target	.02	$n.s.$	$n.s.$

Table 4: Significant interactions and ΔR^2 for IAP, F-measure and correlation with the human ratings.

Figure 4 displays the interaction between *measure* and *model* when modeling IAP. The best models, DSM and iMWN$_{cos}$, had a different performance on the worst measure (Min) but converged on the two best ones (Mean and Centroid). On the other side, iMWN$_{freq}$ showed a less dramatic improvement and reached a plateau after moving away from the Min setting.

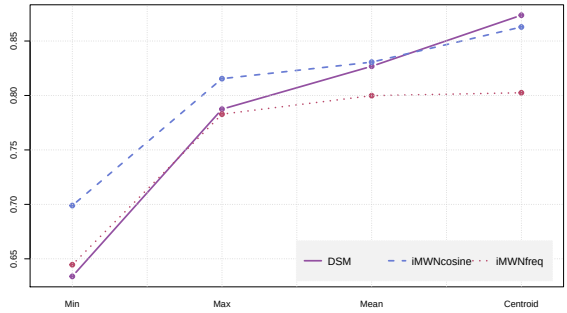

Figure 4: IAP, measure / model.

Figure 5 shows that in the F-measure setting the DSM model had a steeper improvement when moving from Min to the other measures, as compared to the iMWN$_{cos}$ and the iMWN$_{freq}$ models.

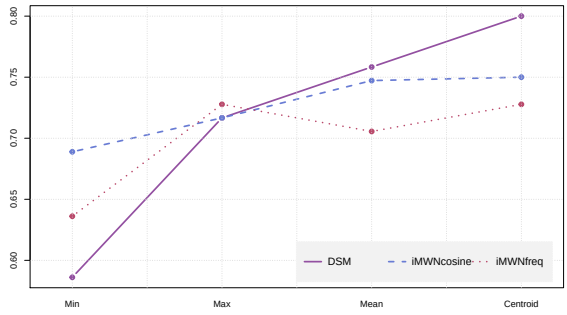

Figure 5: F-measure, measure / model.

Figure 6 shows that in the correlation setting the iMWN$_{cos}$ and the DSM models outperformed the iMWN$_{freq}$ model only when exploiting the Min and the Mean measures. It is worth remarking that the correlational scores with the human ratings are negative and therefore points that are positioned lower on the *y*-axis indicate better performance.

Figures 7 and 8 plot the interaction between *model* and the way of encoding *non-attested variants* in the IAP and in the ρ models, respectively. In both cases, only the two iMWN models appeared to be sensitive to the way non-attested variants are handled. In the IAP model, zero variants appeared to be the outperforming setting, while the ρ model showed the opposite pattern. In both

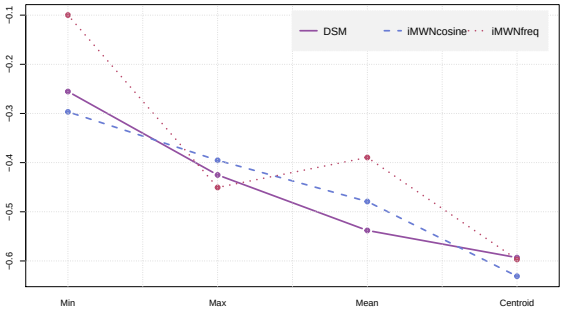

Figure 6: ρ, measure / model.

models, moreover, the best overall setting always involve the iMWN$_{cos}$ model.

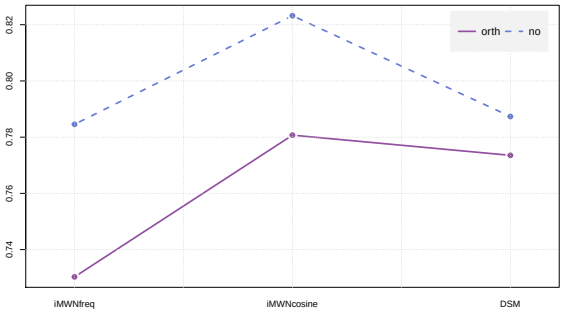

Figure 7: IAP, model / non-attested variants.

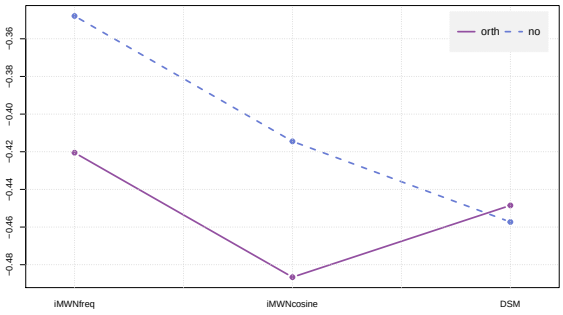

Figure 8: ρ, model / non-attested variants.

Figures 9 and 10 display the interactions between *measure* and the way of encoding *non-attested variants*. In the IAP model, ignoring the non-attested variants resulted in a significantly better performance only when using the Max and Centroid measures. In the ρ model, however, accounting for the effects of non-attested variants outperformed the other setting only when using the Min and Mean measures.

The interaction between the number of *variants per target* and the *model* when modeling IAP is displayed in Figure 11. We observed a strong effect of the variants number on the performance of

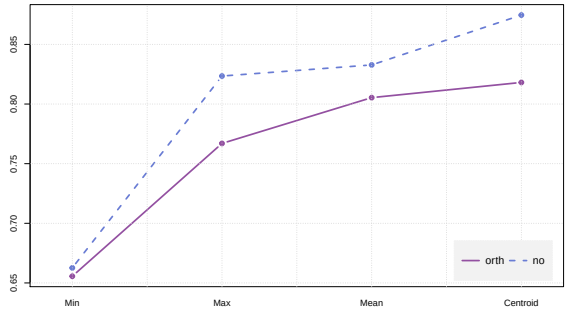

Figure 9: IAP, measure / non-attested variants.

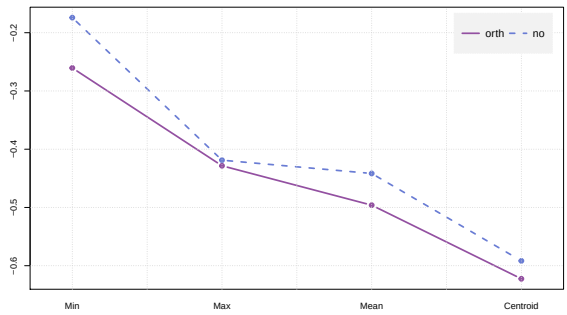

Figure 10: ρ, measure / non-attested variants.

iMWN$_{freq}$, with more variants leading to a better performance. There was a significant advantage of iMWN$_{cos}$ over the other models when using 15 variants, but this advantage was lost as the number of variants increased.

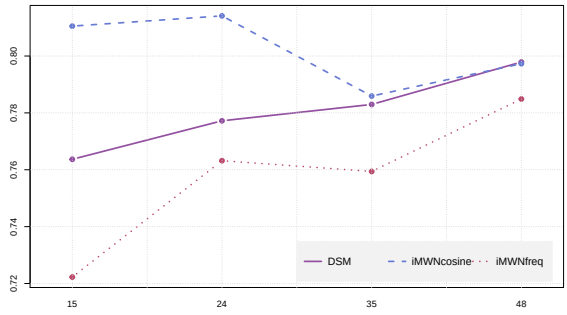

Figure 11: IAP, variants per target / model.

All in all, the Centroid measure appeared to perform better than the other three measures, with Min obtaining the worst results. The DSM and the iMWN$_{cos}$ models performed consistently better than iMWN$_{freq}$, while the advantage of either way of encoding non-attested variants (*no* vs. *orth*) over the other depended on the evaluation setting. Finally, the number of variants per target did not appear to consistently influence the performance of our models.

Error Analysis. A qualitative inspection of the data revealed that the most frequent false positives (i.e. non-idioms classified as idioms) include expressions like *giocare a carte* ('to play cards') or *mostrare interesse* ('to show interest'). Despite being literal and compositional, these word combinations display some form of collocational behavior, being less lexically free than the other literal combinations. Conversely, among the most common false negatives (i.e. idioms that were classified as non-idioms), we find expressions like *cadere dal cielo* ('to fall from the sky, to be heaven-sent') or *aprire gli occhi* ('to open one's eyes') that happen to be highly ambiguous in that they make both an idiomatic and a literal reading possible according to the context. It is possible that the evidence available in our corpus privileged a literal reading for them. Such ambiguous expressions should be analyzed in more detail in following contributions by means of *token detection* algorithms that might tell apart idiomatic and literal usages of these expressions in context.

6 Conclusions

In this paper we carried out an idiom type identification task based on the idea that idiomatic expressions tend to allow for more restricted variability in the lexical choice of their subparts with respect to non-idiomatic ones. Starting from a list of target Italian V-NP and V-PP constructions, comprising both idioms and non-idioms, we generated a set of lexical variants by replacing their components with semantically related words extracted distributionally or from Italian MultiWordNet. We then measured the cosine similarity between the vectors of the original expressions and the vectors of their variants, expecting idioms to be less similar to their variants with respect to non-idioms. All in all, this proved to be the case. More specifically, cosine similarity between the vector of the original expressions and the centroid of their variants stood out as the best performing measure. The best models used DSM variants or iMWN variants filtered by their cosine similarity with the components of the target expressions. In the second place, our methods proved to be successful also when applied to idioms most of which had many scarcely or not at all attested variants. In devising our variant-based distributional idiomaticity measures we also tried to take this negative evidence into consideration, still achieving high and reliable performances.

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.

Colin Bannard. 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.

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 and Sam Glucksberg. 1991. Understanding idiomatic expressions: The contribution of word meanings. *Advances in Psychology*, 77:217–240.

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

Noam Chomsky. 1980. Rules and representations. *Behavioral and Brain Sciences*, 3:1–15, 3.

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.

Patrick J. Duffley. 2013. How creativity strains conventionality in the use of idiomatic expressions. In Mike Borkent, Barbara Dancygier, and Jennifer Hinnell, editors, *Language and the creative mind*, pages 49–61. CSLI Publications.

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 / Rivista di Linguistica*, 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.

John R. Firth. 1957. *Papers in Linguistics*. Oxford University Press.

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

Raymond W. Gibbs, Nandini P. Nayak, John L. Bolton, and Melissa E. Keppel. 1989. Speakers' assumptions about the lexical flexibility of idioms. *Memory & Cognition*, 17(1):58–68.

Adele E. Goldberg. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Filip Graliński. 2012. Mining the web for idiomatic expressions using metalinguistic markers. In *Proceedings of Text, Speech and Dialogue: 15th International Conference*, pages 112–118.

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

Thomas Hoffmann and Graeme Trousdale, editors. 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press.

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

Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.

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.

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

Matthew S. McGlone, Sam Glucksberg, and Cristina Cacciari. 1994. Semantic productivity and idiom comprehension. *Discourse Processes*, 17(2):167–190.

Rosamund Moon. 1998. *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.

Grace Muzny 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.

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

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.

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

Ivan A. Sag, 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.

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

John Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press.

Anna Siyanova-Chanturia and Ron Martinez. 2014. The idiom principle revisited. *Applied Linguistics*, pages 1–22.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

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.

Dominic Widdows 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.