Modeling Idiom Variability with Entropy and Distributional Semantics

Gianluca E. Lebani[†], Marco S. G. Senaldi[†‡], Alessandro Lenci[†]
[†]University of Pisa, Pisa, Italy
[‡]Scuola Normale Superiore, Pisa, Italy
gianluca.lebani@for.unipi.it, marco.senaldi@sns.it, alessandro.lenci@unipi.it

***Abstract*—In this paper we assessed the cognitive plausibility of corpus-based flexibility measures for a sample of Italian idioms taken from the normative data by Tabossi et al. (2011). We found that psycholinguistic judgments on idiom predictability, literality and syntactic flexibility can be modeled by distributional representations of idiom semantics, entropy-based formal flexibility measures, frequency and the number of fully lexicalized arguments of idioms.**

I. Introduction and Related Work

Idioms ([1], [2], [3]) are mainly characterized by non-compositionality, restricted formal flexibility, figurativity and proverbiality ([4], [5]). Due to their challenging nature for any model of grammar ([6], [7]), they have been the subject of intensive (neuro)cognitive and computational linguistic studies. Most computational researches have focused on *type* and *token identification*, the former consisting in separating potentially idiomatic constructions, like *spill the beans*, from expressions that can only have a literal meaning, like *write a letter* ([8], [9], [10], [11], [12], [13], [14]), and the latter consisting in telling apart idiomatic and literal uses of a given token expression in context ([15], [16], [17], [14], [18], [19]). These and other studies focused on *Multiword Expressions* in general ([20], [21]), exploit the linguistic features that are considered typical of these kinds of constructions by the theoretical literature, like their restricted lexical exchangeability, the limited presence of modifying adjectives and PPs, and the constrained occurrence of syntactic variants (e.g. passivized or dislocated forms), for classificational purposes.

More similarly to the present work, other quantitative studies have tried to verify whether corpus-based flexibility measures identical or similar to those exploited by the aforementioned studies actually exhibit psycholinguistic validity. In a broader sense, this means to observe whether these indices can generally model and predict speaker-elicited idiomaticity judgments. More specifically, it also means to detect the single measures that have the greatest weight in modeling these judgments among a large set of factors. Wulff [22] devises computational indices of compositionality and syntactic flexibility for a sample of English V-NP constructions, including both idiomatic and non-idiomatic ones. Compositionality is computed with a collocation-based index that compares the number of collocates shared by a construction with those shared by its component words, while morphosyntactic flexibility is captured by comparing the distributional behavior of a target V-NP construction with that of a typical V-NP construction along a series of morphosyntactic variational dimensions (e.g. verbal morphology, presence and type of adverbs, presence and type of determiners, etc.). These corpus statistics are then used as predictors in a regression analysis with human idiomaticity ratings assigned to the same constructions. Corpus statistics pertaining to verbal morphology turn out to have the greatest weight in modeling the human judgments. Interestingly, this findings stand at odds with the theoretical assumption that idiom morphosyntactic idiosyncrasy is more evidently manifested by the idiomatic arguments rather than the verb, which can normally inflect in tense and mood just like in literal expressions [23]. One of the aims of our experiments with different corpus-based measures was indeed to investigate the importance of verb-related parameters *vis-à-vis* the role of other dimensions identified in the theoretical literature.

Quantitative analyses have essentially addressed the contrast between idiomatic and non-idiomatic expressions. However, it is well known that idioms form a very heterogeneous class, which greatly varies in terms of syntactic flexibility and semantic transparency ([5], [24]). These degrees of variations deeply affect the processing and representation of idiomatic expressions ([25]). Nunberg *et al.* distinguish *idiomatic phrases* like *kick the bucket* that are strictly non compositional, and *idiomatic combinations* like *pop the question*, for which we can detect a direct or metaphorical mapping between the idiom components and the idiomatic referent components. Although this semantic decomposability has long been associated with greater syntactic flexibility ([5], [26]), other works have nonetheless stressed that almost all kinds of idioms are formally flexible if an appropriate context is provided ([4], [27], [28]). Therefore, the factors that are most relevant for the speakers to assess the idiomaticity and the formal flexibility of a construction still need to be clearly identified.

In light of this debate, the aim of the present paper is to carry out an extensive analysis of the cognitive plausibility of corpus-based idiom flexibility measures, by analyzing a sample of 87 Italian idioms, taken from the descriptive norms by Tabossi et al. [24]. This sample includes both fully lexicalized idiomatic expressions like *gettare acqua sul fuoco* 'to minimize', and partially lexicalized ones like *andare a genio a NP* 'to sit well with NP', which have an open syntactic slot. These idioms were automatically extracted from a corpus with SYMPAThy ([29], [30]), a data representation format that includes various kinds of morphosyntactic information to describe the combinatory potential of verbal and nominal lexemes. For each idiom, we computed a *variational profile* consisting of entropy-based measures for lexical and morphosyntactic variability and distributional representations to capture the idiom semantics. We then applied stepwise multiple regression to find out what

kind of corpus-driven measures are most useful to explain psycholinguistic judgments on idioms. Entropic indexes, cosine semantic similarity and frequency were used as predictors of human ratings on idiom predictability, literality and syntactic flexibility collected in [24].

II. Assessing the cognitive plausibility of corpus-based idiomaticity measures

A. SYMPAThy: a computational method to extract and represent word combinations

Idioms contribute to define the combinatory potential of a target lexeme (TL). By word combinations we broadly refer to the range of constructions typically associated with a lexical item. In Construction Grammar, constructions (Cxn) are conventionalized form-meaning pairings that can vary in both complexity, schematicity and idiomaticity ([6], [31]). Word combinations can be defined and observed at a more constrained, surface POS-pattern level (P-based), *and* at the more abstract level of syntactic structures (S-based). These two levels are often kept separate, not only theoretically, but also computationally, as their performance varies according to the different types of combinations that we want to track.

S-based methods are most suitable to identify the subcategorization frames TLs occur in, the lexical items (fillers) typically appearing in each frame slot, and the selectional preferences constraining the semantic type of fillers. On the flip side, the difference between a literal string like *non vedere l'uscita* 'not to see the exit' and a syntactically identical but idiomatic one like *non vedere l'ora* 'to look forward to something' can only be captured by a POS-based method, which, focusing on the surface pattern "Neg V Det N" would detect a stronger association between the words in the second string. Anyway, such a method alone would in turn miss higher-level generalizations, such as the fact that there are partially lexicalized idioms with open syntactic slots.

To overcome such limitations, SYMPAThy (SYntactically Marked PATerns) [29], [30] acts as a distributional knowledge representation that combines P-based and S-based information. First of all, the sentences containing a TL are extracted from a version of the "la Repubblica" corpus [32] (about 331M tokens), POS tagged with the Part-Of-Speech tagger described in [33] and dependency parsed with DeSR [34]. For each sentence and for each terminal node that depends on it we extract the following P-based information: (i) its lemma; (ii) its POS tag; (iii) its morphosyntactic features (gender and number for nouns and adjectives, person, number, tense and mood for verbs): (iv) its linear distance from the TL; (v) the dependency path linking it to the TL. This information is represented in a pattern that preserves the linear order of the words in the sentence. Then, to capture S-based information we extract: (i) the subcategorization frame of the TL; (ii) the fillers occurring in the frame slots. Statistical measures are then applied to P-based and S-based information to determine a *variational profile* for each construction the TL occurs in, which summarizes 1) the variability of the fillers that instantiate the syntactic slots of the construction; 2) the morphological variability of the TL and of its fillers; 3) the variability of the fillers definiteness; 4) the variability in the presence of adjectives and PPs modifying the slots, or adverbs modifying the TL; 5) the variability in the linear order of the slots with respect to the TL. Each of these values capture a different aspect of the lexical and morphosyntactic flexibility of constructions.

B. Measuring idiom flexibility with entropy

Shannon entropy measures the average uncertainty of a random variable X:

$$H(X) = -\sum_{x \in X} p(x) \log_2(p(x)) \tag{1}$$

For each dimension of variation below, we replace the variable X with the Cxn of interest and take the states of the system x as its values on that dimension. The higher the entropy, the higher the variability of Cxn along a particular lexical or morphosyntactic dimension. Noteworthily, we cannot compare the entropic values of different variational dimensions for a given idiom, nor the entropies of a specific dimension across all the idioms, since each dimension has a different number of states. Therefore, we follow Wulff [35] in using relative entropy, computed as the ratio between the observed entropy for the variable X and the maximum entropy H_{max} for X ($|X|$ is the number of states of X):

$$H_{rel}(X) = \frac{H(X)}{H_{max}(X)} = \frac{H(X)}{\log_2(|X|)} \tag{2}$$

Hence we obtain a flexibility index ranging from 0 to 1 for each of the following dimensions of variation:

Lexical Entropy. If the idiomatic Cxn has free slots, this index estimates the variability of their fillers. For instance, for *gettare#luce#su_*X 'to cast light on X' the open PP slot can be filled with *vicenda* 'affair', *mistero* 'mistery', etc.

Morphological Entropy. This index measures the morphological variability of the slots of an idiomatic Cxn: for *gettare#ombra#su* 'to cast shadow on', we can have *gettare#ombra-fs* 'cast shadow-singular' and *gettare#ombra-fp* 'cast shadow-plural'. For free slots, we only consider the different combinations of morphological features they can display, regardless of the fillers (e.g., *gettare#ombra-fs#su_vicenda-fs* and *gettare#ombra-fs#su_questione-fs* would count as two instances of the same state).

Verbal Morphological Entropy. This index measures the morphological variability of the TL.

Articles Entropy. This index measures the variation in the presence or absence of articles specifying the slots in a Cxn, and, if appropriate, their type (DEFinite vs. INDefinite): for instance, *gettare#*∅*+acqua#*su_*DEF+fuoco*.

Modifiers Entropy. This index measures the variability in the presence of adjectives and PPs modifying the slots of an idiom. For a Cxn like *gettare#acqua#su_fuoco* 'to minimize' (lit. 'to throw water on the fire'), possible states are *gettare#molta+acqua#*su_∅*+fuoco* 'to minimize a lot' (lit. 'to throw a lot of water on the fire') or *gettare#molta+acqua#*su_*fuoco+di_polemica* 'to minimize the controversies a lot' (lit. 'to throw a lot of water on the controversies fire').

VERBAL MODIFIERS ENTROPY. This index measures the variability of the TL with respect to adverbial and PP modifiers.

ORDER ENTROPY. This index measures the order variability of the slots with respect to the TL (e.g. *acqua#gettare#su_fuoco*, *su_fuoco#gettare#acqua*, etc.).

C. Capturing idiom semantics with distributional vectors

A crucial aspect of idioms is their idiosyncratic semantic behavior. In particular, they differ for the identifiability of the meaning of their parts, as well as for the availability of a literal interpretation, besides the idiomatic one. While the entropic indexes defined above explore the formal behavior of an idiom, we used distributional semantics to estimate how the usage of an idiom in context diverges from the prototypical usage of its components. Distributional Semantic Models (DSMs) represent the content of lexemes with vectors containing their distributional statistics with linguistic contexts ([36]). We represented idioms and their components with vectors recording co-occurrences with the content words (noun, verbs, adjectives and adverbs) appearing within the same sentence. In our models we considered the first 10,000 dimensions. Co-occurrences were also extracted from "la Repubblica" corpus with SYMPAThy. Then, we calculated the average cosine similarity between an idiom vector and each of its constituent word vectors (for a similar approach, cf. [20]). We trained two DSMs, PPMI and PLMI, by weighting co-occurrences respectively with Positive Pointwise Mutual Information and with Positive Local Mutual Information [37]. The former association measure is more biased towards low-frequency co-occurrences, while the latter favors high frequency ones. In both cases, the weighted co-occurrence matrix was reduced to 300 dimensions with Singular Value Decomposition (SVD), before measuring cosine similarities.

D. Basic idiom statistics

Besides the measures above, we also used the following basic statistics of the idioms:

LOG FREQUENCY. The logarithm of the raw frequency of an idiom.

LOG RELATIVE FREQUENCY. The logarithm of the ratio between the raw frequency of an idiom and the raw frequency of its verb head.

FIXED ARGUMENTS NUMBER. The number of fixed (i.e. fully lexicalized) arguments of an idiom.

E. The normative data by Tabossi et al. (2011)

Tabossi et al. [24] elicited normative judgments for 245 Italian verbal idioms from a group of 740 native subjects. For each idiomatic expression, they collected at least 40 judgments on a series of psycholinguistically relevant variables. In this paper, we focused our regression study to model three of them:

PREDICTABILITY. The proportion of idiomatic completions given for a certain idiom, which is presented to the subjects in an example sentence and with the final word missing.

LITERALITY. The plausibility of a literal interpretation for an idiom. For instance, *Perdere il treno* 'to miss an opportunity' (lit. 'to miss the train') has also a clear literal meaning beside the figurative one, while *andare in rosso* 'to go into the red' does not have a plausible literal meaning and can only be idiomatically interpreted.

SYNTACTIC FLEXIBILITY. Each idiom was inserted in a sentence containing one of the following five syntactic modifications: *adverb insertion*, *adjective insertion*, *left dislocation*, *passivization* and *movement*. Participants evaluated how much the meaning of the idiom in the syntactically modified version was similar to its unmarked meaning, expressed in the form of a paraphrase prepared by the authors.

III. IDIOM EXTRACTION AND ANALYSIS

We have analyzed a sample of 87 idioms from Tabossi et al. [24], consisting of 60 fully lexicalized idioms, and 27 with open slots. The data were automatically extracted with the following procedure:

1) for each verbal TL appearing in the idiom list by Tabossi et al., we extracted its SYMPAThy patterns and subcategorization frames from a dependency parsed version of "la Repubblica" corpus;
2) the frames corresponding to our target idioms were identified and selected (e.g. *gettare*#obj:*spugna* for *gettare la spugna* 'to throw in the towel');
3) idioms with frame frequency < 75 were discarded, in order to filter out statistically unreliable data. This threshold has been empirically estimated, and downsized our final dataset to 87 verbal idioms;
4) for each idiom, we calculated the basic statistics, the entropic scores and the distributional semantic measures described in Section II. In our dataset, we distinguish between idioms having free slots, for which we calculate also the *lexical entropy* (*H_{lex} idioms*), and fully lexically specified ones, for which this index is not computable (*No-H_{lex} idioms*).

Using our entropic, cosine and basic statistics as predictors, we have run a distinct stepwise multiple regression analysis for each psycholinguistic variable of interest as a dependent variable: Predictability, Literality, and Syntactic Flexibility. All predictors were mean-centered to ensure more reliable parameter estimation, and human ratings were standardized. The analyses were carried out for the *H_{lex} idioms* and the *No-H_{lex} idioms* separately, obtaining the six models that are described below.

IV. RESULTS AND DISCUSSION

The dendrogram in Figure 1, obtained with R [38], shows the correlational structure of our predictors. To obtain such a dendrogram, we first extracted the correlation matrix for our predictors using Spearman's ρ. The elements of this matrix were then squared to obtain a similarity metric that was sensitive to many types of dependence, including non-monotonic relationships. Divisive hierarchical clustering was then performed on the resulting matrix. The obtained clusters

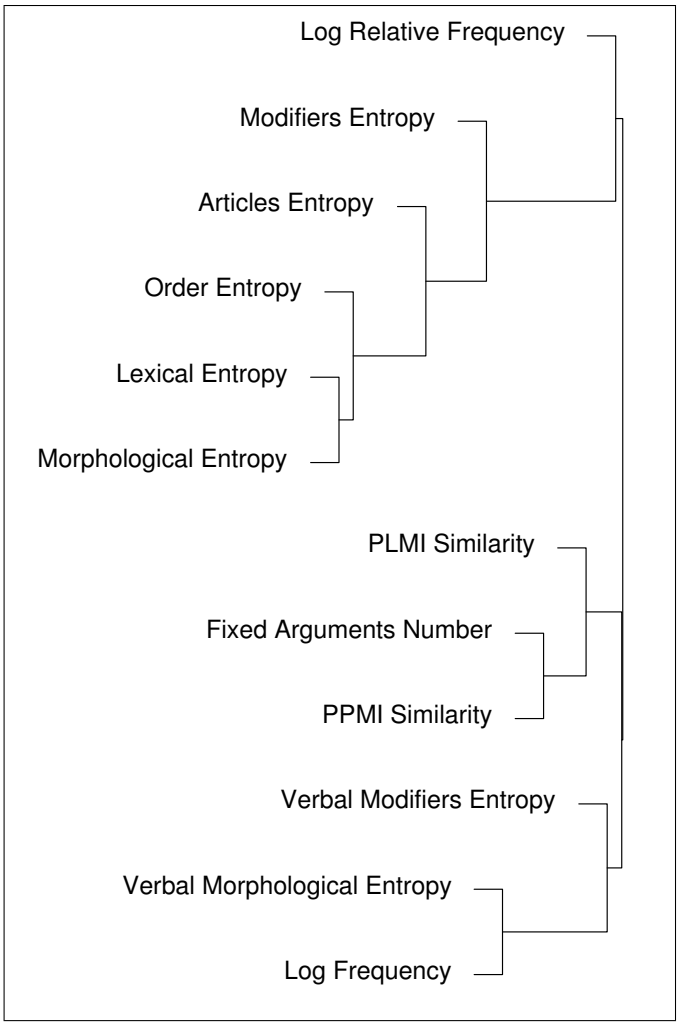

Fig. 1. Divisive clustering analysis of our predictors, using Spearman's ρ^2 as similarity metric

mirrored our theoretical assumptions. Nearly all the entropic measures clustered together, as well as the two distributional indices. The appearance of Fixed Arguments Number together with the cosine measures may be due to the fact that the latter are influenced by the number of slots to be averaged. Interestingly, the verbal entropies clustered with Log Frequency. In effect, we expect that the more often an idiom occurs, the more different are the morphosyntactic contexts in which its verb appears. Interestingly, Log Frequency and Log Relative Frequency do not appear in the same group: while the first captures the absolute frequency of an idiom, the second one basically corresponds to the probability to encounter an idiom given a certain verbal lexeme.

The best model in each regression was chosen via the AIC criterion, which allows minor residual errors, but disadvantages the inclusion of further predictors and helps avoiding overfitting. Each final model is bootstrap-validated by 200 resampling runs. Predictors that were excluded or did not have significant coefficients ($p > 0.05$) in the final models are not commented. Tables with the coefficient, the standard error and the t and p-values for each predictor in a model are reported in the Appendix. The label assigned to each model below refers to the dataset used (*H_{lex} vs No-H_{lex}* idioms) and to the modeled variable.

NO-$\mathbf{H}_{lex}$ PREDICTABILITY (Table I). Bootstrapped R^2 is 0.5792, two outliers were removed. As the coefficients show, the greater the Fixed Arguments Number and the PPMI Similarity between an idiom and its components, the more predictable an idiom. Log Relative Frequency also receives a positive coefficient.

Bootstrapped **NO-$\mathbf{H}_{lex}$ LITERALITY** (Table II) accounts for approximately 20% of the variance of Literality; two outliers were removed from the model. Both PPMI Similarity and Fixed Arguments Number have positive betas.

NO-$\mathbf{H}_{lex}$ SYNTACTIC FLEXIBILITY (Table III). Bias-corrected R^2 is 0.0713, with just one outlier excluded. The only predictor that reaches significance is Article Entropy, which has a positive coefficient.

$\mathbf{H}_{lex}$ PREDICTABILITY (Table IV) has a bootstrapped R^2 of 0.68, with two outliers cut out. Fixed Arguments Number, PLMI Similarity and Order Entropy appear to increase in parallel with Predictability, while PPMI Similarity, Morphological Entropy and Modifiers Entropy show the opposite tendency.

$\mathbf{H}_{lex}$ LITERALITY (Table V) after bootstrapping accounts for about 35% of the variance of Literality. Among significant predictors, PLMI Similarity obtains a positive coefficient.

Bootstrapped R^2 for **$\mathbf{H}_{lex}$ SYNTACTIC FLEXIBILITY** (Table VI) is equal to 0.6875, after the removal of three outliers. Significant predictors are Morphological Entropy, Modifiers Entropy and Log Frequency, with positive betas. Conversely, Order Entropy, Articles Entropy, Verbal Modifiers Entropy and Log Relative Frequency have all negative coefficients.

V. CONCLUSIONS

Large part of the variance in Predictability judgments turns out to be explained by a distributional semantic representation of idioms, and by the number of their fixed arguments: the more complex an idiom and the greater the similarity between its usage and the usage of its components, the more easily subjects can predict it. Formal variability measured with entropy appears to be relevant only for idioms with free slots, while Relative Frequency models Predictability only for lexically specified idioms. Literality is accounted for by distributional semantic similarity indices and, for fully lexicalized idioms, by the number of fixed arguments too. The portion of predicted variance (about 35% for H_{lex} and 20% for *No-H_{lex}* idioms) is however smaller than for Predictability. Further improvements can be expected by a better tuning of the DSMs parameters as well as by testing more advanced methods to estimate the semantic proximity between idioms and their components. As for Syntactic Flexibility, our model for lexically specified idioms explains just a restricted part of the variance (about 7%), with information on articles variability as our only significant predictor. Results are instead more promising for idioms with free slots: 68% of the variance is predicted by almost all our surface measures. The significance of most of our entropic measures therefore seems to be due to the

presence of a lexically free slot, except for the Order Entropy, which receives a negative coefficient. This latter fact can be explained by the way the Syntactic Flexibility index is defined in Tabossi *et al.* (cf. also the discussion below), as a sort of *semantic preservability* index, in that it actually captures how much the meaning of an idiom is preserved despite the syntactic transformations it undergoes. Consistently, the less variable the order of constituents is, the more the figurative meaning is preserved. By contrast, the positivity of other predictors (e.g. Morphological Entropy or Modifiers Entropy) needs to be further investigated.

Interestingly, while Wulff [22] finds that parameters related to the morphological variability of the idiom verbal head have the highest weight in predicting idiomaticity judgments, Verbal Morphological Entropy never appears as a significant predictor in our models. It must be noted, however, that Wulff [22] predicts idiomaticity ratings assigned to a set of literal and figurative V-NP constructions, while the judgments we have modeled only concern idiomatic expressions. We can therefore speculate that morphological variability of the verbal head is only relevant to discriminate idiomatic vs. non-idiomatic expressions. On the other hand, our dendrogram in Figure 1 shows that Verbal Morphology clusters together with Log Frequency, which instead appears among the significant predictors. A possible explanation of the results obtained by Wulff [22] could therefore be that variation in the verbal morphology is actually influenced by the frequency of an expression: the more frequent an idiom, the more various the syntactic contexts it occurs in and the more variable its verbal inflection.

In conclusion, we want to highlight some methodological issues that have emerged from the analysis of the normative data by Tabossi and colleagues [24] and that we intend to address in future research. As we have reported above, Syntactic Flexibility is only partially modeled by our corpus-based indexes. Noteworthily, the Syntactic Flexibility judgments are biased towards the high end of the 1-7 scale (mean 5.1, SD 0.5), that is to say idioms are generally regarded by the subjects as very flexible from the syntactic point of view. We speculate that this could be due to the kind of task the subjects had to perform: Asking the participants to compare the syntactically modified version of an idiomatic string with its figurative meaning could have biased them towards an idiomatic interpretation, influencing their acceptability judgments. We consequently propose to elaborate a different syntactic flexibility test consisting in collecting an acceptability score on a 1-7 scale for each sentence, without any comparison with the idiomatic meaning of the constructions under investigation. This way, non-idiomatic expressions could also be added as control stimuli to verify whether the ratings assigned to the two kinds of expressions would be significantly different. While theoretical studies have underlined that some syntactic variations are tolerated by idioms and literals alike (e.g. adverbial insertion [23]), other operations, like passivization or left dislocation, are expected to exhibit considerable differences in their acceptability [5], [14]. Finally, other syntactic operations that the literature describes as particularly puzzling for idioms could be tested, like the inversion of the arguments order (e.g. *gettare acqua sul fuoco* vs. *gettare sul fuoco acqua*) or the opposition between *internal* and *external* modification (e.g. *gettare molta acqua sul fuoco* 'to throw a lot of water on the fire' vs. *gettare proverbiale acqua sul fuoco* 'to throw proverbial water on the fire', where the second modification acts as a metalinguistic comment that embraces the idiom as a whole). The results of a pilot study we have performed using the crowdsourcing platform Crowdflower to collect the acceptability judgments indeed seem to confirm these predictions: the acceptability ratings for the idiomatic expressions are significantly lower than the ones collected by Tabossi *et al.*, and the ones for non-idiomatic control sentences. We leave to future research to investigate to what extent this new type of syntactic flexibility judgments can be accounted by corpus-based data.

We also intend to experiment with a wider range of distributional measures of compositionality [39], different ways to construct our distributional vectors, the refinement of the entropic indices to model idiomatic variability, as well as the use of other statistical methods (e.g. hierarchical multiple regression) to estimate the role of different formal and semantic distributional factors in the complex variation patterns of idiomatic expressions revealed by psycholinguistic data.

REFERENCES

[1] I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, "Multiword Expressions: A Pain in the Neck for NLP," in *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, 2002, pp. 1–15.

[2] N. Calzolari, C. J. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod, and A. Zampolli, "Towards Best Practice for Multiword Expressions in Computational Lexicons," in *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, 2002, pp. 1934–1940.

[3] A. Siyanova-Chanturia and R. Martinez, "The idiom principle revisited," *Applied Linguistics*, pp. 1–22, 2014.

[4] C. Cacciari and S. Glucksberg, "Understanding idiomatic expressions: The contribution of word meanings," *Advances in Psychology*, vol. 77, pp. 217–240, 1991.

[5] G. Nunberg, I. Sag, and T. Wasow, "Idioms," *Language*, vol. 70, no. 3, pp. 491–538, 1994.

[6] A. E. Goldberg, *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago, IL: University of Chicago Press, 1995.

[7] R. Jackendoff, *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press, 1997.

[8] D. Lin, "Automatic identification of non-compositional phrases," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999, pp. 317–324.

[9] D. McCarthy, B. Keller, and J. Carroll, "Detecting a continuum of compositionality in phrasal verbs," in *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 2003, pp. 73–80.

[10] T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows, "An empirical model of multiword expression decomposability," in *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, 2003, pp. 89–96.

[11] S. Evert, U. Heid, and K. Spranger, "Identifying morphosyntactic preferences in collocations," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004, pp. 907–910.

[12] S. Venkatapathy and A. Joshi, "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*, 2005, pp. 899–906.

[13] J. Ritz and U. Heid, "Extraction tools for collocations and their morphosyntactic specificities," in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006, pp. 1925–1930.

[14] A. Fazly, P. Cook, and S. Stevenson, "Unsupervised type and token identification of idiomatic expressions," *Computational Linguistics*, vol. 1, no. 35, pp. 61–103, 2009.

[15] G. Katz and E. Giesbrecht, "Automatic identification of non-compositional multi-word expressions using latent semantic analysi," in *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 2006, pp. 12–16.

[16] J. Birke and A. Sarkar, "A clustering approach to the nearly unsupervised recognition of nonliteral language," in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006, pp. 329–336.

[17] M. T. Diab and M. Krishna, "Unsupervised Classification of Verb Noun Multi-Word Expression Tokens," in *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, 2009, pp. 98–110.

[18] L. Li, B. Roth, and C. Sporleder, "Topic models for word sense disambiguation and token-based idiom detection," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1138–1147.

[19] J. Peng, A. Feldman, and E. Vylomova, "Classifying Idiomatic and Literal Expressions Using Topic Models and Intensity of Emotions," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 2019–2027.

[20] A. Fazly and S. Stevenson, "A distributional account of the semantics of multiword expressions," *Italian Journal of Linguistics / Rivista di Linguistica*, vol. 1, no. 20, pp. 157–179, 2008.

[21] L. Squillante, "Polirematiche e collocazioni dell'italiano. Uno studio llinguistico e computazionale," Ph.D. dissertation, Università di Roma "La Sapienza", 2013.

[22] S. Wulff, "Converging evidence from corpus and experimental data to capture idiomaticity," *Corpus Linguistics and Linguistic Theory*, vol. 5, no. 1, pp. 131–159, 2009.

[23] V. Bianchi, "An empirical contribution to the study of idiomatic expressions," *Italian Journal of Linguistics / Rivista di Linguistica*, vol. 5, no. 1, pp. 349–385, 1993.

[24] P. Tabossi, L. Arduino, and R. Fanari, "Descriptive norms for 245 italian idiomatic expressions," *Behavior Research Methods*, vol. 43, pp. 110–123, 2011.

[25] C. Cacciari, "Processing multiword idiomatic strings: Many words in one?" *The Mental Lexicon*, vol. 9, no. 2, pp. 267–293, 2014.

[26] R. W. Gibbs and N. P. Nayak, "Psycholinguistic studies on the syntactic behavior of idioms," *Cognitive Psychology*, vol. 21, no. 1, pp. 100–138, 1989.

[27] E. Holsinger, "Representing idioms: Syntactic and contextual effects on idiom processing," *Language and speech*, vol. 56, no. 3, pp. 373–394, 2013.

[28] S. Vietri, *Idiomatic Constructions in Italian: A Lexicon-grammar Approach*. John Benjamins Publishing Company, 2014.

[29] A. Lenci, G. E. Lebani, S. Castagnoli, F. Masini, and M. Nissim, "Sympathy: Towards a comprehensive approach to the extraction of italian word combinations," in *Proceedings of the First Italian Conference on Computational Linguistics*, 2014, pp. 234–238.

[30] A. Lenci, G. E. Lebani, M. S. G. Senaldi, S. Castagnoli, F. Masini, and M. Nissim, "Mapping the constructicon with sympathy. italian word combinations between fixedness and productivity," *NetWordS 2015 Word Knowledge and Word Usage*, pp. 144–149.

[31] T. Hoffmann and G. Trousdale, Eds., *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, 2013.

[32] M. Baroni, S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston, and M. Mazzoleni, "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*, 2004, pp. 1771–1774.

[33] F. Dell'Orletta, "Ensemble system for Part-of-Speech tagging," in *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian*, 2009.

[34] G. Attardi and F. Dell'Orletta, "Reverse revision and linear tree combination for dependency parsing," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 261–264.

[35] S. Wulff, *Rethinking Idiomaticity: A Usage-based Approach*. Continuum, 2008.

[36] P. D. Turney and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research*, vol. 37, pp. 141–188, 2010.

[37] S. Evert, "Corpora and collocations," in *Corpus Linguistics. An International Handbook*, A. Lüdeling and M. Kytö, Eds. Mouton de Gruyter, 2008, vol. 2, pp. 1212–1248.

[38] R Core Team, *R: A Language and Environment for Statistical Computing*, 2015. [Online]. Available: https://www.R-project.org

[39] J. Mitchell and M. Lapata, "Composition in Distributional Models of Semantics," *Cognitive Science*, vol. 34, no. 8, pp. 1388–1429, 2010.

VI. Appendix

Predictor	Beta	S.E.	t	p-value
Fixed Arguments Number	1.3772	0.2271	6.06	<0.0001
PPMI Similarity	2.7532	1.3188	2.09	0.0418
Articles Entropy	-0.7334	0.4321	-1.70	0.0958
Modifiers Entropy	-0.6363	0.4754	-1.34	0.1867
Log Frequency	-0.1843	0.1140	-1.62	0.1123
Log Relative Frequency	0.1635	0.0474	3.45	0.0011

TABLE I. *No-H_{lex} Predictability*

Predictor	Beta	S.E.	t	p-value
Fixed Arguments Number	0.9335	0.2673	3.49	0.0010
PPMI Similarity	5.5378	1.3371	4.14	0.0001
Modifiers Entropy	0.9099	0.5891	1.54	0.1283

TABLE II. *No-H_{lex} Literality*

Predictor	Beta	S.E.	t	p-value
PPMI Similarity	1.7949	1.2880	1.39	0.1690
Articles Entropy	1.4892	0.5633	2.64	0.0106

TABLE III. *No-H_{lex} Syntactic Flexibility*

Predictor	Beta	S.E.	t	p-value
Fixed Arguments Number	1.0871	0.2404	4.52	0.0002
PPMI Similarity	-5.1421	1.2230	-4.20	0.0005
PLMI Similarity	3.3898	1.1332	2.99	0.0075
Morphological Entropy	-4.2254	0.8282	-5.10	<0.0001
Order Entropy	6.9023	1.1494	6.01	<0.0001
Modifiers Entropy	-10.1364	1.6806	-6.03	<0.0001

TABLE IV. *H_{lex} Predictability*

Predictor	Beta	S.E.	t	p-value
PLMI Similarity	8.0489	2.0975	3.84	0.0009
Articles Entropy	-1.4271	0.9537	-1.50	0.1488

TABLE V. *H_{lex} Literality*

Predictor	Beta	S.E.	t	p-value
Morphological Entropy	5.7881	0.8633	6.70	<0.0001
Order Entropy	-2.5676	0.9290	-2.76	0.0138
Articles Entropy	-3.9033	0.9265	-4.21	0.0007
Modifiers Entropy	3.2573	1.4436	2.26	0.0384
Verbal Modifiers Entropy	-3.8651	0.9054	-4.27	0.0006
Log Frequency	0.4351	0.1635	2.66	0.0171
Log Relative Frequency	-0.3921	0.0624	-6.28	<0.0001

TABLE VI. *H_{lex} Syntactic Flexibility*