Objective Frequency Values of Canonical and Syntactically Modified Idioms: Preliminary Normative Data

Azzurra Mancuso & Alessandro Laudanna
LaPSUS, Laboratory of Experimental Psychology, University of Salerno
Via Giovanni Paolo II, 132 Fisciano, SA, 84084, Italy
amancuso@unisa.it; alaudanna@unisa.it

Abstract

In this study we collected several objective frequency values for 124 Italian idiomatic expressions, in order to verify the relation among these measures of frequency and a set of subjective variables (e.g., familiarity, meaning knowledge, age of acquisition, etc.) which are relevant from a psycholinguistic perspective, since they are supposed to play a role in idiom processing. Specifically, we calculated the following frequency types: occurrences of content words, (lemma and word-form values), occurrences of canonical idioms (e.g., Paolo broke the ice), occurrences of syntactically manipulated idioms (e.g., The ice was suddenly broken by Paolo). We discuss the results of correlational analyses.

1. Introduction

Several psycholinguistic norms are available for pictures and words (e.g., Barca, Burani, & Arduino, 2002; De Martino, Mancuso and Laudanna, 2017; Janssen, Pajtas, & Caramazza, 2011; Montefinese, Ambrosini, Fairfield, & Mammarella, 2014). However, this is less frequent for longer word-combinations, such as idiomatic expressions. An idiomatic expression comprises several words whose overall figurative meaning is not a direct function of its components (Tabossi, Arduino, & Fanari, 2011). For instance, the Italian idiomatic expression *rompere il ghiaccio* (“break the ice”) means “to take the initiative in an embarrassing situation” and thus its global meaning is far from the meaning of its components.

Some norms are available in English (Abel, 2003; Cronk, Lima, & Schweigert, 1993; Libben & Titone, 2008; Titone & Connine, 1994b), in French (Caillies, 2009; Bonin, Méot, & Bugaiska, 2013), in Bulgarian (Nordmann & Jambazova, 2017), in German (Citron et al., 2016) and in Italian (Tabossi et al., 2011). These databases collect mean values obtained from subjective ratings for some relevant psycholinguistic variables (such as age of acquisition, familiarity, meaning knowledge, etc.).

The existence of norms for idiomatic expressions has made it possible to account for issues concerning the comprehension, the production and the lexical storage of idioms (e.g., Cutting & Bock, 1997; Konopka & Bock, 2009; Sprenger, Levelt, & Kempen, 2006).

There are different theories on the topic of how idioms are stored in memory. According to some authors, idioms correspond to lexical units (e.g., Swinney & Cutler, 1979), whereas for others, they are stored as configurations of words (Cacciari & Tabossi, 1988; 2014). As claimed by Bonin et al. (2013), “it is therefore obvious that no empirical test of the different views of idiom processing is possible without first collecting norms for idioms”.

2. The present study

In the present research, we computed the frequency of 124 Italian idiomatic expressions in text corpora, in order to verify the relation among objective measures of frequency and a set of subjective variables which are available for Italian (Tabossi et al., 2011).

Some studies have underlined the influence exerted by the frequency values in the processing of these strings (Cronk et al., 1993; Libben & Titone, 2008; Bonin et al., 2013). In these works, the frequency values were obtained by calculating the familiarity of the expressions or the objective frequency (occurrence) of the individual words that compose the strings. Between the two methods, the first proved to be a better predictor of the complexity of processing (Bonin et al., 2013; Libben & Titone, 2008). The authors attributed this effect to the fact that the idiomatic meaning is often arbitrarily related to that of the individual constituents.

In our study, we pursued three main goals. The first was to collect the objective frequency of the isolated words that make up the Italian idiomatic expressions. Word frequency is certainly one of most important variables to have been considered by studies investigating reading or speaking. For instance, all influential models of word reading (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Harm & Seidenberg, 2004) are able to account for the finding that high-frequency words are processed faster and more accurately than low-frequency words in experimental tasks such as lexical decision and reading aloud. However, the influence of objective word frequency in idiom processing has received little attention (Cronk et al., 1993; Libben & Titone, 2008; Bonin et al., 2013). In the Italian normative study of idiomatic expressions (Tabossi et al., 2011), this variable was not taken into account.

The second goal was to obtain the objective frequency of idiomatic expressions, intended as the frequency of use of the idiomatic expression considered in its entirety. To our knowledge, all previous studied had not calculated this variable but focused exclusively on the subjective frequency of idioms. We claim that this methodology could offer several advantages to the research on idiom processing. First of all, it provides an objective measure of the degree of exposure to a given idiomatic expression by speakers, without being affected by any distortion or idiosyncrasy coming from subjective evaluations of familiarity. Some studies have shown that subjective frequency is a good index of the frequency of encounter of the words (Balota, Pilotti, & Cortese, 2001). However, the reliability of estimates of other-based familiarity (as considered in Bonin et al., 2013 and Tabossi et al., 2011) can be problematic, since it is more likely that participants can reliably estimate their own frequency of exposure to an idiomatic expression than how well other people know such expressions (Cronk et al., 1993; Libben & Titone, 2008; Titone & Connine, 1994b).

Moreover, the availability of corpus-based frequency values may offer an ideal shortcut to the preparation of psycholinguistic experiments, since familiarity estimates are often difficult to obtain, as they typically require running pre-studies to collect ratings. In this direction, recent studies claimed that subjective frequency ratings are no longer needed when objective word frequency norms are available (Brysbaert et al., 2011).

The third purpose of our study was to obtain objective frequency values of idioms used in a not canonical form (e.g., passive form, adjective and adverb insertion, etc.). Idioms have been traditionally described as fixed expressions, highly restricted in their realization (Cacciari & Tabossi, 1988; Gibbs, 1980; Swinney & Cutler, 1979; Titone & Connine, 1999). However, more recent corpus and experimental studies have shown that they are more flexible than previously thought (Moon, 1998; Barlow, 2000; Geeraert, Baayen, & Newman, 2017; Langlotz, 2006; Tabossi, Wolf, & Koterle, 2009; Vietri, 2014; Mancuso, Elia, Laudanna, & Vietri, 2019; Kyriacou, Conklin, & Thompson, 2019). The issue of idiom syntactic flexibility has received a renewed interest, since it also addresses the problem of how idioms are mentally stored.

3. Method

Materials.The idiomatic expressions used in the present work were taken from a study by Tabossi and colleagues (2011), who elicited normative judgments for Italian verbal idioms on the following variables:

- meaning knowledge, the proportion of correct meaning definitions given for each idiom;
- familiarity, the subjective frequency with which speakers encounter an idiom in its written or spoken form, regardless of their familiarity with the actual meaning of the phrase;
- age of acquisition, which indicates at what age the subjects thought they had learnt the expressions;
- predictability, the proportion of idiomatic completions given for a certain idiom, which was presented with the final word missing;
- syntactic flexibility, obtained by asking how much the meaning of the idiom in the syntactically modified version[1] was similar to its unmarked meaning, expressed in the form of a paraphrase;

[1]Each idiom was inserted in a sentence containing one of the following five syntactic modifications: adverb insertion, adjective insertion, left dislocation, passivization and wh-movement.

- literality, the plausibility of a literal interpretation for an idiom[2];
- compositionality, obtained by asking how much the component words of the idioms contribute to their overall meaning.

Each idiom was also associated with a length value calculated in words.

Procedure. In order to assess the frequency of content words that compose the idiomatic expressions we calculated their cumulative frequency, namely, the summed frequencies of the individual words divided by the number of words, as in Cronk et al. (1993) and Bonin et al. (2013). Differently from previous studies, we took into account both word-form and lemma frequencies; values were taken from CoLFIS (Bertinetto et al., 2005) and ItWaC (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009).

Moreover, we calculated the overall objective frequency of the expressions, intended as the frequency of co-occurrence of all words that make up the string, by means of ad-hoc queries within ITWaC.

We extracted the occurrence values of the idiomatic expressions in all the inflected forms of the verb (e.g., ' break/broke/breaks/etc. the ice'), by searching for the lemma (e.g., 'to break') and filtering the query by specifying one or more constituents (e.g., 'ice'). We adopted a context window of 7/10 elements (depending on the length of idioms), both to the right and left of the lemma, in order to obtain not only the frequency values of canonical idioms, but also the frequency of any possible syntactic manipulations where the order of presentation of the elements is modified (as in passive form, e.g., 'the ice was broken') or other lexical elements are inserted (as in adjective/adverb insertion, e.g., 'he has suddenly broke the ice', etc.). The results of each query were manually checked in order to eliminate casual co-occurrences (as instance, the sentence *la macchina si ruppe con il ghiaccio*, 'the car broke because of the ice' contains all words adopted as filters but does not correspond to the given idiomatic expression).

An example of a query is reported in Figure 1.

[2]For instance, *perdere il treno* "to miss the boat" (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.

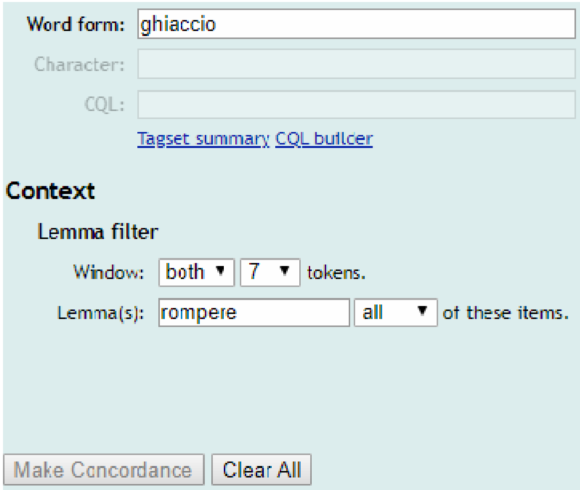

Figure 1. An example of query in ItWaC
(The idiomatic expression *rompere il ghiaccio* ('break the ice') is searched by filtering for the lemma *rompere* (to break) and the word-form *ghiaccio* (ice), within a context window of 7 tokens, both to the right and the left of the lemma)

4. Results

Data are now available for 124 idiomatic expressions with different degrees of length.

For each idiom, we collected several frequency values:

- Total frequency of idioms;
- Frequency of idioms occurring in a canonical form;
- Frequency of idioms occurring in a transformed form;
- Frequency in CoLFIS of word-forms and lemmas related to content-words appearing in idioms;
- Frequency in ItWaC of word-forms and lemmas related to content-words appearing in idioms.

Table 1 shows the means and the range of all frequency values calculated.

	means	range
TotFq	2,4	0-27
CanonFq	1,9	0-19
VariedFq	0,5	0-9
%varied	23%	0-100%
Ff CoLFIS	1.218	17 - 23.322
Fl CoLFIS	6.939	28 - 72.546
Ff ItWAC	281.642	3.741 - 4.512.480
Fl ItWAC	1.813.494	7.618 - 9.700.850

Table 1: Descriptive statistics (means and range) for the set of 124 idioms. **TotFq**=total frequency of idioms; **CanonFq**=frequency of canonical idioms; **VariedFq**=frequency of manipulated idioms; **FfCoLFIS**=word-form frequency in CoLFIS; **FlCoLFIS**=lemma

frequency in CoLFIS; **FfItWaC**=word-form frequency in ItWaC; **FlCoLFIS**=lemma frequency in ItWaC

Hereafter, we report some examples of very frequent idioms:

[1] *Cantar vittoria*, ‘to sing victory’

[2] *Guardarsi allo specchio,* ‘to look in a mirror’

and some examples of infrequent idioms:

[3] *Passare la misura* ‘to cross the line’

[4] *Avere ancora i denti da latte,* ‘to still have baby teeth’

For each idiom, all context occurrences are available in an Excel file. For ambiguous idioms (e.g., *break the ice*), we computed the frequency of all uses, both idiomatic and literal. Data about the syntactic flexibility of each idiom (the percentage of manipulations and the types of manipulation) can also be extracted. In this way, it will be possible for future research to obtain detailed information about the syntactic behavior of each idiomatic expression. Moreover, by analyzing context occurrences of expressions, it will be possible to disambiguate the figurative vs. literal use of ambiguous idioms, in order to derive objective frequency dominance values, in addition to subjective literal plausibility estimates, which are already available in Tabossi et al. (2011).

Below we report some examples of idioms which rarely occur in a manipulated form (less than 5%):

[5] *Battere la fiacca*, ‘to loaf about’

[6] *Mettere il carro davanti ai buoi,* ‘to put the cart before the horse’

and some examples of much flexible idioms (more than 30%):

[7] *Ingoiare la pillola*, ‘to swallow a bitter pill’

[8] *Mettersi nei panni di qualcuno*, ‘to put yourself in someone’s shoes’.

We carried out some correlational analyses in order to evaluate the relationship among objective frequency values and subjective variables, which are available for this set of idiomatic expressions (Tabossi et al., 2011). Hereafter, we will discuss most interesting results.

Relationship among subjective and objective frequency. As shown by Table 2, there is not a correlation between the frequency values of idioms and the frequency values of content words that compose the expressions: most used idioms are not necessarily made up by frequent words; rather, it often happens that frequent idiomatic expressions are composed by words that are used predominantly – if not exclusively – within such expressions (e.g., 'cuoia ' in ‘tirare le cuoia', ‘pull the skins'). Nevertheless, there are positive correlations between frequency values of words (both taken by CoLFIS and ItWaC) and subjective variables of familiarity and meaning knowledge: in other words, idiomatic expressions which have been rated more familiar and known by speakers are made up by frequent words. Interestingly, more frequent idioms are also more familiar but there is not a correlation between the frequency of idioms and meaning knowledge. We may interpret this finding as an evidence that speakers do not always know the exact meaning of idioms, independently by the fact that they occur very frequently in their language. As regards the frequency of manipulated idioms, we found a positive correlation with the frequency of lemmas (taken by CoLFIS): idioms which more often occur in corpora in a manipulated form are made up by frequent words. As expected, there are strong positive correlations among frequency values of words (both lemmas and word-forms) collected in CoLFIS and ItWaC.

Correlations between objective and subjective frequency

	2	3	4	5	6	7	8	9
1.TotFq	.99***	.87***	-.01	-.02	-.03	-.01	-.04	.21***
2.CanonFq		.77***	-.03	-.05	.01	-.06	.01	.23***
3.VariedFq			.06	.19**	.14	.16	.01	.09
4.Ff CoLFIS				.71***	.76***	.62***	.21***	.14
5.Fl CoLFIS					.78***	.94***	.24***	.18***
6.Ff ItWAC						.74***	.24***	.18**
7.Fl ItWAC							.26***	.20***
8.Know								.45***
9.Famil								1.00

Table 2. TotFq=total frequency of idioms; **CanonFq**=frequency of canonical idioms; **VariedFq**=frequency of manipulated idioms; **FfCoLFIS**=word-form frequency in CoLFIS; **FlCoLFIS**=lemma frequency in CoLFIS; **FfItWaC**=word-form frequency in ItWaC; **FlCoLFIS**=lemma frequency in ItWaC; **Know**=meaning knowledge; **Famil**=familiarity

Relationship among objective frequency values and psycholinguistic variables. As reported in Table 3, there is a negative correlation between the frequency and the age of acquisition of idioms: the idiomatic expressions acquired earlier are also the most frequent in corpora. Also, more frequent idioms are the shorter ones (negative correlation with the length, even in the case of manipulated idioms). Interestingly, all frequency values of words correlate negatively with literality: idioms containing frequent words have been judged less literally plausible by speakers.

Correlations between objective frequency and psycholinguistic variables						
	Length	AoA	Pred	Flex	Lit	Comp
1.TotFq	-.39***	-.21***	-.07	.05	.04	-.06
2.CanonFq	-.39***	-.22***	-.05	.04	.03	-.07
3.VariedFq	-.32***	-.13	-.10	.07	.05	-.02
4.Ff CoLFIS	.21***	-.04	-.04	.12	-.25***	-.05
5.Fl CoLFIS	.10	-.12	-.12	.17	-.29***	-.05
6.Ff ItWAC	.19	-.11	.03	.16	-.19***	-.03
7.Fl ItWAC	.10	-.15	-.13	.17	-.30***	-.06

Table 3. TotFq=total frequency of idioms; **CanonFq**=frequency of canonical idioms; **VariedFq**=frequency of manipulated idioms; **FfCoLFIS**=word-form frequency in CoLFIS; **FlCoLFIS**=lemma frequency in CoLFIS; **FfItWaC**=word-form frequency in ItWaC; **FlCoLFIS**=lemma frequency in ItWaC; **Length=**number of words; AoA=age of acquisition; Predic=predictability; Flex=syntactic flexibility; Lit=literality

5. Conclusions

In the present study, we pursued the main goal of collecting objective frequency values of idioms and evaluating their relation with a set of subjective variables available for Italian idiomatic (Tabossi et al., 2011). The novelty of our methodology allowed us to obtain corpus-based frequency values not only for content-words composing idioms (as reported in other normative data available for other languages, e.g., Cronk et al., 1993; Libben & Titone, 2008; Bonin et al., 2013), but also for idioms considered in their entirety. Furthermore, frequency values took into account also the occurrences of syntactically manipulated idioms (passive form, left dislocation, etc.).

The possibility of having objective frequency values of idiomatic expression can be an important support for directing future research on idiom processing. Recent psycholinguistic studies (e..g, Tabossi, Fanari, & Wolf, 2009) have questioned the hypothesis that the so-called 'idiom superiority effect' - namely, the established fact that idiomatic expressions are faster to process than literal sentences -is due to the idiomaticity itself of the expressions. According to the authors, the phenomenon could depend, more simply, on the fact that the idiomatic expressions adopted in most of the existing experimental studies were much more familiar than the literal sentences of control to which they were compared, which, in many cases, were completely new expressions, obtained by manipulating in part the idiomatic expressions of origin. A possible continuation of these studies could involve the implementation of experiments, in which idiomatic and literal expressions are matched for the objective frequency of occurrence, as well as a series of other well-known parameters. Moreover, studies aiming to explore the syntactic behavior of idioms might rely on objective frequency values of idioms occurring in a non-canonical form and explore the type and the percentage of manipulations for each idiomatic expression.

Acknowledgments

The authors would like to thank Simonetta Vietri for her constructive comments and recommendations on an earlier version of the paper and to Annibale Elia for his constant support to our research work.

Reference

Abel, B. (2003). English idioms in the first language and second language lexicon: A dual representation approach. *Second language research*, *19*(4), 329-358.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, *29*(4), 639-647.

Barca, L., Burani, C., & Arduino, L. S. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, & Computers*, *34*(3), 424-434.

Barlow, M. (2000). Usage, blends and grammar. *Usage-based models of language*, 315-345.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation, 43*(3), 209-226.

Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., & Thornton, A. M. (2005). *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS).* http://linguistica.sns.it/CoLFIS/Home.htm

Bonin, P., Méot, A., & Bugaiska, A. (2013). Norms and comprehension times for 305 French idiomatic expressions. *Behavior research methods*, *45*(4), 1259-1271.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental psychology*.

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of memory and language*, *27*(6), 668-683.

Cacciari, C., & Tabossi, P. (2014). *Idioms: Processing, structure, and interpretation*. Psychology Press.

Caillies, S. (2009). Descriptions de 300 expressions idiomatiques: Familiarité, connaissance de leur signification, plausibilitélittérale, "décomposabilité" et "prédictibilité". *L'AnnéePsychologique,* 109, 463–508.

Citron, F. M., Cacciari, C., Kucharski, M., Beck, L., Conrad, M., & Jacobs, A. M. (2016). When emo-

tions are expressed figuratively: Psycholinguistic and Affective Norms of 619 Idioms for German (PANIG). *Behavior research methods*, *48*(1), 91-111.
Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, *108*(1), 204.
Cronk, B. C., Lima, S. D., & Schweigert, W. A. (1993). Idioms in sentences: Effects of frequency, literalness, and familiarity. *Journal of Psycholinguistic Research*, *22*(1), 59-82.
Cutting, J. C., & Bock, K. (1997). That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory & Cognition, 25*(1), 57-71.
De Martino, M., Mancuso, A., & Laudanna, A. (2017). Variabili Rilevanti nella Rappresentazione delle Parole nel Lessico Mentale: Dati Psicolinguistici da una Banca-Dati di Nomi e Verbi Italiani. In Basili, R., Nissim, M., & Satta, G. (Eds.), *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017: 11-12 December 2017, Rome.* Torino: Accademia University Press.
Geeraert, K., Baayen, R. H., & Newman, J. (2017). Understanding idiomatic variation. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (pp. 80-90).
Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition*, *8*(2), 149-156.
Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological review*, *111*(3), 662.
Janssen, N., Pajtas, P. E., & Caramazza, A. (2011). A set of 150 pictures with morphologically complex English compound names: Norms for name agreement, familiarity, image agreement, and visual complexity. *Behavior Research Methods*, *43*(2), 478-490.
Konopka, A. E., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology*, *58*(1), 68-101.
Kyriacou, M., Conklin, K., & Thompson, D. (2019). Passivizability of Idioms: Has the Wrong Tree Been Barked Up?. *Language and speech*. https://doi.org/10.1177/0023830919847691
Langlotz, A. (2006). *Idiomatic creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English* (Vol. 17). John Benjamins Publishing.
Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. *Memory & Cognition*, *36*(6), 1103-1121.
Mancuso, A., Elia, A., Laudanna, A., & Vietri, S. (2019). The Role of Syntactic Variability and Literal Interpretation Plausibility in Idiom Comprehension. *Journal of Psycholinguistic Research*, https://doi.org/10.1007/s10936-019-09673-8
Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). The adaptation of the affective norms for English words (ANEW) for Italian. *Behavior research methods*, *46*(3), 887-903.
Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.
Nordmann, E., & Jambazova, A. A. (2017). Normative data for idiomatic expressions. *Behavior research methods*, *49*(1), 198-215.
Sprenger, S. A., Levelt, W. J., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of memory and language*, *54*(2), 161-184.
Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of verbal learning and verbal behavior*, *18*(5), 523-534.
Tabossi, P., Arduino, L., & Fanari, R. (2011). Descriptive norms for 245 Italian idiomatic expressions. *Behavior Research Methods*, *43*(1), 110-123.
Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast? *Memory & Cognition*, *37*(4), 529-540.
Tabossi, P., Wolf, K., & Koterle, S. (2009). Idiom syntax: Idiosyncratic or principled?. *Journal of Memory and Language*, *61*(1), 77-96.
Titone, D. A., & Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbol*, *9*(4), 247-270.
Vietri, S. (2014). *Idiomatic constructions in Italian: a lexicon-grammar approach* (Vol. 31). John Benjamins Publishing Company.