



Bridging the Gap: Creation of a Lexicon of 150 Pairs of English and Italian Idioms Including Normed Variables for the Exploration of Idiomatic Ambiguity

RESEARCH PAPER

IRENE PAGLIAI 

ubiquity press

ABSTRACT

This paper describes the creation of a normed lexicon of 150 pairs of English and Italian idioms annotated by translatability level (Beck, 2020). The dataset was created through the implementation of a cross-linguistic norming study conducted online via a novel combination of tools and design, which was validated by data reliability measurement. The lexicon contains information with respect to two groups of idiom variables (Hubers, Cucchiaroni, Strik, & Dijkstra, 2019): Experience-Based Variables (“EBVs”: familiarity, meaningfulness and objective knowledge) and Content-Based Variables (“CBVs”: literal plausibility, decomposability and transparency). These variables are particularly relevant for the analysis of ambiguous contexts where there is an interaction between the literal and figurative meaning of an idiom (Wagner, 2021). In particular, the sum of the mean ratings obtained for the three CBVs provides the Potential Idiomatic Ambiguity (“PIA”) index of each idiom: the higher the PIA, the more likely the idiom should be to occur in ambiguous contexts. This index can therefore be exploited in the setup of future psycholinguistic experiments to verify the relationship between the internal distribution of idiom features and ambiguous contexts. Moreover, the lexicon is a valuable research tool per se, since it contributes to bridging the gap in cross-linguistic research in idiom norming studies (Nordmann, Cleland, & Bull, 2014; Pastor, 2021), enabling systematic comparative analyses and enhancing understanding of the relationships among the examined variables.

CORRESPONDING AUTHOR:

Irene Pagliai

RTG 2636: Form-meaning mismatches, Georg-August Universität, Göttingen, Germany

irene.pagliai@uni-goettingen.de

KEYWORDS:

idiom norming study; cross-linguistic idioms; ambiguous contexts

TO CITE THIS ARTICLE:

Pagliai, I. (2023). Bridging the Gap: Creation of a Lexicon of 150 Pairs of English and Italian Idioms Including Normed Variables for the Exploration of Idiomatic Ambiguity. *Journal of Open Humanities Data*, 9: 16, pp. 1–13. DOI: <https://doi.org/10.5334/johd.123>

Idioms are multiword expressions whose prototypical meaning is figurative (Cacciari & Tabossi, 2014; Giora, 1997). For instance, the Italian idiom *essere la ciliegina sulla torta* (“be the cherry on the cake”) means “be something extra that makes a good thing even better”, and the English *kick the bucket* means “die”. As idioms are non-compositional (Monti & di Buono, 2019), the linguistic contexts in which they occur typically show no reference to their literal constituents. Note, however, examples 1a and 1b:

- (1) a. La lettura è un fattore di modernità, **non è la ciliegina sulla torta** ma è la **torta stessa**.
‘Reading is a factor of modernity; **it is not the icing on the cake** but **the cake itself**.’
la Repubblica
- b. But, if I do end up on some opiate, **the bucket has been kicked** and **leaking** significantly.
pilotsofamerica.com

Such examples show an interesting contextual variation whereby the literal constituents of the idioms are exploited in the co-text to create ambiguity where there is interaction between literal and figurative meaning; such uses are also referred to as “creative” (Langlotz, 2006).

All idioms are characterised by some degree of non-compositionality and formal complexity. However, they are also heterogeneous elements that differ from each other with respect to several variables (Wulff, 2013). In particular, idioms that are positively characterised by specific features seem to be more likely to occur in ambiguous contexts (see section 1.1.1). Consequently, to investigate the relationship between idiom external variation (i.e. contextual) and idiom internal variation (i.e. how each idiom is characterised with respect to the features of interests), it is necessary to collect quantifiable information on the idiom features that are relevant for the study of ambiguous idiomatic contexts. This is achieved by implementing a cross-linguistic norming study on a selection of English and Italian idioms. Norming studies are questionnaire-style tools by which participants are asked to provide judgements (typically Likert-scale ratings or categorical choices) with respect to the relevant variables. Our norming study led to the creation of the dataset described in this paper: a normed cross-linguistic lexicon of 150 English – Italian idiom pairs.

In the remainder of the introduction, we will define the variables included in the study (1.1) and the motivations supporting the creation of the dataset (1.2). Following this, the methodology adopted will be described in detail in 2. The results of the norming study will then be presented and discussed in section 3, including an in-depth look at ethical decision-making in crowdsourcing scenarios (3.1), the description of the cross-linguistic lexicon (3.2) and its reliability measurement (3.3). The paper will subsequently conclude in 4 with a concise summary and the outlines of future research.

1.1 THE VARIABLES

1.1.1 Content-Based Variables

Following Hubers et al. (2019), we define Content-Based Variables (“CBVs”) as variables whose assessment is based on an idiom’s linguistic content. The CBVs considered are *literal plausibility*, *decomposability* and *transparency*; they were selected because idioms positively characterised by such features seem to be more likely to occur in ambiguous contexts (Haagsma, 2020; Vulchanova, Milburn, Vulchanov, & Baggio, 2019; Wagner, 2021).

Literal plausibility is also referred to as “literality” (Tabossi, Arduino, & Fanari, 2011) or “ambiguity” (Citron et al., 2016); it refers to the possibility that the literal-compositional meaning of an idiom denotes an event consistent with our world knowledge (Wagner, 2021). Since it is reasonable to imagine someone literally kicking a bucket, *kick the bucket* is a literally plausible idiom. In contrast, when considering the idiom *be in seventh heaven* (“be very happy”), it is unrealistic that someone could literally be in a heaven labelled number seven. In other words, literally plausible idioms have “possible alternate non-idiomatic readings” (Fellbaum, 2019: 736). According to Vulchanova et al. (2019), such a feature implies that these idioms are particularly prone to trigger competition between literal and figurative meaning; conversely, implausible idioms have no such internal competition between the two meanings, hence they “do not have a strong potential ambiguity” (Wagner, 2021: 35).

Decomposability is also referred to as “compositionality” (Michl, 2019b; Titone & Connine, 1994) and “analysability” (Gibbs Jr & Nayak, 1989); it refers to the possibility of decomposing the figurative meaning of an idiom and associating the parts with its literal constituents (Sailer, 2021). Decomposability is, therefore, a variable that investigates the interface between the syntactic and semantic level of an idiom (Maher, 2013). In fact, both Geeraerts (1995) and Fadlon, Horvath, Siloni, and Wexler (2013) argue that decomposable idioms are characterised by isomorphism between the syntactic and semantic structure. Consider two examples: *kick the bucket* is typically considered non-decomposable, since the figurative meaning “die” cannot be associated with both the constituents “kick” and “the bucket”; conversely, *spill the beans*, whose meaning is “reveal secret information”, is usually considered decomposable, since the syntactic-semantic bridge is rather easily identifiable: “spill” = “reveal” and “the beans” = “secret information” (see Nunberg, Sag, & Wasow, 1994).

Given its relevance to both syntax and semantics, decomposability has been included in most norming studies. It has also been pointed out as a key factor in explaining the degree of linguistic flexibility of idioms (Gibbs Jr & Nayak, 1989). However, recent developments in linguistics and psycholinguistics invite to resize the overall explanatory power of decomposability (Maher, 2013; Michl, 2019b) and to consider it in a multi-layered scenario where other variables also play a key role in the analysis of idiomatic variation.

Transparency is also referred to as “motivation” (Geeraerts, 1995), “relation” (Michl, 2019a), “semantic bridge” (Dobrovol’Skij, 2016); it refers to the possibility of tracing a synchronic relationship between the literal and figurative meaning of an idiom (Sailer, 2021). Such relationship serves as rationale for the figurative use, which can consequently be inferred from the literal constituents even when one is not familiar with an idiom (Moreno, 2005). *Kick the bucket* can be considered opaque (i.e., non-transparent), since the scene evoked by the literal meaning has no link with the scene evoked by the figurative meaning (there is an “incongruence”, in the terms of Boers & Webb, 2015); in contrast, the idiom *pull strings*, whose meaning is “leverage influential connections”, can be considered transparent when one thinks of the links between the two semantic poles: the first link is the conceptual metaphor CONTROL IS OBJECT MANIPULATION (present in the MetaNet¹ database), while the second link is the presence in one’s world knowledge of the figure of the puppeteer, who imposes control over a puppet with the aid of some strings (Sailer, 2021). Transparency has been overall less investigated, as it is often subsumed within decomposability itself (like in Libben & Titone, 2008). However, Geeraerts (1995) and later Hubers et al. (2019) argue that these variables are distinct constructs: decomposability draws attention to the relationship between the syntactic and semantic dimensions of an idiom, while transparency highlights any conceptual links (typically provided by metaphor, metonymy and world knowledge, see Boers & Webb, 2015; Kovecses & Szabco, 1996) that may exist between the figurative and literal meaning of an idiom.

Only few studies have considered both decomposability and transparency: see Carrol, Littlemore, and Dowens (2018) and Michl (2019a, 2019b). However, these studies differ from each other in their definitions and operationalisations. Carrol et al. (2018) defines transparency as the ease with which the meaning of an idiom can be guessed on the basis of individual words, and is operationalised by means of a 1 to 7 Likert scale. In their study, transparency is presented pre-meaning, i.e. it is rated before the figurative meaning of the idiom is made explicit. As for decomposability, it is considered as the ease with which the connection between individual words and the figurative meaning can be seen once the figurative meaning of the idiom is made explicit. It is hence presented post-meaning. Like transparency, it is operationalised on a 1 to 7 Likert scale.

Differently, Michl (2019a) divides transparency into two sub-variables: comprehensibility and relation. The first is defined as “the ease with which the meaning of an idiomatic unit can be recovered” (p. 102), and the author herself equates it to the variable meaningfulness introduced in other studies (see 1.1.2). The latter describes “the relatedness between the literal and the non-literal meaning of an idiom” (p. 102). It refers to the strength of the semantic-conceptual link between the two meaning poles of the idiom. Relation is therefore closer to transparency as understood in this paper. Both sub-variables are operationalised on a 1 to 5 Likert scale and are presented post-meaning. As for decomposability, it is examined in Michl (2019b), and is

1 Last accessed in May 2023.

defined as “the degree to which the meaning of an idiom can be composed from the meaning of its constituents” (p. 1288). It is presented post-meaning, but this time it is operationalised on a 1 to 7 Likert scale.

Overall, such different definitions and operationalisations may lead to different results, making a comparative analysis of transparency and decomposability complex.

Transparency becomes especially relevant for the present norming study when one considers that the identification of the relationship between the two meanings allows speakers to move more easily along the continuum between literalness and figurativeness (Moreno, 2005: 410–11), hence fostering the possibility of using transparent idioms in ambiguous contexts. Finally, the results reported by van Ginkel and Dijkstra (2020) show how transparency interacts with literal plausibility in creating competitiveness between the literal and figurative meaning of an idiom.

1.1.2 Experience-Based Variables

Following Hubers et al. (2019), we define Experience-Based Variables (“EBVs”) as variables whose assessment strictly depends on speakers’ subjective experience with idioms. The EBVs considered are *familiarity*, *meaningfulness* and *objective knowledge*. EBVs have been consistently reported to influence CBVs perception and assessment (Libben & Titone, 2008; Nordmann et al., 2014), which is why it is essential to consider both groups to obtain a comprehensive understanding of idioms’ internal variation.

Familiarity refers to the subjective frequency with which an idiom is used and heard/read (Titone & Connine, 1994). In general, there is widespread consensus on the impact of familiarity on the perception of the other variables (see especially Nordmann et al., 2014; Nordmann & Jambazova, 2017), as well as on idiom cognitive processing (Libben & Titone, 2008).

Meaningfulness is also referred to as “knowledge” (Tabossi et al., 2011); it refers to the perceived, subjective knowledge of the figurative meaning of an idiom (Libben & Titone, 2008; Nordmann et al., 2014). As both familiarity and meaningfulness measure a speaker’s subjective experience with an idiom, the two EBVs are related. This does not mean, however, that they describe the same thing (Nordmann et al., 2014), since one may know well the figurative meaning of an idiom that is nevertheless used infrequently.

Objective knowledge refers to the actual knowledge of the figurative meaning of an idiom; it is the objective counterpart of meaningfulness, as it verifies the correctness of the figurative meaning associated with an idiom. Citron et al. (2016) highlight that speakers can confidently believe to know the figurative meaning (high meaningfulness) while having an incorrect paraphrase in mind (no objective knowledge). This is why it is appropriate to include objective knowledge as balancer.

1.2 MOTIVATIONS

The lexicon is a validation tool for future research on ambiguous idiomatic contexts. More precisely, for each idiom, the sum of the mean ratings of the three CBVs (literal plausibility, decomposability and transparency) provides the index of “Potential Idiomatic Ambiguity” (“PIA”), whereby the higher this is, the more suitable the syntactic, semantic and lexical configuration of the idiom is for occurring in ambiguous contexts like 1a and 1b. The lexicon is therefore a “structured tool” containing “empirically determined variables” (Citron et al., 2016: 104), which makes it an ideal starting point for the setup of future experiments.

Secondly, the dataset is a rich source of information to explore per se. Indeed, the cross-linguistic lexicon enables us to:

- a) Further enhance research in the field of idiom norming studies, especially because it bridges the current gap in cross-linguistic research (Nordmann et al., 2014; Pastor, 2021).
- b) Enrich knowledge on Italian idioms, which, as far as we know, have only been investigated in the norming study by Tabossi et al. (2011).
- c) Shed light in a more systematic way on two of the most complex variables considered in the literature, decomposability and transparency. As norming studies differ greatly in

variable definitions and operationalisations (Hubers et al., 2019), the knowledge available on the two dimensions is fragmentary and diverse (Carrol et al., 2018; Hubers et al., 2019; Libben & Titone, 2008; Michl, 2019a). Since the norming study that led to the creation of the cross-linguistic lexicon exploited the same definitions and operationalisations for both English and Italian, it is possible to make progress in this regard too.

LL PAIRS	SL PAIRS	PL PAIRS
be green with envy – essere verdi d'invidia	be under someone's feet – stare tra i piedi a qualcuno ("stay among someone's feet")	spill the beans – sputare il rospo ("spit the toad")
pick up the pieces – raccogliere i pezzi	put your foot down – puntare i piedi ("point the feet")	go through the wringer – passare le pene dell'inferno ("pass the hell's pains")

Table 1 Idioms illustrating the three levels of translatability.

2 METHOD

2.1 MATERIALS

One hundred and fifty pairs of English and Italian idioms with similar meanings were selected for the study. Initially, a larger sample (=600) of English idioms was pooled from several existing resources (the MAGPIE and TInCAP corpora by Haagsma, 2020 and Wagner, 2021, plus the dataset resulting from the norming study by Bulkes & Tanner, 2017). Afterwards, the corresponding Italian idioms were manually selected by exploiting online resources and corpora.² This resulted in 300 cross-linguistic idiom pairs.

The cross-linguistic pairs were subsequently manually annotated by categories of translatability (Beck, 2020), which describes "how closely the idioms match in their word-for-word translations" (p. 27), meaning the degree of lexical similarity within each idiom pair. Following Irujo (1986) and Liontas (2002), Beck (2020) describes three levels of translatability: lexical ("LL"), in which there is an exact overlap of lexical material between two idioms; semi-lexical ("SL"), in which the lexical overlap is only partial; post-lexical ("PL"), where there is no lexical overlap. To obtain a balanced dataset, it was decided to include 50 pairs for each translatability category in the lexicon, resulting in a total of 150 pairs of cross-linguistic idioms. See Table 1 for some examples of idiom pairs belonging to each category.

All idioms comprised in the lexicon are presented within a minimal verb structure. Idioms that typically occur with the verb "to be" are also included.

2.2 PARTICIPANTS

Thirty participants per language were recruited via the crowdsourcing platform Prolific. Hubers et al. (2019) showed that a minimum of twenty participants is necessary to obtain reliable data in norming studies. Furthermore, their analyses together with the ones in Nordmann and Jambazova (2017) point out that an increase in the number of participants does not result in higher data reliability. For this reason, thirty seemed the appropriate number in view of possible exclusions due to data quality issues.

The selection of potential participants was filtered by native language (English or Italian), minimum number of submissions on Prolific (15–40), minimum success rate in previous studies (75–100%).

Before deciding whether to take part in the study, participants were informed of the relevant details on Prolific; these consisted of a brief description of the tasks to be completed, the estimated completion time (based on a previously conducted pilot), the way we would keep track of data quality, the devices allowed to take part in the study, and payment details. In this regard, it was decided to meet a standard of £9 per hour, which Prolific labels as "good payment".

² The online resources exploited are WordReference, Reverso Context, the De Mauro Italian Dictionary and the Dizionario dei Modi di Dire. The corpora used are CORIS, Paisà and enTenTen20 on SketchEngine. Last accessed in January 2023.

All participants provided informed consent before starting the study. They were then asked to provide some demographic information: sex assigned at birth, age, country of residence, highest degree obtained, number and type of languages known. Data collection was divided into three batches: ten participants at a time were recruited for each language.

As a result of data quality control (2.3.1), one Italian-speaking participant and two English-speaking participants were excluded. See [Table 2](#) for a summary of the demographic details of the workers whose data have been included in the lexicon. To view the demographic details of each participant, consult the “participant_details_ENG” and “participant_details_ITA” files in the [repository](#).

	ENGLISH	ITALIAN
Sex	20 F; 8 M	17 F; 12 M
Age – mean (s.d.)	33.3 (9.5)	30.8 (11.1)
Residence countries	UK (14); South Africa (5); Ireland (3); Canada (3); Poland (1); Portugal (1); USA (1)	Italy (26); UK (3)
Education	High school diploma (10); Bachelors' degree (10); Master's degree (7); PhD (1)	High school diploma (15); Bachelor's degree (8); Master's degree (5); Junior high school diploma (1)
Number of languages per participant – median (IQR)	1 (1)	2 (1)

Table 2 Overview of participants' demographic details.

2.3 DESIGN

The data collection was carried out entirely online via the [Gorilla Experiment Builder](#) research platform. For each language, a two-part norming study was developed. During the first part, participants were asked to assess the EBVs and literal plausibility: familiarity, meaningfulness (which was defined “knowledge of the meaning” in the study, for the sake of clarity) and literal plausibility were operationalised using 1 to 5 Likert scales, while objective knowledge was operationalised via a dropdown option in which to choose the correct idiom paraphrase from three alternatives. In the second part, participants assessed decomposability and transparency, both operationalised using 1 to 5 Likert scales (see Supplementary Files to access operationalisation examples in the following files: “example_ratings_part1.pdf” and “example_ratings_part2.pdf”).

Before accessing the evaluation phase, participants read detailed instructions including definitions of all the relevant terms and careful illustrations of the variables for each section. Where appropriate, variable explanation was accompanied by examples showing the extremes of the Likert scale. In addition, a reference question was formulated for each variable, which participants were invited to ask themselves in case of doubts during the rating phase. For instance, the reference question for transparency was “how easily can I infer the figurative meaning from the literal meaning – based on a possible relationship between the two?” (see Supplementary Files: “instructions_part1.pdf” and “instructions_part2.pdf”). Each reference question was then placed next to each variable operationalised by means of a Likert scale, so that it was readily available.

For each section, idioms were spread over five randomised pages comprising approximately 30 idioms each. Upon completion of each page, participants were invited to take a short break before continuing. After completing the first section, participants only had access to the second one after a minimum of 24 hours; they were informed in advance on Prolific that compensation for their work would be paid once both parts of the study were completed.

Setting up and monitoring a multi-part online study poses challenges with regard to potential technical issues and participant engagement. Despite this, the choice of such complex design is motivated. The division of the variables into two separate sections is motivated by the need to present the idioms without their figurative meaning paraphrase in the first part. This ensures that the tasks (the selection of the correct meaning and the rating of the plausibility of the literal meaning) are not interfered with. In the second part, the idioms are presented with their figurative meaning paraphrase to obtain reliable ratings for decomposability and

transparency (Carrol et al., 2018). Furthermore, the multi-part design made it possible to dilute the cognitive load of the participants. Therefore, the selected structure seemed the most suitable option to have a reliable fully within-subject design in which each participant rated each idiom for each variable.

It should now be pointed out that the most popular design in idiom norming studies is the between-subject one, where different groups of participants evaluate a single variable or a group of them, after which the data are aggregated (see, among others, Citron et al., 2016; Gavilán, Haro, Hinojosa, Fraga, & Ferré, 2021; Tabossi et al., 2011; Titone & Connine, 1994). In this regard, Nordmann and Jambazova (2017) have shown that design choice does not impact the correlations between the variables. However, they also claim that “it is important to collect these ratings within subjects, because they can never be independent and should not be treated as such” (p. 200). Following the same line, Carrol et al. (2018) also stated that “familiarity has a direct influence on perceptions of transparency, and semantic judgements therefore cannot be treated as independent” (p. 40). Therefore, adopting a within-subject design is the most rigorous approach to obtain data that accurately reflect the relationships between idiom variables.

2.3.1 Data quality tracking

To ensure the quality of the answers provided by the crowdworkers, some control items were included in the study. For their creation, the norming study by Bulkes and Tanner (2017) was taken as a model.

In the first section, ten idioms (two per page) from other languages (Thai, Japanese, Korean, Arabic, Hindi) translated into Italian and English were included. These control idioms do not exist in English and Italian, so the expectation was to obtain low ratings (≤ 2) on the Likert scale associated with familiarity and meaningfulness. In the second section, ten (two per page) literal expressions (such as “eat a sandwich”, which has been paraphrased as “nourish oneself with some stuffed bread”) were included. As these expressions do not have a figurative meaning, the expectation was to obtain high ratings (≥ 4) for both decomposability and transparency. The control elements were considered successfully passed if participants answered seven out of 10 per section correctly (but see 3.1). In case of failure, participants were contacted on Prolific, given detailed explanations of their responses and of the contradicted expectations, and finally asked to “Return” their submission. Note that at the end of each section, a debriefing was included in which the control elements were explained to the workers (see Supplementary Files).

3 RESULTS AND DISCUSSION

3.1 CROWDSOURCING OUTCOME AND ETHICS

Table 3 shows the average completion times of the study, divided into first part, second part and total. To view the completion times of each participant, consult the “participant_details_ENG” and “participant_details_ITA” files in the repository.

COMPLETION TIMES IN MINUTES: MEAN (S.D.)		
FIRST PART	SECOND PART	TOTAL
58 (32)	47 (23)	105 (52)

Table 3 Average completion times.

To the best of our knowledge, the combination of tools used (Gorilla and Prolific) and design adopted (fully within-subject and multi-part) is novel in the field. For this reason, it seems appropriate to devote a section to exemplify through two case studies the strategies followed to make ethical decisions towards the crowdworkers. The first episode involves a participant who failed the control items of the second part (=3/10). For this reason, they were contacted and requested to “Return” the submission. The worker subsequently wrote a detailed explanation for their ratings, revealing a lack of understanding of decomposability and transparency. For this reason, it was decided to collect more information on their behaviour: firstly, a widespread tendency to give low ratings of decomposability and transparency for all idioms was noted.

The time taken to read the instructions and to complete the ratings was also checked, and it was evident that the worker had invested their time in completing the study. Therefore, the combination of quantitative and qualitative observations revealed that their responses were intentional and consistent throughout the second section of the study. As a result, it was decided to compensate the participant, but their data was not included in the lexicon.

The second episode involves a participant who passed only six out of 10 control items in the first section. Upon examining the languages the participant claimed to know, it was noticed that these included Malay and Chinese, which are languages geographically close to Thai and Japanese, from which two idioms each were translated. Interestingly, when checking which control idioms the participant considered to be relatively familiar and meaningful, we found the two Thai idioms and one Japanese idiom. It was hypothesized that linguistic interference might be at play, as the participant may be familiar with similar idioms in the known languages. As a result, the participant was granted access to the second section, where they performed perfectly on the control items. Consequently, their data was included in the normed idiom lexicon.

In conclusion, when crowdsourcing is exploited to collect data, it is advisable to have both quantitative and qualitative tools to track data quality. When no problems arise, quantitative tools are sufficient to declare the goodness of data. However, when facing doubts and/or complex situations, qualitative observations regarding workers' behaviour and responses should also be exploited, so that researchers can make informed, ethical decisions with respect to the hired workers.

3.2 DATASET DESCRIPTION

The normed lexicon of cross-linguistic idioms is in the [repository of the University of Göttingen](#), it is searchable and downloadable. Its Digital Object Identifier (DOI) is [10.25625/EPswDY](#), the license applied is CC BY-NC 4.0. The data files are in the non-proprietary csv format.

The lexicon in its entirety is contained in the "whole_dataset" file. Let us see the columns it contains:

- **IDIOM_ID**: the id associated with each idiom. Note that each pair of cross-linguistic idioms has the same numeric id, differentiated by the language (e.g., the idioms 15eng and 15ita constitute a cross-linguistic pair).
- **IDIOM**: the idiom.
- **LANGUAGE**: two-level variable ("ENG" or "ITA") indicating whether the idiom is Italian or English.
- **TRANSLATABILITY**: three-level variable ("LL", "SL", "PL") indicating the translatability relationship within the cross-linguistic idiom pair.
- **SYNTAX**: shallow parsing of the syntactic structure of each idiom. This column was included for its potential in future research, especially regarding decomposability, as the variable is particularly relevant to the analysis of the interface between idiom syntax and semantics. Note that idioms presented within a verbal structure beginning with the verb "to be" are differentiated: *be a breath of fresh air* is analysed as VPbe_NP_PP, while *get a slice of the pie* as VP_NP_PP.
- **SAME_SYNTAX**: two-level variable ("YES" or "NO"), it indicates whether the idioms within a cross-linguistic pair have the same syntactic structure.
- **PARTICIPANT**: the id of the participant who provided the data in the row.
- **Familiarity; Meaningfulness; Literal_Plausibility; Decomposability; Transparency**: the rating from 1 to 5 provided by the participant for the relevant variable.
- **Objective_Knowledge**: two-level variable (0 or 1) indicating whether the participant correctly guessed (= 1) or not (= 0) the figurative meaning of the idiom.

The normed lexicon in its entirety is ready to be used for descriptive and/or inferential statistical analyses.

It should be noted that releasing all raw ratings of an idiom norming study is not common practice: in most cases, only aggregate data for each variable is released, typically the mean

associated with the standard deviation (some examples: Beck, 2020; Bulkes & Tanner, 2017; Citron et al., 2016; Hubers et al., 2019; Libben & Titone, 2008). To ensure maximum transparency and adhere as closely as possible to the open science guidelines, it was decided to release the dataset in its entirety.

Both because it is common practice in the field and because we want to allow for observations of a more qualitative nature, it was nevertheless decided to release the dataset also in the form of aggregate data. In particular, two perspectives are present in the repository: the monolingual and the cross-linguistic one.

The “monolingual overviews” folder contains two csv files, one for English and the other for Italian (“eng_overview” and “ita_overview”). These files are designed for those interested in only one of the two languages. Let us see the columns of “ita_overview” as an example, i.e. the aggregate data regarding Italian idioms.

- **IDIOM_ID, IDIOM_ITA, SYNTAX:** same as in “whole_dataset” (“IDIOM_ITA” instead of just “IDIOM”).
- **IDIOM_LITERAL_TRANSLATION:** the literal English translation of the Italian idiom, to make the content accessible to everybody (absent in the “eng_overview” file).
- **IDIOM_MEANING_ITA:** the figurative meaning of the idiom in Italian. Note that this is the paraphrase that was read by the participants while completing the study.
- **IDIOM_MEANING_ENG:** the paraphrase in English, to make it accessible to everybody.
- **Familiarity_MEAN:** mean value resulting from all raw ratings the idiom received on the familiarity scale. The same applies to “Meaningfulness_MEAN”, “Literal_Plausibility_MEAN”, “Decomposability_MEAN” and “Transparency_MEAN”.
- **Familiarity_SD:** standard deviation of the mean value of familiarity. The same applies to “Meaningfulness_SD”, “Literal_Plausibility_SD”, “Decomposability_SD” and “Transparency_SD”.
- **Objective_Knowledge_PROPORTION:** proportion ranging from 0 to 1 indicating how often the meaning of the idiom was correctly selected.
- **PIA:** the index of Potential Idiomatic Ambiguity associated with the idiom, resulting from the sum of the mean values of literal plausibility, decomposability and transparency.

Finally, the “cross-linguistic overviews” folder contains three csv files, one for each translatability relation (“crossling_overview_LL”, “crossling_overview_SL”, “crossling_overview_PL”). These files bring together the pairs of idioms that share the same translatability relationship, and allow for an immediate comparison based on the aggregate data between the idioms that are within a cross-linguistic pair. The files are arranged horizontally, so that on the left-hand side are the English idioms and on the right-hand side the corresponding Italian idioms. The columns included are “IDIOM”, “IDIOM_ID”, “SYNTAX”, “SAME_SYNTAX”, “IDIOM_MEANING” (for the Italian part also “IDIOM_LITERAL_TRANSLATION”), all “MEAN” values and associated “SD” for idiom features on Likert scales, the “PROPORTION” for objective knowledge, and finally the “PIA”.

3.3 RELIABILITY

By data reliability we refer to the degree to which the ratings obtained covary in a sensible way (see Hubers et al., 2019). Measuring reliability is in this case especially important, since the combination of tools (Gorilla and Prolific) and design (fully within-subject, based on a two-part structure) has not been previously tested in the literature. It is therefore a way to validate or invalidate the proposed methodology through statistical analysis.

A well established metrics is the Intraclass Correlation Coefficient (ICC), a statistical measure spanning from 0 to 1 that quantifies the degree of consistency among ratings made by different raters (see Shrout & Fleiss, 1979). There are different types of ICCs: the most appropriate for the present case is the ICC (2,k), being a tool for checking rating consistency while assuming random effects for both participants and items (Hubers et al., 2019).

Furthermore, Hubers et al. (2019) introduced generalizability theory³ for assessing data reliability in idiom norming studies. This theory exploits two coefficients: the generalizability

3 For more details on generalizability theory see Brennan (2001, 2003).

(“g-coefficient”) and the dependability coefficient (“d-coefficient”). Following Clayson, Carbine, Baldwin, Olsen, and Larson (2021), the former is low when inter-individual ratings are inconsistent; the latter “is low when measurements of the same individuals are inconsistent” (p. 8). Like the ICC, both coefficients span from 0 to 1.

For each language, we report the ICC (2,k) (calculated using R in the RStudio environment,⁴ by exploiting the psych package), the g-coefficient and the d-coefficient (obtained via the gtheory package. The R script is available in the “supplementary files” folder in the repository). See Table 4 for results on English and Table 5 for results on Italian.

	GENERALIZABILITY THEORY		INTRACLASS CORRELATION COEFFICIENT
	g-COEFFICIENT	d-COEFFICIENT	ICC (2,K)
Familiarity	0.95	0.93	0.93
Meaningfulness	0.95	0.93	0.93
Literal plausibility	0.97	0.97	0.97
Decomposability	0.92	0.91	0.91
Transparency	0.93	0.92	0.92

Table 4 English reliability ratings.

	GENERALIZABILITY THEORY		INTRACLASS CORRELATION COEFFICIENT
	g-COEFFICIENT	d-COEFFICIENT	ICC (2,K)
Familiarity	0.93	0.89	0.89
Meaningfulness	0.94	0.93	0.93
Literal plausibility	0.98	0.97	0.97
Decomposability	0.94	0.92	0.92
Transparency	0.94	0.92	0.92

Table 5 Italian reliability ratings.

Results between 0.75 and 0.9 are indicative of good reliability; results above 0.9 are indicative of excellent reliability (Koo & Li 2016). Tables 4 and 5 report very high rating reliability for both languages, including the typically challenging CBVs, whose assessment is considered to be more fluctuating and difficult (Hubers et al., 2019; Libben & Titone, 2008). These results therefore validate the selected research tools and the multi-part design adopted. Factors contributing to these results may include: the clear variable definitions; the explicit attention to the crowdworkers, manifested through ethical decisions towards them and the valorisation of their work; an effective distribution of cognitive load. These elements ensure a high-quality dataset reusable for future research.

4 CONCLUSIONS AND FUTURE RESEARCH

This paper described the creation of a normed lexicon of 150 pairs of English and Italian idioms annotated by translatability level. The lexicon contains variables that are particularly relevant for the analysis of ambiguous idiomatic contexts where there is interference between literal and figurative meaning. It is stored in the repository of the University of Göttingen and can be explored through three different perspectives. The dataset was created through the implementation of a cross-linguistic norming study carried out entirely online and resorting to crowdsourcing. The reliability analysis validates the research tools used and the design selected.

Future research is two-fold: firstly, the lexicon will be exploited for the selection of idioms to be used in experimental settings to test the relationship between the internal structure of idioms and ambiguous contexts. This will be particularly feasible thanks to the PIA index, which is, for each idiom, the sum of the mean ratings obtained for literal plausibility, decomposability and transparency. Idioms can therefore be selected on the basis of this index, to verify whether idioms that have a high PIA are indeed particularly suitable for occurrence in ambiguous contexts.

⁴ Posit team (2023).

Second, the lexicon serves as a valuable research tool per se, representing the first step to bridge the gap in cross-linguistic research in idiom norming studies. Since the methodology has been explained in detail in the present paper, it can serve as a starting point for the creation of further cross-linguistic idiom datasets. In addition, the present lexicon allows for testing correlations between variables within each language and across languages. It also enables the application of statistical models to explore the relationships between variable sets, facilitating reliable cross-linguistic comparisons, as the definitions and operationalisations used are the same for both languages. Finally, these analyses will facilitate a more systematic understanding of both decomposability and transparency.

DATA ACCESSIBILITY STATEMENT

The normed lexicon of cross-linguistic idioms is in the [repository of the University of Göttingen](#), it is searchable and downloadable. Its Digital Object Identifier (DOI) is [10.25625/EPSWDY](#), the license applied is CC BY-NC 4.0. The data files are in the non-proprietary csv format.

SUPPLEMENTARY FILES

Excerpts of the norming study in pdf format are available in the “supplementary files” folder in the [data repository](#). These files include the consent form, instructions, data quality tracking strategies, Likert scales examples and final debriefing for both sections of the study. The same folder includes the R script to measure data reliability.

ETHICS AND CONSENT

Written informed consent was obtained for each participant. They were guaranteed complete anonymisation of the data. They were fully informed about the purposes of this research, how their responses would be stored, and the possibility of using their data for future publications. The informed consent form can be consulted in the supplementary files.

ACKNOWLEDGEMENTS

Thank you to my supervisors Anke Holler, Caroline Sporleder and Alessandro Lenci for their support. Sincere thanks also to Thomas Weskott for his help in defining the methodology to be implemented, and to Yana Strakatova for her suggestions and constant support. Thank you also to Hedde Zeijlstra for his help and valuable feedback.


FUNDING STATEMENT

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK2636.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATIONS

Irene Pagliai  orcid.org/0009-0003-7994-3916
RTG 2636: Form-meaning mismatches, Georg-August Universität, Göttingen, Germany

REFERENCES

- Beck, S. D. (2020). *Native and non-native idiom processing: Same difference* (Doctoral dissertation, University of Tübingen). DOI: <https://doi.org/10.15496/publikation-46248>
- Boers, F., & Webb, S. (2015). Gauging the semantic transparency of idioms: Do natives and learners see eye to eye. *Bilingual figurative language processing*, 368–392. DOI: <https://doi.org/10.1017/CBO9781139342100.018>

- Brennan, R.** (2001). *Generalizability theory*. DOI: <https://doi.org/10.1007/978-1-4757-3456-0>
- Brennan, R.** (2003). Coefficients and indices in generalizability theory. *Center for advanced studies in measurement and assessment, CASMA research report, 1*, 1–44. DOI: <https://doi.org/10.1111/j.1745-3984.2003.tb01098.x>
- Bulkes, N. Z., & Tanner, D.** (2017). “going to town”: Large-scale norming and statistical analysis of 870 american english idioms. *Behavior research methods, 49*, 772–783. DOI: <https://doi.org/10.3758/s13428-016-0747-8>
- Cacciari, C., & Tabossi, P.** (2014). *Idioms: Processing, structure, and interpretation*. Psychology Press. DOI: <https://doi.org/10.4324/9781315807133>
- Carrol, G., Littlemore, J., & Dowens, M. G.** (2018). Of false friends and familiar foes: Comparing native and non-native understanding of figurative phrases. *Lingua, 204*, 21–44. DOI: <https://doi.org/10.1016/j.lingua.2017.11.001>
- Citron, F. M., Cacciari, C., Kucharski, M., Beck, L., Conrad, M., & Jacobs, A. M.** (2016). When emotions are expressed figuratively: Psycholinguistic and affective norms of 619 idioms for german (panig). *Behavior research methods, 48*, 91–111. DOI: <https://doi.org/10.3758/s13428-015-0581-4>
- Clayson, P. E., Carbine, K. A., Baldwin, S. A., Olsen, J. A., & Larson, M. J.** (2021). Using generalizability theory and the erp reliability analysis (era) toolbox for assessing test-retest reliability of erp scores part 1: Algorithms, framework, and implementation. *International Journal of Psychophysiology, 166*, 174–187. DOI: <https://doi.org/10.1016/j.ijpsycho.2021.01.006>
- Dobrovol'Skij, D.** (2016). The notion of “inner form” and idiom semantics. *Études et travaux d'Eur'ORBEM, 1*(1), 21–35.
- Fadlon, J., Horvath, J., Siloni, T., & Wexler, K.** (2013). The acquisition of idioms: Stages and theoretical implications. *Poster Presentation, Generative Approaches to Language Acquisition, The University of Oldenburg, Oldenburg, Germany.*
- Fellbaum, C.** (2019). How flexible are idioms? A corpus-based study. *Linguistics, 57*(4), 735–767. DOI: <https://doi.org/10.1515/ling-2019-0015>
- Gavilán, J. M., Haro, J., Hinojosa, J. A., Fraga, I., & Ferré, P.** (2021, 07). Psycholinguistic and affective norms for 1,252 spanish idiomatic expressions. *PLOS ONE, 16*, 1–14. DOI: <https://doi.org/10.1371/journal.pone.0254484>
- Geeraerts, D.** (1995). Specialization and reinterpretation in idioms. *Idioms: Structural and psychological perspectives, 57*, 1–14.
- Gibbs, Jr, R. W., & Nayak, N. P.** (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive psychology, 21*(1), 100–138. DOI: [https://doi.org/10.1016/0010-0285\(89\)90004-2](https://doi.org/10.1016/0010-0285(89)90004-2)
- Giora, R.** (1997). Understanding figurative and literal language: The graded salience hypothesis. *cogl, 8*(3), 183–206. DOI: <https://doi.org/10.1515/cogl.1997.8.3.183>
- Haagsma, H.** (2020). *A bigger fish to fry: Scaling up the automatic understanding of idiomatic expressions* (Doctoral dissertation, University of Groningen). DOI: <https://doi.org/10.33612/diss.131057087>
- Hubers, F., Cucchiarini, C., Strik, H., & Dijkstra, T.** (2019). Normative data of dutch idiomatic expressions: Subjective judgments you can bank on. *Frontiers in Psychology, 10*, 1075. DOI: <https://doi.org/10.3389/fpsyg.2019.01075>
- Irujo, S.** (1986, 01). Steering clear: Avoidance in the production of idioms. *IRAL-international Review of Applied Linguistics in Language Teaching – IRAL-INT REV APPL LINGUIST, 31*. DOI: <https://doi.org/10.1515/iral.1993.31.3.205>
- Koo, T. K., & Li, M. Y.** (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine, 15*(2), 155–163. DOI: <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kovecses, Z., & Szabco, P.** (1996). Idioms: A view from cognitive semantics. *Applied linguistics, 17*(3), 326–355. DOI: <https://doi.org/10.1093/applin/17.3.326>
- Langlotz, A.** (2006). Idiomatic creativity. *Idiomatic Creativity* (pp. 1–339). DOI: <https://doi.org/10.1075/hcp.17>
- Libben, M. R., & Titone, D. A.** (2008). The multidetermined nature of idiom processing. *Memory & cognition, 36*, 1103–1121. DOI: <https://doi.org/10.3758/MC.36.6.1103>
- Liontas, J.** (2002). Context and idiom understanding in second languages. *EUROSLA yearbook, 2*(1), 155–185. DOI: <https://doi.org/10.1075/eurosla.2.11lio>
- Maher, Z.** (2013). Opening a can of worms: Idiom flexibility, decomposability, and the mental lexicon. *Yale University MA thesis.*
- Michl, D.** (2019a). Metonymies are more literal than metaphors: Evidence from ratings of german idioms. *Language and Cognition, 11*(1), 98–124. DOI: <https://doi.org/10.1017/langcog.2019.7>
- Michl, D.** (2019b). Speedy metonymy, tricky metaphor, irrelevant compositionality: How nonliteralness affects idioms in reading and rating. *Journal of psycholinguistic research, 48*(6), 1285–1310. DOI: <https://doi.org/10.1007/s10936-019-09658-7>
- Monti, J., & di Buono, M. P.** (2019). Parseme-it: an italian corpus annotated with verbal multiword expressions. *IJCoL. Italian Journal of Computational Linguistics, 5*(5–2), 61–93. DOI: <https://doi.org/10.4000/ijcol.483>

- Moreno, R. E. V.** (2005). Idioms, transparency and pragmatic inference (Tech. Rep.). *UCL Working Papers in Linguistics*, 17, 389–425.
- Nordmann, E., Cleland, A. A., & Bull, R.** (2014). Familiarity breeds dissent: Reliability analyses for british-english idioms on measures of familiarity, meaning, literality, and decomposability. *Acta psychologica*, 149, 87–95. DOI: <https://doi.org/10.1016/j.actpsy.2014.03.009>
- Nordmann, E., & Jambazova, A. A.** (2017). Normative data for idiomatic expressions. *Behavior research methods*, 49, 198–215. DOI: <https://doi.org/10.3758/s13428-016-0705-5>
- Nunberg, G., Sag, I. A., & Wasow, T.** (1994). Idioms. *Language*, 70(3), 491–538. DOI: <https://doi.org/10.1353/lan.1994.0007>
- Pastor, G. C.** (2021, April). Constructional idioms of ‘insanity’ in english and spanish: A corpus-based study. *Lingua*, 254, 103013. DOI: <https://doi.org/10.1016/j.lingua.2020.103013>
- Posit team.** (2023). Rstudio: Integrated development environment for R [Computer software manual]. Boston, MA. Retrieved from <http://www.posit.co/>
- Sailer, M.** (2021, October). *Idioms*. Language Science Press. DOI: <https://doi.org/10.5281/zenodo.5599850>
- Shrout, P. E., & Fleiss, J. L.** (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2), 420. DOI: <https://doi.org/10.1037/0033-2909.86.2.420>
- Tabossi, P., Arduino, L., & Fanari, R.** (2011). Descriptive norms for 245 italian idiomatic expressions. *Behavior Research Methods*, 43, 110–123. DOI: <https://doi.org/10.3758/s13428-010-0018-z>
- 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. DOI: https://doi.org/10.1207/s15327868ms0904_1
- van Ginkel, W., & Dijkstra, T.** (2020). The tug of war between an idiom’s figurative and literal meanings: Evidence from native and bilingual speakers. *Bilingualism: Language and Cognition*, 23(1), 131–147. DOI: <https://doi.org/10.1017/S1366728918001219>
- Vulchanova, M., Milburn, E., Vulchanov, V., & Baggio, G.** (2019). Boon or burden? the role of compositional meaning in figurative language processing and acquisition. *Journal of Logic, Language and Information*, 28, 359–387. DOI: <https://doi.org/10.1007/s10849-019-09282-7>
- Wagner, W.** (2021). *Idioms and ambiguity in context*. De Gruyter. DOI: <https://doi.org/10.1515/9783110685459>
- Wulff, S.** (2013). Words and Idioms. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 274–289). Oxford University Press. DOI: <https://doi.org/10.1093/oxfordhb/9780195396683.013.0015>

TO CITE THIS ARTICLE:

Pagliai, I. (2023). Bridging the Gap: Creation of a Lexicon of 150 Pairs of English and Italian Idioms Including Normed Variables for the Exploration of Idiomatic Ambiguity. *Journal of Open Humanities Data*, 9: 16, pp. 1–13. DOI: <https://doi.org/10.5334/johd.123>

Submitted: 26 July 2023

Accepted: 28 September 2023

Published: 25 October 2023

COPYRIGHT:

© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.