**This is a pre-print of an article published in Behavior Research Methods. The final authenticated version is available online at https://doi.org/10.3758/s13428-019-01337-8**

# Perceptual modality norms for 1121 Italian words: a comparison with concreteness and imageability scores and an analysis of their impact in word processing tasks.

Alessandra Vergallito[1,2], Marco Alessandro Petilli[1], Marco Marelli[1,2]

[1] Dipartimento di Psicologia, Università degli Studi di Milano-Bicocca

[2] NeuroMi, Milan Center for Neuroscience

Corresponding author

Alessandra Vergallito

alessandra.vergallito@unimib.it,

Department of Psychology, University of Milano Bicocca,

Piazza Ateneo Nuovo, 1, 20126 Milano, Italy.

**Abstract**

Normative measures of verbal material are fundamental in psycholinguistic and cognitive research to control for confounding in experimental procedures and achieve a better comprehension of our conceptual system. Traditionally, normative studies focused on classical psycholinguistic variables, such as concreteness and imageability. Recent works shifted researchers' focus to perceptual strength, in which items are separately rated for each of the five senses.

We present a resource including perceptual norms for 1121 Italian words extracted from the Italian version of ANEW. Norms were collected from 57 native-speakers. For each word, participants provided perceptual strength ratings for each of the five perceptual modalities. Perceptual norms performance in predicting human behavior was tested in two novel experiments, a lexical decision and a naming task. Concreteness, imageability and different composite variables representing perceptual strength scores were considered as competing predictors in a series of linear regressions, evaluating the goodness-of-fit of each model.

For both tasks, the model with *imageability* as predictor was found to be the best fitting model according to AIC, while the model with the separately considered *five modalities* better described data according to the explained variance. These results differ from the ones previously reported for English, in which maximum perceptual strength emerged as the best predictor of behavior. We investigated this discrepancy by comparing Italian and English data on the same set of translated items, thus confirming a genuine cross-linguistic effect. We conclude confirming that perceptual experience influences linguistic processing, even though evaluations from different languages are needed to generalize this claim.

*keywords*

embodied cognition, perceptual strength, concreteness, imageability, modality exclusivity

## 1. Introduction

Normative measures of verbal material are of special interest in psycholinguistics and cognitive research, where they are used to control for confounding variables and create balance item sets in experimental procedures, and to achieve a better comprehension of the organization of our conceptual system. Traditionally, normative studies include classical psycholinguistic variables, such as word frequency, affective properties, orthographic/phonological metrics, concreteness, or imageability ratings. Referring to concreteness and imageability, the two constructs (and their associated ratings) have been often used interchangeably by the literature in the field (e.g. Kousta et al., 2011; Connell and Lynott, 2012), due to their high correlation and theoretical relationship. However, the two concepts reflect, at least partially, different aspects of semantic representations, with concreteness representing the degree to which word referent refers to a perceptible entity and imageability scores strongly correlating with a concept visual property (Brysbaert et al. 2014; Connell & Lynott, 2012, 2015).

Despite the importance assigned to these variables in facilitating word processing (the well-known *concreteness effect*, e.g. Paivio, 1991), imageability and concreteness failed to explain and predict human behaviour in a conclusive way, with evidence pointing out the opposite facilitation (i.e. *abstractness effect*, e.g. Kousta et al., 2011) or no effects (e.g. Barca et al., 2002). The inconsistency of empirical data brought to the idea that both concreteness and imageability could be considered noisy measures (Connell and Lynott, 2012), which do not offer an accurate approximation of the perceptual basis of concepts.

At the same time, the last decades saw the prospering of research within the embodied cognition framework, suggesting a strong involvement of the sensorimotor system in language comprehension[1] (see Meteyard et al., 2012; Pulvermüller, 2018 for a recent review). This evidence determined the prospering of questionnaires investigating the perceptual and motor features of a word's referent (e.g., Juhasz et al., 2011; Lynott and Connell, 2009, 2013; see Lynott et al., 2019 for the largest norm dataset). For example, sensory experience ratings (SER, Juhasz et al., 2011) are aimed at capturing the extent to which a certain word evokes a sensory and/or perceptual experience in the reader's mind. To validate the obtained resources and provide evidence in favor of their relevance for psychological studies, such variables are typically tested against human performance, in particular response latencies obtained in chronometric studies with word stimuli. Juhasz and

[1] In its strongest formulation, indeed, embodied theory claims that conceptual representations are encoded in a sensorimotor format (e.g. Glenberg, 2015) and language comprehension involves the re-activation of the sensorimotor states acquired during previous experiences or interactions with word referents (Cappa & Pulvermuller, 2012; Glenberg & Gallese, 2012).

colleagues (2011), for example, collected SER for over 2850 words and tested it against lexical-decision data for monosyllabic words from two English mega-studies (Balota et al., 2004; Keuleers et al., 2012). Authors found that words with higher SER elicited faster and more accurate responses as compared to words with lower SER. Several studies addressed and replicated this point (Juhasz & Yap, 2013; see Bonin et al., 2015 for similar results in French) extending results also to noun-noun compounds (Kuperman, 2013) and semantic tasks (Zdrazilova & Pexman, 2013).

It is crucial to note that in the aforementioned studies, participants were instructed to evaluate the degree to which a certain word evoked a general sensory experience, without distinguishing among the five senses. Such choice leaves to the participants' initiative to consider all the different modalities through which an object can be experienced, with the potential limitation of leading to an underspecified characterization of the variable of interest or to an overestimate of one perceptual modality as compared to the others (Connell and Lynott, 2016; Lynott et al., 2019). A stronger measure, in this respect, is obtained by asking participants to rate the perceptual strength of a given word separately for the five senses (Lynott and Connell, 2009, 2013).

In the last few years perceptual modality norms of this kind have spread widely, becoming available in many different languages, such as Dutch (Speed and Majid, 2017), Russian (Miklashevsky, 2018), and Mandarin (Chen et al., 2019), and their validity has been tested with several experimental paradigms. For example, Speed and Majid used perceptual strength norms in a similarity judgment task, finding that words from the same dominant modality were rated more similar than words from different dominant modalities, and such effect was enhanced for word pairs with higher ratings. Moreover, they investigated whether perceptual modalities were differently experienced in spatial terms, thus running a lexical decision experiment with word spatial position presented in proximal or distal space. Interestingly, they found that words dominant in olfaction were processed faster in the proximal than distal space as compared to the other modalities, suggesting that olfactory information is mentally simulated as close to the body. Moreover, perceptual norms have been validated in modality-switch costs tasks, in which participants are typically asked to verify a series of properties of a concept (e.g. TIGER-striped *visual*). The behavioral pattern shows that participants are slower when the following target concerns a different modality (e.g. WHISTLE-shrill, *auditory*) as compared to the same perceptual modality (e.g. CANDLE-flickering) (e.g. Pecher et al., 2003; Van Dantzig et al., 2008; Vermeulen et al., 2007).

But to what extent perceptual strength ratings reflect concreteness and imageability and can explain human performance?

To the best of our knowledge, only Connell and Lynott (2012) investigated this issue, comparing perceptual modality ratings with concreteness and imageability scores and testing the three measures as competing predictors of participants' performance in word recognition tasks. Their findings suggested that the *maximum perceptual strength*, namely the rating value of the dominant perceptual modality, predicted accuracy and reaction times better than concreteness and imageability. However, at present, these results have not been replicated on languages different from English. This rests uncomfortably with the evidence that ratings concerning the properties of word-denoted objects also reflect lexical statistics (as captured from models trained on text corpora, Hollis & Westbury, 2016). In fact, if when producing intuitions about referents participants are influenced by distributional properties of their associated words, it is conceivable that different linguistic experiences (as being exposed to a given language or another) might result in slightly different distributions in semantic norms. Given these considerations, it becomes crucial to search for cross-linguistic evidence concerning the impact of rating norms on language-processing data.

In the present work, we first describe a new resource including perceptual-modality norms for Italian (following Lynott and Connell, 2013). These new data ideally complement the largest norming work currently available in Italian, namely the *Affective norms for English words* (*ANEW*, Bradley & Lang, 1999) Italian adaptation by Montefinese and colleagues (2014). The dataset is composed of 1121 words, of which 1034 are the Italian translations of the ANEW stimuli, and 87 are based on a previously published database (Montefinese et al., 2013). The Italian ANEW includes rating-based norms for three affective variables, namely valence, arousal, and dominance, as well as familiarity, imageability, and concreteness. However, it lacks more specific estimates concerning the perceptual properties of the included words: no information concerning the perceptual experience associated with the five senses is provided, hence the perceptual strength of the stimuli cannot be estimated. The norms presented here include perceptual-strength estimates for the 1121 words of the ANEW database. In the second section, we specifically investigate whether Connell and Lynott's results (2012) can be extended to other languages by comparing the effect of perceptual strength to the one of concreteness and imageability in two novel experiments (lexical decision and word naming tasks) on Italian. Having found this is not the case, in the third section of the paper, we test whether the emerged dissociation between English and Italian can be considered a genuine cross-linguistic effect or is more trivially due to differences in item selection, thus comparing Italian and English datasets including the same (translated) words.

## 2. Part 1: Perceptual modality norms for 1121 Italian words

In this section, we present the perceptual-modality ratings collected for 1121 words from Italian native speakers. We also compare these ratings to concreteness and imageability scores (as measured by Montefinese et al., 2014). If concreteness and imageability are a pure reflection of the degree of perceptual information in a concept, their scores should be positively related to perceptual strength ratings in all the five modalities. On the other hand, following the findings of Lynott and Connell (2013), it is also conceivable that concreteness and imageability reflect some perceptual modality more than the others.

## 2.1 Methods

### *2.1.1 Participants*

57 students (males = 28; *Age* = 23.6 ± 5.2) of the University of Milano-Bicocca took part in the experiment in exchange of course credit. Participants were Italian native speakers. The study was approved by the local ethical committee and participants' ethical treatment was in accordance with the principles stated in the Declaration of Helsinki.

### *2.1.2 Materials*

The item set contained the 1121 items from the Montefinese dataset (Montefinese et al., 2014). It comprises 20% of adjectives, 69% of nouns, 5% of verbs and a 6% of words which could be considered both as an adjective or a noun. Trial-by-trial data were released as Supplementary materials (https://osf.io/zdg59/).

### *2.1.3 Procedure*

Items were randomly presented to participants for perceptual strength ratings in a norming procedure based on Lynott & Connell (2009). Each word was presented on a separated screen, in a sentence that reported "To what extent do you experience WORD" (with the WORD slot being filled with a noun or verb target) or "to what extent do you experience something being WORD" (with the WORD slot being filled with an adjective target or a target that could be considered both an adjective

and a noun[2]). The sentence was completed underneath by five endings, corresponding to the five perceptual modalities: "by feeling through touch", "by hearing", "by seeing", "by smelling", and "by tasting". Each of these endings was paired with a rating scale. An example of trial is reported in Figure 1.


Figure 1. Screen capture of an experiment trial. In the example, the participant was asked "To what extent do you experience *MANICHINO* (dummy) by hearing, by tasting, by feeling through touch, by smelling, by seeing. Participants replied on a Likert scale going from 0 (not at all) to 5 (greatly). At the top of the screen participants could check the questionnaire progression. At the bottom of the screen, left and right arrows allowed participants to go back to the previous item or to move on once they have completed the trial.

Participants were hence asked to rate the extent to which they would perceive the referent of each word through each of the five senses, on a scale ranging from 0 (not at all) to 5 (greatly). The numerical rating scale was displayed with no default value selected, and participants clicked on a number to indicate their preference. Once each word had been rated on all five modalities, participants clicked an arrow placed at the bottom of the screen, in order to move to the following item. Participants were told to evaluate each item using their own judgment because there was no pre-determined right or wrong answer. They were also instructed to skip items with which they were unfamiliar, moving directly to the following item. The experiment was self-paced as participants were able to take a break every time they desire.

[2] Since the task-question formulation induced an adjectival interpretation for the ambiguous stimuli, from now on we classify as adjectives words which could be deemed as both adjectives and nouns. Hence, our dataset includes 69.4% nouns, 25.4% adjectives, 5.2 % verbs.

Differently from previous studies, which divided the item set in different sub-questionnaires (e.g., Bonin et al., 2015) or asked participants to rate only one dimension (e.g., color or smell, Díez-Álamo et al., 2018), in the present experiment, all participants rated the full item set. The order of words was randomized across participants. The order of the modalities was fixed across items for each participant but was counterbalanced across participants in a latin-square design. The experiment was administered online using the software Qualtrics (Provo, UT). Experiment links were sent to participants via e-mail so that they could fill questions using their personal laptop, tablet or smartphone.

As sanity-check, we selected 81 items that were unambiguously experienced through one sense more than one other (e.g. a *tavolo* - *table* is more likely to be experienced through sight than through taste). On these items we evaluated participants' accuracy to control that they paid attention to the task and did not answer randomly. Response accuracy for these sanity-check stimuli was higher than 80% for all participants (mean = 96.8, SE = 0.54).

## 2.2 Results

### *2.2.1 Perceptual modality norms*

Participants' ratings were collapsed, excluding missing trials (0.6 % of the data), and for each word, average values were calculated separately for each modality, resulting in a dataset comprising 5605 unique data points. In Table 1, we report rating means, standard deviations and standard errors for each of the five modalities.

| Modality | M | SD | SE |
|---|---|---|---|
| Auditory | 2.28 | 1.29 | 0.04 |
| Gustatory | 0.55 | 0.93 | 0.03 |
| Haptic | 2.11 | 1.42 | 0.04 |
| Olfactory | 0.95 | 1.02 | 0.03 |
| Visual | 4.01 | 0.86 | 0.03 |

Table 1. Mean ratings, standard deviations, and standard errors of perceptual strength (on a 5-point scale) across the five modalities.

Each item was assigned as dominant modality (visual, haptic, auditory, olfactory or gustatory), the modality which received the highest mean rating (Lynott and Connell, 2009). As in Lynott and Connell (2009), where ties existed for the strongest modality (11 items out of 1121, see Table 2) one of the tied modalities was randomly chosen as the dominant one.

| Stimulus | English translation | Grammatical class | Auditory | Gustatory | Haptic | Olfactory | Visual |
|---|---|---|---|---|---|---|---|
| colpa | fault | noun | 2.49* | 0.07 | 0.56 | 0.12 | 2.49 |
| commedia | comedy | noun | 4.25* | 0.04 | 0.09 | 0.07 | 4.25 |
| cuscino | pillow | noun | 0.72 | 0.07 | 4.75 | 1.33 | 4.75* |
| infastidire | annoy | verb | 3.68 | 0.95 | 2.45 | 1.93 | 3.68* |
| promozione | promotion | noun | 2.86* | 0.14 | 0.18 | 0.14 | 2.86 |
| disperato | despairing | adjective | 3.93 | 0.23 | 0.55 | 0.32 | 3.93* |
| gentile | gentle | adjective | 3.75 | 0.47 | 1.51 | 0.61 | 3.75* |
| idiota | idiot | adjective | 3.75* | 0.19 | 0.56 | 0.27 | 3.75 |
| onesto | honest | adjective | 3.23* | 0.21 | 0.60 | 0.28 | 3.23 |
| sculacciata | spanking | noun | 3.54 | 0.02 | 4.35* | 0.09 | 4.35 |
| ulcera | ulcer | noun | 0.32 | 0.16 | 2.02 | 0.16 | 2.02* |

Table 2. Words in which ratings of perceptual strength revealed non-unique dominant modality. The asterisk following the mean rating indicates the randomly assigned dominant modality.

Table 3 represents the distribution of modality dominance across items, showing their strength with respect to the other perceptual modalities and their exclusivity scores. Modality exclusivity indicates the extent to which a certain item is perceived through a single perceptual modality. Where each item has a vector containing mean ratings for the five modalities, modality exclusivity is calculated as the range of values divided by their sum, according to the formula

$$\frac{\max(x) - \min(x)}{\sum(x)} * 100$$

Where *x* is a vector of mean ratings for each of the five perceptual modalities. In such a way, modality exclusivity scores in principle can range from 0% to 100%, where an entirely multimodal property

(scoring equally strongly on all perceptual modalities) will have the lowest modality exclusivity score of 0% and an entirely unimodal properties (scoring zero on all but one perceptual modality) will have the highest modality exclusivity score of 100%. The actual resulting scores ranged from 2.8% (item *piacere, pleasure*) to 96.1% (item *arcobaleno, rainbow*), with an overall mean of 40.6% (SD= 12.8%, see Table 3).

| | | N. item | Average Modality exclusivity score | Mean Auditory Rating | Mean Gustatory Rating | Mean Haptic Rating | Mean Olfactory Rating | Mean Visual Rating |
|---|---|---|---|---|---|---|---|---|
| *Dominant modality* | Auditory | 106 | 45% | 3.70 | 0.25 | 0.81 | 0.31 | 2.93 |
| | Gustatory | 27 | 27.9% | 0.82 | 4.71 | 2.62 | 3.27 | 3.92 |
| | Haptic | 32 | 40% | 1.39 | 0.57 | 4.20 | 0.54 | 3.28 |
| | Olfactory | 9 | 46.1% | 0.52 | 1.27 | 1.16 | 4.51 | 2.37 |
| | Visual | 947 | 40.5% | 2.21 | 0.46 | 2.18 | 0.93 | 4.17 |

Table 3. Numbers of words and exclusivity scores (as percentage) per dominant modality, along with the mean ratings of perceptual strength (0-5) in each modality.

Table 4 represents the distribution of the obtained ratings separately for the grammatical class of the item. A chi-squared test of the modality distributions across grammatical classes was not significant ($X_{2(8)} = 12.8$, $p = .120$), suggesting that item distribution did not differ among the three linguistics categories.

| | Modality | M | SD | SE | N. Item |
|---|---|---|---|---|---|
| *Nouns (n= 778)* | Auditory | 2.14 | 1.34 | 0.05 | 70 |
| | Gustatory | 0.53 | 0.97 | 0.03 | 21 |
| | Haptic | 2.37 | 1.44 | 0.05 | 19 |
| | Olfactory | 1.03 | 1.07 | 0.04 | 7 |
| | Visual | 4.09 | 0.93 | 0.03 | 661 |
| *Adjectives (n=285)* | Auditory | 2.64 | 1.12 | 0.07 | 24 |
| | Gustatory | 0.61 | 0.85 | 0.05 | 5 |
| | Haptic | 1.42 | 1.15 | 0.07 | 10 |
| | Olfactory | 0.72 | 0.85 | 0.05 | 2 |
| | Visual | 3.85 | 0.65 | 0.04 | 244 |
| *Verbs (n=58)* | Auditory | 2.47 | 0.98 | 0.13 | 12 |
| | Gustatory | 0.57 | 0.82 | 0.11 | 1 |
| | Haptic | 1.96 | 1.11 | 0.15 | 3 |
| | Olfactory | 0.92 | 0.92 | 0.12 | 0 |
| | Visual | 3.64 | 0.56 | 0.07 | 42 |

Table 4. Mean ratings of perceptual strength (0-5) across the five modalities in each grammatical class, with standard deviations, standard errors, and number of dominant items for each modality.

Concerning the relationship among different modalities, not all perceptual modalities were equally distinct, as it is shown in the correlation matrix (Bonferroni corrected) reported in Table 5, as well as the scatterplot of dominant-modality clusters reported in Figure 2. Significant correlations were found for most of the modality pairs, although most of them were weak to moderate.

| Rated Modality | Auditory | Gustatory | Haptic | Olfactory | Visual |
|---|---|---|---|---|---|
| Auditory | -- | -.08 | -.36*** | -.11** | -.17*** |
| Gustatory | | -- | . 14*** | . 59*** | -.01 |
| Haptic | | | -- | .29*** | .57*** |
| Olfactory | | | | -- | .22*** |
| Visual | | | | | -- |

Table 5. Correlations between perceptual-strength scores in different modalities (Bonferroni corrected). ***p<.001 **p<.01 *p<.05

In Figure 2 ratings on the five modalities have been reduced to 2 dimensions using principal components analyses (singular value decomposition, explaining 68 % of the original variance).

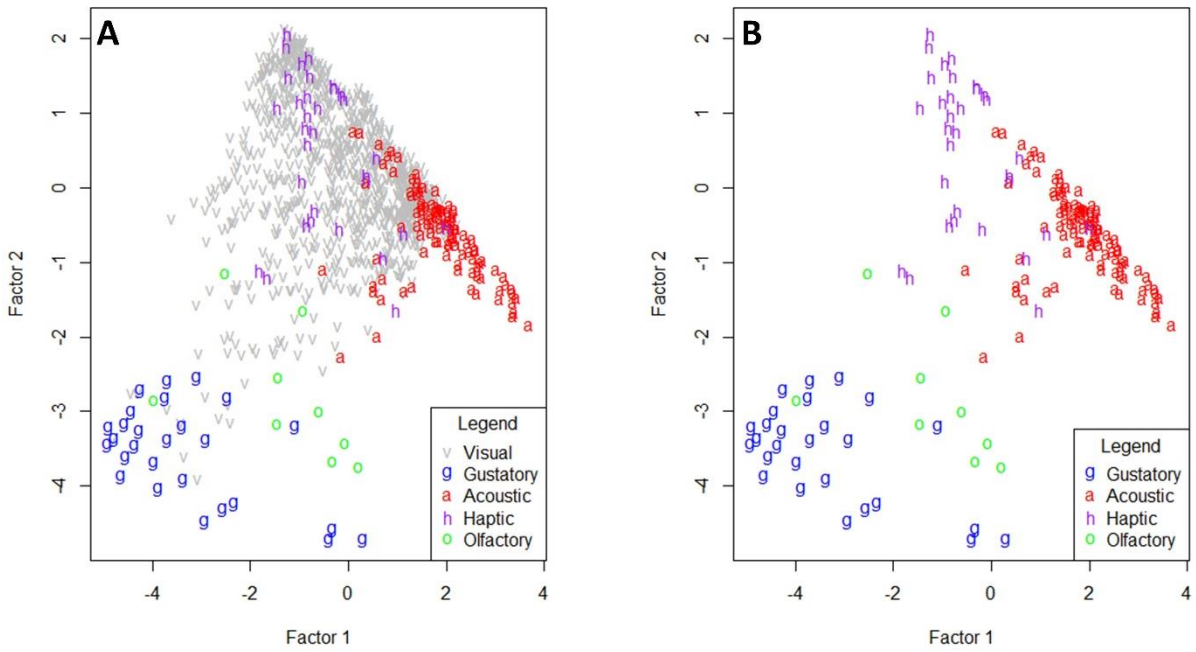


Fig. 2 Clustering of words dominant in auditory, haptic, gustatory, olfactory, and visual modalities on two factors extracted from the factor analysis with principal components (Panel A). To better appreciate the relationships between the other modalities, the same plot is reported in Panel B excluding visually dominant items.

The comparison between panels 2A and 2B clearly highlights that most items were rated by participants as most experienced through the visual modality. This modality was so preponderant that

it encapsulated most of the other modalities, especially the haptic and auditory ones, indicating that many items which are experienced through the touch and the hearing can also be experienced by the visual modality. Gustatory and olfactory modalities, instead, were relatively separated from visual elements, in line with the results indicating a non-significant correlation between gustatory and visual scores and a weak correlation between olfactory and visual ones. When not considering items classified as *visually dominant* (panel B), the four remaining modalities showed patterns that are relatively segregated from each other.

Correlation-wise (Table 5), the strongest positive relationships were observed between the olfactory and gustatory modalities, which is not surprising given their *chemical-sense* status, and between haptic and visual modalities, showing that objects which can be touched can also be seen. Auditory modality correlated negatively with all the other modalities, which, together with its distinct cluster in Figure 2, suggests that the higher a given word is experienced through hearing, the lower the same word is experienced through the other sensory modalities.

Exclusivity scores differed across dominant modalities according to an ANOVA ($F(4,1116)=10.5$, $p<.001$): post-hoc analysis with Bonferroni corrections showed that properties with gustatory dominance scored lower in modality exclusivity as compared to those for all other perceptual modalities (all *ps*>.528).

### *2.2.2 Relationship between perceptual strength ratings and concreteness / imageability scores*

As second step, we investigated whether concreteness and imageability reflect the perceptual properties of a word referent, or they represent different information. More specifically, if concreteness and imageability summarize the perceptual features of a word, we should find concreteness and imageability to be positively correlated with ratings for all the perceptual modality.

In line with this reasoning, we found imageability and concreteness being highly correlated (see Table 6), suggesting that in our database they capture the same latent variable. Haptic and visual perceptual modalities had a strong correlation with both concreteness and imageability. Interestingly, visual perceptual ratings correlated in the same way with concreteness and imageability, while haptic modality correlated more with concreteness scores as compared to imageability. The relationship between olfactory modality and concreteness and imageability was significant but weak, while auditory modality was negatively correlated to both concreteness and imageability, suggesting that word-denoted objects which can be experienced through hearing are considered more abstract and

less imaginable. Gustatory modality did not show significant correlation with concreteness nor with imageability.

| | Concreteness | Imageability | Auditory | Gustatory | Haptic | Olfactory | Visual |
|---|---|---|---|---|---|---|---|
| Concreteness | - | .88*** | -.30*** | .00 | .69*** | .26*** | .66*** |
| Imageability | | - | -.21*** | .02 | .59*** | .23*** | .66*** |
| Auditory | | | - | -.08 | -.36*** | -.11** | -.17*** |
| Gustatory | | | | - | .14*** | .59*** | -.01 |
| Haptic | | | | | - | .29*** | .57*** |
| Olfactory | | | | | | - | .22*** |
| Visual | | | | | | | - |

Table 6. Correlation between concreteness, imageability, and mean perceptual strength ratings for each modality predictor in study 1 (N =1121). Asterisks represent p-value adjusted with Bonferroni corrections, *p<.0.5, **p<.01, ***p<.001.

As second step, we investigated whether our ratings were good predictors of concreteness and imageability. We ran stepwise regression analysis using a backward procedure with either concreteness or imageability ratings as dependent variable and ratings of auditory, gustatory, haptic, olfactory, and visual perceptual strength as predictors. In both cases the model comprising the five perceptual modalities was found to be the best one.

While all the five perceptual modalities contributed to the regression model, the direction of the relationship varied across modalities (see Table 7). Auditory and gustatory ratings were negatively related to concreteness: the more strongly a word referent was related to taste and sound experiences, the less concrete it was. At the opposite, haptic and visual modalities showed the strongest positive relation with concreteness, followed by olfactory ratings.

Imageability predictors were not totally overlapping with the ones of concreteness. Despite also in this regression visual and haptic modality were the best predictor for imageability scores, followed by olfaction, auditory modality did not predict imageability ratings. The gustatory modality

was the only one to have a negative effect on imageability, suggesting that words that are experienced through the taste are less easy to be imagined.

| Dependent variable | Auditory | Gustatory | Haptic | Olfactory | Visual |
|---|---|---|---|---|---|
| Concreteness | -3.804*** | -6.129*** | 17.489*** | 5.476*** | 15.891*** |
| Imageability | -0.727 | -2.442* | 11.254*** | 2.666** | 17.455*** |

Table 7. T-values for each modality of perceptual strength as model predictor of concreteness and imageability. Asterisks indicate p-values ***p<.001 **p<.01 *p<.05

## 2.3 Discussion

In the present work, we collected perceptual modality ratings for Italian, with the aim of releasing a new resource as a complement of Italian adaptation of the ANEW database.

Following the original works by Lynott and Connell on English (2009, 2013), perceptual strength norms have been also collected in different languages including Russian (Miklashevsky, 2018), Dutch (Speed and Majid, 2017), Mandarin (Chen et al., 2019). In order to allow a more straightforward comparison among studies, we reported in Table 8 a summary of the results on the available studies conducted with perceptual strength across languages.

| Perceptual modality | English Adjectives[a] | | English Nouns[b] | | Dutch Nouns[c] | | Mandarin Adjectives[d] | | Russian nouns[e] | | Italian words | | English Lancaster scale[f] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M.E. | N. | M.E. | N. | M.E. | N. | M.E. | N. | M.E. | N. | M.E. | N. | M.E. | N. |
| Auditory | 57% | 68 | 44.10% | 42 | 51% | 37 | 47% | 13 | 21.5% | 81 | 45% | 106 | 44.20% | 4528 |
| Gustatory | 35% | 55 | 24.60% | 6 | 36% | 120 | 37% | 19 | 19.5% | 48 | 27.9% | 27 | 29.50% | 890 |
| Haptic | 37% | 70 | 35.30% | 14 | 46% | 35 | 35% | 43 | 21.1% | 108 | 40% | 32 | 37.40% | 975 |
| Olfactory | 43% | 25 | 14.60% | 2 | 41% | 27 | 38% | 5 | 18.2% | 16 | 46.1% | 9 | 40.70% | 216 |
| Visual | 49% | 205 | 39.1% | 336 | 52% | 261 | 54% | 91 | 20.2% | 253 | 40.5% | 947 | 44.80% | 29552 |

Table 8. Modality exclusivity norms and number of items for each of the five perceptual modalities. a) Lynott and Connell, 2009; b) Lynott and Connell, 2013; c) Speed and Majid, 2017; d) Chen et al., 2019; e) Miklashevsky, 2018; f) Lynott et al., 2019.

First, we demonstrated that our sample of words was experienced in a multimodal way. The multimodal composition of words has been supported by perceptual ratings of English adjectives, nouns and verbs (Lynott and Connell, 2009, 2013; Van Dantzig et al., 2011; Winter, 2016) and Dutch and Russian nouns (Speed and Majid, 2017; Miklashevsky, 2018). In line with previous norming ratings, we replicated a visual dominance effect (Lynott and Connell, 2009, 2013; Van Dantzig et al., 2011; Winter, 2016; Chen et al., 2019) with Italian speakers. Moreover, similarly to previous studies by Lynott et al. (2019), Lynott and Connell, (2009) and Speed and Majid (2017), we found gustation to be the most multimodal sense. Different findings were reported for Mandarin, in which the haptic modality was the most multimodal one, and English nouns, in which olfactory modality was the less exclusive modality. Russian nouns, on the other hand, received generally higher multimodal scores as compared to other norming datasets. Only in our norming dataset olfaction received the highest exclusivity rating, indicating that concepts that can be experienced by smelling are less experienced with the other four perceptual modalities. Auditory perceptual modality received high exclusivity scores as well, in line with previous studies (Connell and Lynott, 2012; Lynott and Connell, 2013; Miklashevsky, 2018). Visual and haptic modalities shared the third position after olfactory and auditory ones. While haptic average-level multimodality was in line with previous results (except Mandarin items), visual modality was rated as the most unimodal in Dutch and Mandarin words. Such

heterogeneous patterns between norming studies may be due to differences in the selection criteria used for the item list composition. Indeed, in Speed and Majid (2017) and Lynott and Connell (2009), items were selected in order to cover equally all the five perceptual modalities, while Miklashevsky (2018) selected items representative of specific categories (e.g. animals, tools, emotions).

Concerning the relationship among the perceptual variables, the strongest positive correlation was observed between olfactory and gustatory modalities, which is consistent across languages. Similarly, haptic and visual perceptual ratings were positively related across the different samples, reflecting that concepts that can be touched can also be seen (e.g., Lynott et al., 2019). Auditory modality correlated negatively with all the other perceptual experiences**.** This negative relationship between the auditory modality and the other ones seems to be a robust pattern across different language and datasets (English, Lynott and Connell, 2009; Connell and Lynott, 2012; Lynott et al., 2019; Russian, Miklashevsky, 2018; Dutch, Speed and Majid, 2017).

As a second step, we compared perceptual ratings with two traditional psycholinguistic variables, namely concreteness, and imageability. Our aim was to investigate whether concreteness and imageability reflected the degree of concepts perceptual information or were predicted by some sensorial modalities more than others. In line with previous studies on English and Russian, we found haptic and visual modalities to be the strongest predictors of both concreteness and imageability.

Taken together, our results highlighted similarities in perceptual ratings across different languages, which may reflect the way in which we experience and interact with our environment.

## 3. Part 2: behavioral evaluation of the collected norms

After norms were collected, we tested their validity in predicting chronometric data. In order to do so, we ran two novel word-processing studies, namely a lexical decision and a naming task, and tested which measures among concreteness, imageability, and different operationalizations of perceptual strength, are better at explaining human performance.

### 3.1 Lexical decision task

#### 3.1.1 Methods

##### *3.1.1.1 Participants*

33 psychology students from the University of Milano-Bicocca (males = 6; *Age* = 23 ± 4.98; *Education*= 14.8 ± 1.49) took part in the experiment in exchange of course credits. Participants were Italian native speakers and were naïve to the experiment purpose. The study was approved by the departmental ethical committee, and participants' ethical treatment was in accordance with the principles stated in the Declaration of Helsinki.

#### *3.1.1.2 Materials*

Word sample comprises the same 1121 words used for the normative ratings (Montefinese et al., 2014), plus 1121 pseudowords matched with the lexical stimuli for orthographic length. Pseudowords were created using the WUGGY software (http://crr.ugent.be/programs-data/wuggy, Keuleers & Brysbaert, 2010), a multilingual pseudoword generator able to create orthographic strings that respect the orthotactic rules of a given language.

For each item, we extracted imageability (M = 6.98, SD ± 1.16, SE = .03), and concreteness (M = 6.21, SD = 1.66, SE = .05) scores from the ANEW dataset (Montefinese et al., 2014). Lexical frequencies (M = 4706.7, SD = 13957.3, SE = 416.9), on the other hand, were obtained from subtlex-it (http://crr.ugent.be/subtlex-it/).[3] Trial-by-trial data of words and pseudowords are released in our Supplementary materials.

#### *3.1.1.3 Procedure*

Participants took part in a two-session experiment, with each session lasting about an hour. The two sessions took place at the same time of the day at a maximum temporal distance of two weeks. After receiving information about the experimental procedure, participants were asked to sign the written informed consent. Participants were then sat in front of a 17" computer screen. They were informed that they would have been presented a string of letter at the center of the screen that could be either a word or a non-word and that they would have been asked to press the "N" key of the keyboard if the stimulus presented was a word and the "C" key if the stimulus was a pseudoword. Participants were asked to keep their index fingers over the two keys and to respond as fast as possible after word presentation.

A practice sequence took place at the beginning of each session, including 10 words and 10 pseudowords in a randomized order. Only in this phase participants received visual feedback after

[3] For ten items, most of them multi-word expressions (e.g. *capro espiatorio, scapegoat*), a frequency norm was not available. In these cases, we assigned to the item a frequency of 0.

each trial informing them about their accuracy and response time. The two experimental sessions were composed of 1120 and 1122 trials (for a total of 2242 trials for each participant), and each of them included a break after the first 560 trials. Each trial started with a fixation cross of 500 ms presented at the center of the screen. Subsequently, a written letter string (a word or a pseudoword) was presented for a maximum duration of 2000 ms (the string disappeared as soon as the software recorded participants' response), followed by a blank screen with a fixed duration of 1500 ms. The order of the stimuli was randomized across participants. The experimental procedure was implemented in E-Prime 3 (Psychology Software Tools Inc., Pittsburgh. PA). Accuracy and reaction times were recorded.

#### *3.1.1.4 Statistical design and analysis*

One participant systematically inverted the response keys in the first session of the experiment, while data of two participants were partially lost because of a power shortage during data collection. Data of these 3 participants were then removed by subsequent analysis. By-item average reaction times (RTs) were then computed. Before aggregating RTs, we removed non-word items, incorrect responses (1874 data points), and RTs inferior to 100 ms (6 data points). Raw RTs were then logarithmically transformed and converted into z-scores (over participant and session), following standard procedures in the literature on word recognition (Baayen, 2008; Balota et al., 2007), and finally by-item average latencies were computed. This procedure ensures a more reliable measure of latency, accounting for individual differences in overall speed and variability (Balota et al., 2007). The dataset used for these analyses is available as Supplementary Material.

Statistical analyses were performed in the statistical environment R (R Core Team, 2008, https://www.R-project.org). We ran a series of linear regressions with RTs[4] as dependent variable. First, we fitted a baseline model with log-transformed item frequency and orthographic length as predictors (for a similar procedure see Brysbaert and New, 2009). Then, we separately investigated the impact on RTs of each of interest, namely concreteness, imageability and different operationalizations of perceptual strength, derived from the scores in the five perceptual modalities.

We considered different measures, following Connell and Lynott (2012) and Lynott et al. (2019) procedure, in order to compute composite variables that reduce the 5-dimension profile.

[4] Given the low rate of errors (average accuracy was 94.4%), accuracy was not further investigated.

a) *Five perceptual modalities:* the five perceptual modalities were separately added in the regression predictors.
b) *Maximum perceptual strength*: it corresponds to the highest score across the five perceptual modalities. It has been suggested to be the best composite variable of perceptual strength (Connell and Lynott, 2012, 2016; Connell et al., 2018).
c) *Minimum perceptual strength*: opposite to the previous one, it returns the minimum score across the five perceptual modalities.
d) *Mean perceptual strength*: it represents the mean value of the ratings in the five perceptual modalities. It is equivalent to the *summed strength* previously used by Connell and Lynott (2012) and Lynott et al. (2019) and considers all dimensions as equally important.
e) *Magnitude of perceptual strength* or *Euclidean vector length:* it corresponds to the length of the vector, including the scores for the five perceptual modalities (for details see Lynott et al., 2019).
f) *Minkowski 3 distance*: it reflects the perceptual strength in all the five dimensions, but the influence of weaker dimensions is attenuated. It has been suggested to be the best composite measure to account for multisensory integration in perception (To et al., 2010), and it has been showed to be the best candidate to predict reaction times and accuracy in lexical decision tasks (Lynott et al., 2019).

The six measures were separately added to the baseline model, and separate regressions were computed. We compared the resulting regression models in terms of goodness-of-fit, i.e., their ability to explain variance in behavioral performance as compared to the baseline model. For each of the regression model we calculated the r-squared value, the Akaike information criterion (AIC; Akaike, 1973; Bozdogan, 1987), the Akaike weights (see Wagenmakers and Farrell, 2004), the Bayesian information criterion (BIC; Schwarz, 1978) and a BIC-derived Bayes factor (Wagenmakers, 2007).

R-squared indicates the proportion of variance of the dependent variable, which is explained by the predictor (or predictors) in the model. AIC and BIC are popular methods used to compare the adequacy of multiple statistical models by estimating which model fits better the data, with both measures penalizing for model complexity thus, ceteris paribus, favoring models with fewer parameters. Lower values of AIC and BIC indicate better models. Akaike weights are a simple transformation of the raw AIC values (see Wagenmakers and Farrell, 2004 for procedural details) and capture the model's probability to be the best one in fitting the data, thus providing greater insight into the model selection procedure. In the same vein, the Bayes factor gives a magnitude of the difference between BIC values belonging to different nested and non-nested models, providing a

more reasonable measure of how likely data arise from one model as compared to another one (Wagenmakers et al., 2018).

### 3.1.2 Results

Regression's results are summarized in Table 9.

| Regression Predictor | AIC | BIC | r-squared | Akaike weight | Bayes factor |
|---|---|---|---|---|---|
| Baseline model | 710.7075 | 730.7954 | .5282 | .0001 | - |
| Concreteness | 705.0458 | 730.1557 | .5314 | .0001 | 1.3769 |
| **Imageability** | **676.8584** | **701.9683** | .5431 | **.9686** | **1818550** |
| Five perceptual modalities | 683.7206 | 728.9183 | **.5435** | .0313 | 2.5563 |
| Maximum perceptual strength | 704.2053 | 729.3152 | .5318 | .0001 | 2.0961 |
| Minimum perceptual strength | 707.8135 | 732.9234 | .5303 | .0001 | 0.3451 |
| Mean perceptual strength | 702.0691 | 727.1789 | .5327 | .0001 | 6.0998 |
| Magnitude of perceptual strength | 698.5383 | 723.6482 | .5342 | .0001 | 35.6447 |
| Minkowski 3 distance | 697.9086 | 723.0184 | .5344 | .0001 | 48.8376 |

Table 9. AIC, BIC, r –squared, Akaike weights, and Bayes factor of the regressions run over the log-transformed reaction times in the lexical decision task. In each regression log-transformed frequency and item's length were fixed predictors (baseline) while we systematically changed the predictor of interest.

Results showed imageability to be the best predictor of RTs in lexical decision. This model is 30.9 times more likely to be the best model in terms of Kullback–Leibler discrepancy than the next-best model with the five perceptual modalities as predictors. In other words, the model with imageability is to be preferred over its nearest competitor with a normalized probability of 0.969.

However, when considering r-squared (hence, not accounting for model complexity), the model with five separate perceptual modalities appears to be on par, if not slightly better, than the model with imageability. Following this consideration, we computed an additional model (*optimized perceptual modalities)* by including as predictors only those perceptual modalities actually contributing to the

model fit. In fact, including all five modalities, irrespective of their contribution, enhances model complexity, hence penalises this model in terms of AIC and BIC by increasing the number of its parameters. Following a backward procedure, we excluded the haptic modality from the model and ran the same analysis described in Table 9.

As compared to the five perceptual modalities model, the optimized one had worse values in terms of AIC (684.299) and r-squared (.5425) but, as expected, better values in terms of BIC (724.4748) and Bayes factor, being the data 9.2 more likely to arise from the optimized model than the five perceptual predictors.

However, the direct comparison between imageability (i.e., previous best predictor) and the optimized perceptual modalities model did not lead to significant changes in model comparison: imageability remains the best predictor in terms of AIC and BIC indexes and explained variance.

## 3.2 Word naming task

### 3.2.1 Methods

#### *3.2.1.1 Participants*

28 psychology students from the University of Milano-Bicocca (males = 4; *Age* = 22.8 ± 2; *Education*= 14.18 ± 1.49) took part in the experiment in exchange of course credit. Participants were Italian native speakers and were naïve to the experiment purpose. The study was approved by the local ethical committee, and participants' ethical treatment was in accordance with the principles stated in the Declaration of Helsinki.

#### *3.2.1.2 Materials*

The item set comprised the same set of 1121 words included in the previous experiment (see the section "Materials" of the lexical decision task for more details). The trial-by-trial database is included in our Supplementary materials.

#### *3.2.1.3 Procedure*

Participants took part in a two-session experiment, lasting about half an hour for each session. The two sessions took place at the same time of the day at a maximum temporal distance of two

weeks. After receiving information about the experimental procedure, participants were asked to sign the written informed consent. They were then sat in front of a 17" computer screen. They were informed that they would have been presented a word at the center of the screen and they were instructed to read it aloud as fast as possible.

A practice phase, including 10 words that were not part of the dataset, took place at the beginning of each experimental session. The two experimental sessions were composed of 560 and 561 trials (for a total of 1121 trials for each participant), and each session included a break after the first 280 trials. Each trial started with a fixation cross of 500 ms presented at the center of the screen. Subsequently, an upper-case word was presented, for a maximum duration of 2000 ms (the word disappeared as soon as the software recorded the participant's response), followed by a blank screen with a fixed duration of 1500 ms. The order of the stimuli was randomized across participants. The experimental procedure was implemented in E-Prime 3 (Psychology Software Tools Inc., Pittsburgh. PA). RTs consisted of voice-onset-times automatically recorded by a microphone connected to the response box. Accuracy was manually recorded by the experimenter.

#### *3.2.1.4 Statistical analysis*

For the computations of by-item aggregated RTs, we eliminated incorrect responses (43 data points), RTs inferior to 100 ms (48 data points), and superior to 1700 ms (7 datapoints), and cases with technical failures in recording the response (2072 data points). We followed the same steps of the statistical analysis of Study 2, keeping as fixed predictors word length and frequency and systematically changing the predictor of interest.

### 3.2.2 Results

Regression's results are summarized in Table 10.

| Regression Predictor | AIC | BIC | r-squared | Akaike weight | Bayes factor |
|---|---|---|---|---|---|
| Baseline model | 749.8709 | 769.9588 | .3205 | .0032 | - |
| Concreteness | 748.7109 | 773.8208 | .3224 | .0057 | 0.1450 |
| **Imageability** | **739.5474** | **764.6572** | .3279 | **.5525** | **14.1654** |
| Five perceptual modalities | 740.1303 | 785.3281 | **.3323** | .4128 | 0.0005 |
| Maximum perceptual strength | 749.1961 | 774.3060 | .3221 | .0044 | 0.1138 |
| Minimum perceptual strength | 751.7584 | 776.8683 | .3206 | .0012 | 0.0316 |
| Mean perceptual strength | 749.6016 | 774.7114 | .3219 | .0036 | 0.0929 |
| Magnitude of perceptual strength | 748.2101 | 773.3199 | .3227 | .0073 | 0.1863 |
| Minkowski 3 distance | 747.7249 | 772.8348 | .3230 | .0093 | 0.2374 |

Table 10. AIC, BIC, r-squared, Akaike weights, and Bayes Factor of the regressions run over the log-transformed reaction times in the naming task. In each regression log-transformed frequency and item's length were fixed predictors (baseline) while we systematically changed the predictor of interest.

Results showed imageability to be the best predictor of naming RTs, being the model 1.34 times more likely to be the best model in terms of Kullback–Leibler discrepancy than the next-best model with the five perceptual modalities as predictor. In other words, the model with imageability is to be preferred over its nearest competitor with a probability of 0.572. Nevertheless, the model including the five perceptual modalities is a close second despite its complexity, with an Akaike weight of 0.4128, and indeed it is associated with a higher r-squared score than the model with imageability.

Comparing results on lexical decision vis-à-vis naming, it is evident that the latter has consistently lower scores in terms of explained variance. This is in line with previous results, showing that variance in naming latencies is typically more difficult to model than variance in lexical decision latencies (Brysbaert and New, 2009; Herdağdelen and Marelli, 2017).

In line with the follow-up analysis for the lexical decision task, we ran an optimized perceptual modalities model, eliminating predictors that did not improve model fit in naming RTs. With this procedure, we removed the haptic and gustatory modalities from the model and ran the same analysis described in Table 10. Compared to the five perceptual predictors, the optimized model was better in terms of AIC (739.4651) and BIC (774.619), being the data 211.6 more likely to arise from the optimized model than the five perceptual predictors. The five perceptual modalities model, however, was better in terms of r-squared (.3323 and .3303, respectively).

Concerning the direct comparison between imageability and the optimized modality model, imageability was a better predictor in terms of BIC values, and consequently of Bayes factor, being data 145.6 more likely to arise from the model with imageability as predictor than the optimized model. Even if the optimized perceptual modalities regressor was slightly better in terms of explained variance (.3303 vs .3279), the two models were essentially equivalent AIC-wise (739.47 vs 739.55).

### 3.3 Discussion

In these further analyses, we have evaluated the performance of perceptual-strength measures based on our ratings in predicting RTs in lexical decision and word naming. To the best of our knowledge, only Connell and Lynott (2012) addressed such issue, finding maximum perceptual strength (i.e. a composite variable of the five perceptual modalities) to be stronger than imageability and concreteness in predicting chronometric data in word recognition.

Despite different authors claiming that maximum perceptual strength is the best composite variable to represent the multidimensional perceptual profile (e.g., Connell and Lynott, 2012; Connell and Lynott, 2016; Connell et al., 2018; Winter et al., 2017), not all researchers agreed on this topic. Đurđević et al. (2016), for example, found summed perceptual strength and vector length to be the best composite variable to reduce perceptual strength ratings. Lynott et al. (2019), instead, found Minkowski 3 distance to be the best composite variable to account for the multimodal profile, in line with To et al. (2010), who suggested it to be the best index to represent multisensory dimensions.

In our analyses, in the lexical decision task, imageability was the best predictor for all model fit indexes except the explained variance, which larger portion was explained by the five perceptual ratings separately added as predictors. In the naming task, the picture was more nuanced, with imageability being the best predictor according to BIC (and consequently according to the Bayes Factor), the five perceptual modalities explaining the larger portion of variance and the optimized

model (comprising, in this case, all but haptic perceptual modalities), being on par with imageability in terms of AIC.

The discrepancy between the two statistical procedures (*AIC and BIC vs. r-squared*) is easily explained: AIC and BIC favor more parsimonious model, i.e. the ones with fewer parameters. This result is particularly interesting: on one side it highlights that the five-perceptual-rating option, despite being more penalized by AIC and BIC indexes, is a close second according to the same measures, suggesting that perceptual ratings play a role in predicting human performance in word recognition tasks. On the other side, we did not replicate previous findings on English: indeed, maximum perceptual strength was not the best predictor of participants' behavior and did not hold a dominant position among the computed regression indexes.

What could have led to the difference between Connell and Lynott's (2012) findings and ours? A first possibility is that our results are different because of the item list composition; in fact, in our item set, differently from Connell and Lynott (2012), we included 58 verbs, which are known to be less imageable as compared to nouns (e.g. Bird et al., 2003). However, even when removing verbs from our item set, the analyses revealed a different pattern of results from English ones (see Supplementary materials – Section A).

A second possible explanation may be looked for in the instruction administered for rating collection. Despite our perceptual norms were collected using the same instruction as in Lynott and Connell (2009, 2013), concreteness and imageability ratings came from the Italian ANEW (Montefinese et al., 2014) in our case, and from MRC dataset (Coltheart, 1981) for the English experiments. Concreteness instructions, indeed, were quite different in the two datasets, with Montefinese et al. suggesting the idea of concreteness as experienced through the five senses, asking participants "*to assess the extent to which a word denotes something that can be directly perceived by the senses"*, while MRC authors linking concreteness to "objects, materials or persons": "*any word that refers to objects, materials or persons should receive a high concreteness rating; any word that refers to an abstract concept that cannot be experienced by the senses should receive a high abstractness rating".* Imageability, however, was very similarly defined in the two datasets, not justifying the inconsistency between the two studies. In both cases, indeed, participants were asked to evaluate imageability based on the easiness to access a "mental image": "*For the imageability scale, we ask you to evaluate how easily you can bring to mind a mental image (e.g., a mental picture, a sound, or other sensorial experience) of a given word when it is presented"* (Montefinese et al., 2014) vs *"any word which, in your estimation, arouses a mental image (i.e., a mental picture, or sound, or other sensory experience) very quickly and easily"* (Coltheart, 1981; Paivio et al., 1968). A

third possible explanation concerns the statistical approach adopted. Connell and Lynott (2012) determined the best fitting model using an index that measures the explained variance of each alternative, namely r squared. We opted to compute different indexes to estimate the best fitting model. Interestingly, Akaike weights suggested the best model to be the one with imageability as predictor, which in lexical decision task has a similar r-squared compared to the model with the five perceptual modalities. In the naming task, Akaike weights indicated imageability as best predictor, while according to r-squared the five perceptual modalities led to a higher explained variance. However, notwithstanding these differences, in our analyses on Italian data the maximum perceptual strength (i.e., the best predictor for English) did not hold a dominant position among the computed regression indexes, suggesting that the cross-linguistic difference we observed does not depend on the adopted evaluation methods.

The last hypothesis is that, indeed, the empirical difference we observe depends on the different language examined: there might be features, differentiating English from Italian, that lead perceptual strength to better capture word processing in the former as compared to the latter, in which imageability seems to provide the best predictions. However, before delving more into this last hypothesis, we need to check that the observed dissociation cannot be trivially explained by the different datasets considered in the Italian vis-à-vis English analysis. In order to address this possibility, we exploited the ANEW translations in order to extract a dataset of RTs for the English words corresponding to our items from available resources (English Lexicon Project; Balota et al., 2007). If the English stimuli led to the same pattern observed for Italian (i.e., better performance of imageability vis-à-vis perceptual strength), the present results would depend on the considered item set. If English data showed a pattern consistent with Connell & Lynott (2012) (i.e., better performance of perceptual strength vis-à-vis imageability), then we would have support for a genuine cross-linguistic difference.

## 4. Part 3: cross-linguistic comparisons

### 4.1 Materials

Relying on existing resources, we retrieved for the English language the same (dependent and independent) variables considered in our analyses of Italian data.

Behavioral data for English stimuli were taken from the English Lexicon Project (ELG) (https://elexicon.wustl.edu/), which includes accuracy and reaction times for both lexical decision and naming. ELP lexical decision RTs were collected from 815 American participants, native

speakers of English, each one presented with 1700 words (and 1700 non-words), with a total of 34 participants per item. ELP naming RTs, instead, were obtained from 444 American speakers in total, each one presented with 2500 items (25 participants per item). From ELP we also collected word length and log-transformed frequency (Brysbaert & New, 2009; Keuleers, Diependaele, & Brysbaert, 2010).

English perceptual norms were obtained from the Lancaster scale (Lynott et al., 2019), which comprises 39707 items rated over six sensorimotor dimensions (visual, haptic, auditory, olfactory, gustatory and interoceptive) and five action effectors (mouth/throat, hand/arm, foot/leg, head excluding mouth/throat, and torso). Sensorimotor dimensions were rated by 2,625 participants, each one completing a mean of 5.99 lists comprising 58 items.

Imageability and concreteness ratings were taken from the MRC machine-usable dictionary (available online at http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm), which comprises 150,837 words and 26 linguistic and psycholinguistics variables (Wilson, 1988).

The three English resources (ELP, Lancaster scale, MRC) were combined to create a *parallel dataset*, comprising translations of our Italian items along with the corresponding normative and behavioral data from English. To create the dataset, we proceeded as follows (i) we selected from the Lancaster scale the items overlapping with the Italian version of ANEW (Montefinese et al., 2014), exploiting the translations provided in the resource; (ii) we compared the remaining 1090 items with the MRC database; (iii) the remaining 658 overlapping items were then compared with words contained in the ELP; this final step did not produce item list reduction.

### 4.2 Statistical approach

We run the same analyses described for the lexical decision and naming tasks, running models with different predictors and considering several indexes of model fit. The first analyses were computed on our lexical decision and naming RTs on the reduced 658-word item set. In this way, we aimed at testing whether the reported results are robust across the two item sets (the original 1121 item set vs. the reduced 658-word item set). Then, in order to test the alleged cross-linguistic dissociation, we ran the same analysis on the parallel dataset we obtained for English.

### 4.3 Results

In Tables 11 and 12, we summarized the results obtained for the Italian dataset over lexical decision and naming tasks, respectively. Analyses were performed to disentangle whether discrepancies emerged between Italian and English were due to differences in item list composition, or they arose from a pure cross-linguistic effect.

| Regression Predictor | AIC | BIC | r-squared | Akaike weight | Bayes Factor |
|---|---|---|---|---|---|
| Baseline model | 249.236 | 267.1928 | .5021 | .0001 | - |
| Concreteness | 242.6777 | 265.1237 | .5085 | .0001 | 2.8138 |
| **Imageability** | **224.0212** | **246.4672** | .5222 | **.5664** | **31659.7** |
| Five perceptual modalities | 224.6475 | 265.0504 | **.5276** | .4140 | 2.9189 |
| Maximum perceptual strength | 241.6713 | 264.1173 | .5093 | .0001 | 4.6541 |
| Minimum perceptual strength | 245.1118 | 267.5578 | .5067 | .0001 | 0.8332 |
| Mean perceptual strength | 236.7528 | 259.1988 | .5129 | .0010 | 54.4346 |
| Magnitude of perceptual strength | 232.1047 | 254.5507 | .5163 | .0099 | 556.1567 |
| Minkowski 3 distance | 232.4043 | 254.8503 | .5161 | .0086 | 478.7842 |

Table 11. AIC, BIC, r-squared, Akaike weights, and Bayes factor of the regressions run over the log-transformed reaction times in the Italian reduced lexical decision dataset. In each regression log-transformed frequency and item's length were included as baseline predictors while we systematically changed the predictor of interest.

| Regression Predictor | AIC | BIC | r-squared | Akaike weight | Bayes factor |
|---|---|---|---|---|---|
| Baseline model | 388.3638 | 406.3206 | .2747 | .0026 | - |
| Concreteness | 383.978 | 406.4242 | .2817 | .0233 | 0.9495 |
| Imageability | **376.8102** | **399.2563** | .2895 | **.8398** | **34.1974** |
| Five perceptual modalities | 381.4267 | 421.8295 | **.2932** | .0835 | 0.0004 |
| Maximum perceptual strength | 386.2129 | 408.6590 | .2793 | .0076 | 0.3106 |
| Minimum perceptual strength | 390.237 | 412.6830 | .2793 | .0010 | 0.0415 |
| Mean perceptual strength | 386.827 | 409.2730 | .2749 | .0056 | 0.2285 |
| Magnitude of perceptual strength | 384.6209 | 407.0669 | .2810 | .0169 | 0.6886 |
| Minkowski 3 distance | 384.3202 | 406.7662 | .2814 | .0197 | 0.8003 |

Table 12. AIC, BIC, r-squared, Akaike weights, and Bayes factor of the regressions run over the log-transformed reaction times in the Italian reduced naming dataset. In each regression log-transformed frequency and item's length were included as baseline predictors while we systematically changed the predictor of interest.

Results for the Italian reduced dataset are in line with the one observed in the full item set: imageability proved to be the best-performing predictor according to the AIC- and BIC-related measures, although the five perceptual modalities, separately introduced in the model, obtained slightly higher explained variance.

Table 13 and 14 reports the results of the same analyses applied to the parallel English dataset.

| Regression Predictor | AIC | BIC | r-squared | Akaike weight | Bayes factor |
|---|---|---|---|---|---|
| Baseline model | -521.7245 | 636.4904 | .7765 | .0047 | - |
| Concreteness | -522.1098 | 640.5943 | .7773 | .0057 | 0.1285 |
| Imageability | -526.7784 | 635.9257 | .7789 | .05866 | 1.3262 |
| Five perceptual modalities | -529.5144 | 651.1465 | **.7824** | .2304 | 0.0007 |
| Maximum perceptual strength | -524.661 | 638.0431 | .7781 | .0203 | 0.4601 |
| Minimum perceptual strength | **-530.9022** | **631.8019** | .7802 | **.4611** | **10.4255** |
| Mean perceptual strength | -528.7148 | 633.9892 | .7795 | .1545 | 3.4924 |
| Magnitude of perceptual strength | -526.0770 | 636.6271 | .7786 | .04131 | 0.9339 |
| Minkowski 3 distance | -524.9412 | 637.7628 | .7782 | .0234 | 0.5293 |

Table 13. AIC, BIC, r-squared, Akaike weights, and Bayes factor of the regressions run over the log-transformed reaction times in the parallel dataset lexical decision task. In each regression log-transformed frequency and item's length were predictors (baseline) while we systematically changed the predictor of interest.

The best-performing model was the one including the minimum perceptual strength as predictor, according to the AIC, BIC, and Akaike weights. The five perceptual modalities, however, obtained an overall larger explained variance.

| Regression Predictor | AIC | BIC | r-squared | Akaike weight | Bayes factor |
|---|---|---|---|---|---|
| Baseline model | -280.4730 | 877.7418 | .6426 | 0.0032 | - |
| Concreteness | -278.8347 | 883.8694 | .6428 | 0.0014 | 0.0467 |
| Imageability | -280.3214 | 882.3827 | .6436 | 0.0030 | 0.0982 |
| Five perceptual modalities | -284.5791 | 896.0817 | **.6502** | 0.0246 | 0.0001 |
| Maximum perceptual strength | -289.8085 | 872.8956 | .6487 | 0.3358 | 11.2808 |
| Minimum perceptual strength | -281.2592 | 881.4449 | .6441 | 0.0047 | 0.1570 |
| Mean perceptual strength | -282.9499 | 879.7542 | .6451 | 0.0109 | 0.3656 |
| Magnitude of perceptual strength | -278.5393 | 884.1648 | .6427 | 0.0012 | 0.0403 |
| **Minkowski 3 distance** | **-291.0199** | **871.6841** | .6494 | **0.6154** | **20.6734** |

Table 14. AIC, BIC, r-squared, Akaike weights, and Bayes factor of the regressions run over the log-transformed reaction times in the parallel naming task. In each regression log-transformed frequency and item's length were fixed predictors (baseline) while we systematically changed the predictor of interest.

Considering the naming RTs, the best-performing model was the one including the Minkowski distance according to all but r-squared model fit-indexes. The model with the five perceptual modalities included as separate predictors was associated to the highest explained variance.

### 4.4 Discussion

The present section aimed at disentangling whether differences between our results and previous findings in English could be due to differences in item selection, rather than a genuine cross-linguistic effect. To investigate this aspect, we compared Italian and English datasets comprising the same items (following Montefinese et al., 2014 translation) and ran the same statistical analyses on both sets. The rationale behind this procedure was that if differences between Italian and English were due to item selection, comparing the datasets including the same (translated) words, would cause the dissociation to disappear. Conversely, if we were observing a genuine cross-linguistic effect, then

such difference should be confirmed through this procedure. Given these premises, we found that results on the reduced Italian dataset were consistent with findings over the complete database. More importantly, we found that differences between Italian and English were robust when considering the same items, thus suggesting the presence of a genuine cross-linguistic effect. However, it must be noted that our analysis on the English items only partially replicated the pattern found by Lynott and colleagues (2019). Indeed, in line with this study, we found Minkowski distance composite variable being the strongest predictor of naming RTs, followed by the maximum perceptual strength, which was found to be the best predictor in Connell and Lynott (2012) and confirmed as a good predictor in Lynott et al. (2019). Considering lexical decision RTs, the pattern observed in our English dataset suggested that minimum perceptual strength, namely the minimum score across the five perceptual modalities, was the best composite variables of perceptual modalities in predicting behavioural performance. This result is partially in line with previous studies, showing that a measure of perceptual strength was stronger in predicting behavioural performance with English items as compared to both concreteness and imageability. However, the specific characterization of perceptual strength (i.e., the minimum score in the modality norms) was never reported as the best predictor in previous norms. With the present data, we cannot evaluate whether this difference is theoretically relevant or simply depends on small variations between different characterizations of the same latent variable. We leave this question to future studies.

How to explain the observed cross-linguistic dissociation? At present, we can only speculate on the reasons leading to such a difference. The source for this effect might be found in the norms themselves, insofar it is known that in semantic rating studies with lexical materials, participants do not only simply evaluate the object denoted by the word; rather, they are also influenced, in their judgments, by statistical distributions in the language. For example, semantic transparency ratings are impacted by the frequency of constituent morphemes of the presented complex word (Bell & Schäfer, 2016). This kind of influence was also described for variables under investigation in the present paper, such as concreteness (Hollis & Westbury, 2016) or perceptual modalities (Louwerse & Connell, 2011). In these works, this piece of evidence was interpreted in terms of language being able to encode grounded information. However, these results also indicate that, when producing semantic ratings, participants might be influenced by aspects of the presented word purely associated with nuanced patterns in its lexical distribution. If that's the case, one can argue that different languages, being associated with different linguistic distributions, will be associated with slightly deviate semantic norms, as produced by their speakers. In other words, it is conceivable that speakers of different languages will provide slightly different rating scores to the same (translated) items because of their different language experiences. These differences could explain the cross-linguistic

dissociation described here: intuitions of Italian speakers, during the rating task, might have produced imageability scores that are more apt at capturing chronometric data than their English counterparts.

An alternative explanation could refer to non-arbitrariness in natural language, and more specifically to *iconicity*, which reflects the resemblance between word form aspects and its meaning (for a review see Dingemanse et al., 2015). Recently iconicity received special attention from both a psycholinguistic and cognitive perspective. Across languages, iconic words are commonly used to drive perceptuomotor analogies between word form and word meaning, such as referent color and shape, size, temperature (Dingemanse et al., 2012). Despite Indo-European vocabularies were considered to be highly arbitrary (e.g. Perniss et al., 2010; Vigliocco et al., 2014) as compared to some African, Asian and America lexicon, converging evidence suggested the emergence of interesting patterns even in these languages. For example, iconicity ratings have been shown to correlate with sensory experience and semantic neighborhood, imageability, frequency, and age of acquisition in English (Juhasz and Yap, 2013; Sidhu & Pexman, 2018; Winter et al., 2017; Perry et al., 2015). Focusing on cross-linguistic differences, this variable has been shown to vary across different languages (English vs Spanish, Perry et al., 2015) and language modality (Perlman et al., 2018), shaped by the cultural evolutionary processes to favour learning, discriminability across categories and communication (Imai and Kita, 2014; Lupyan and Casasanto, 2014; Digenmanse et al., 2015).

Adopting a similar perspective, it is possible that differences in iconicity between English and Italian could explain the cross-linguistic dissociation reported in the present study. That is, Italian words may be overall more iconic, and hence Italian imageability scores may drive some unique information, only partially overlapping perceptual ones. This idea is partially supported by our data: indeed, in line with Connell and Lynott (2012) adding perceptual modality measures in a model already containing imageability increased the model fit, but in Italian the opposite was also true, with the imageability scores significantly improving the regression already containing the five perceptual modalities, thus suggesting that both imageability and perceptual ratings are adding some unique components to the models (see Supplementary materials - section B for the analysis and a more detailed discussion); moreover, as shown in Table 7, imageability scores were predicted not only by visual and olfactory perceptual ratings (as it was in English) but also by haptic perceptual values. Further investigation will be needed in order to shed light on the observed cross-linguistic dissociation. However, irrespective of how the effect should be explained, the present results have important methodological implications, stressing the importance of having modality norms and, more in general, semantic norms in different languages: indeed, we can't take for granted that results on

English are generalizable to all the other languages. In this sense, a standardized dataset like ANEW, which is translated in different languages, can be a useful and powerful instrument to make cross-linguistic comparisons in a more controlled way.

## 5. Conclusions

In the present paper, we present and release perceptual ratings for 1121 Italian words. The richness and usefulness of the present resource is that it ideally complements the largest norming work currently available in Italian (Montefinese et al., 2014). Moreover, given the fact that the dataset comprises items from the English ANEW, which has adaptations in different language (e.g. Spanish, Redondo et al., 2007; Portuguese, Soares et al., 2012), our resource is particularly important from a cross-linguistic comparison perspective, as clearly exemplified by the third experiment described in the present work.

The strength of the present resource, however, is not limited to the perceptual rating validation. Indeed, we released a set of trial-by-trial chronometric data collected with a word naming and a lexical decision task, which does not have any comparable example in Italian. The two novel experiments were collected to investigate to what extent perceptual strength compares with concreteness and imageability in predicting human behavior, and specifically reaction times in word processing tasks. Our results provide evidence for cross-linguistic differences in the impact of these rating-based measures, suggesting caution in generalizing results obtained on English studies to other languages.

Beyond the ability in terms of predicting RTs, the present rating norms are crucial to study how the conceptual system is organized and how different words can imply different semantic representations due to the "channels" available for the acquisition of the information they refer to.

### Acknowledgments

This work was supported by a Fondazione Cariplo-Regione Lombardia (grant number 2017-1633) grant to Marco Marelli. We thank Simona Marnini and Beatrice Capoferri for their help in data collection.

### Open Practices Statement

The database of the present study is publicly archived at https://osf.io/zdg59/. The experiment was not preregistered.

**References**

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika, 60(2), 255-265.

Baayen, R. H. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics using R. Cambridge University Press.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. Journal of Experimental Psychology: General, 133(2), 283.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. Behavior Research Methods, 39, 445–459.

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.

Beavers, J., Levin, B., & Tham, S. W. (2010). The typology of motion expressions revisited. Journal of linguistics, 46(2), 331-377.

Bell, M. J., & Schäfer, M. (2016). Modelling semantic transparency. Morphology, 26(2), 157-199.

Bird, H., Howard, D., & Franklin, S. (2003). Verbs and nouns: The importance of being imageable. Journal of Neurolinguistics, 16(2-3), 113-149.

Bonin, P., Méot. A., Ferrand. L., & Bugaïska. A. (2015). Sensory experience ratings (SERs) for 1.659 French words: Relationships with other psycholinguistic variables and visual word recognition. Behavior research methods. 47(3). 813-825.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52(3), 345-370.

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings (Vol. 30, No. 1, pp. 25-36). Technical report C-1, the center for research in psychophysiology, University of Florida.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior research methods, 41(4), 977-990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior research methods, 46(3), 904-911.

Chen, I. H., Zhao, Q., Long, Y., Lu, Q., & Huang, C. R. (2019). Mandarin Chinese modality exclusivity norms. PloS one, 14(2), e0211336.

Coltheart, M. (1981). The MRC psycholinguistic database. The Quarterly Journal of Experimental Psychology Section A, 33(4), 497-505.

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. Cognition, 125(3), 452-465.

Connell, L., & Lynott, D. (2015). Embodied semantic effects in visual word recognition. Foundations of embodied cognition, 2, 71-89.

Connell, L., & Lynott, D. (2016). Do we know what we're simulating? Information loss on transferring unconscious perceptual simulation to conscious imagery. Journal of Experimental Psychology: Learning, Memory, and Cognition, 42(8), 1218.

Connell, L., Lynott, D., & Banks, B. (2018). Interoception: the forgotten modality in perceptual grounding of abstract and concrete concepts. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1752), 20170143.

Díez-Álamo, A. M., Díez. E., Wojcik. D. Z., Alonso. M. A., & Fernandez. A. (2018). Sensory experience ratings for 5.500 Spanish words. Behavior Research Methods. 1-11.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. Language and Linguistics compass, 6(10), 654-672.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. Trends in cognitive sciences, 19(10), 603-615.

Đurđević, D. F., Popović Stijačić, M., & Karapandžić, J. (2016) A quest for sources of perceptual richness: Several candidates. In S. Halupka-Rešetar and S. MartínezFerreiro (Eds.) Studies in Language and Mind (pp. 187-238). RS, Novi Sad: Filozofski fakultet u Novom Sadu.

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. Cognitive science, 41(4), 976-995.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. Psychonomic bulletin & review, 23(6), 1744-1756.

Juhasz, B. J., & Yap, M. J. (2013). Sensory experience ratings for over 5,000 mono-and disyllabic words. Behavior Research Methods, 45(1), 160-168.

Juhasz, B. J., Yap, M. J., Dicke, J., Taylor, S. C., & Gullick, M. M. (2011). Tangible words are recognized faster: The grounding of meaning in sensory and perceptual systems. The Quarterly Journal of Experimental Psychology, 64(9), 1683-1691.

Keuleers. E., & Brysbaert. M. (2010). Wuggy: A multilingual pseudoword generator. Behavior Research Methods 42(3). 627-633.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. Frontiers in psychology, 1, 174.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. Behavior research methods, 44(1), 287-304.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. Journal of Experimental Psychology: General, 140(1), 14.

Kuperman, V. (2013). Accentuate the positive: Semantic access in English compounds. Frontiers in Psychology, 4, 203.

Louwerse, M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. Cognitive science, 35(2), 381-398.

Lupyan, G., & Casasanto, D. (2015). Meaningless words promote meaningful categorization. Language and Cognition, 7(2), 167-193.

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. Behavior Research Methods, 41(2), 558-564.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. Behavior research methods, 45(2), 516-526.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). Lancaster Sensorimotor Norms-Pre-print.

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. Cortex, 48(7), 788-804.

Miklashevsky, A. (2018). Perceptual Experience Norms for 506 Russian Nouns: Modality Rating. Spatial Localization. Manipulability. Imageability and Other Variables. Journal of psycholinguistic research. 47(3). 641-661.

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013). Semantic memory: A feature-based analysis and new norms for Italian. Behavior research methods, 45(2), 440-461.

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.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. Canadian Journal of Psychology/Revue canadienne de psychologie, 45(3), 255.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. Journal of experimental psychology, 76(1p2), 1.

Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2003). Verifying different-modality properties for concepts produces switching costs. Psychological science, 14(2), 119-124.

Perniss, P., Thompson, R., & Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. Frontiers in psychology, 1, 227.

Perlman, M., Little, H., Thompson, B., & Thompson, R. L. (2018). Iconicity in Signed and Spoken Vocabulary: A Comparison Between American Sign Language, British Sign Language, English, and Spanish. Frontiers in psychology, 9, 1433.

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. PloS one, 10(9), e0137147.

Pulvermüller, F. (2018). Neural reuse of action perception circuits for language, concepts and communication. Progress in neurobiology, 160, 1-44.

R Core Team. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of ANEW (affective norms for English words). Behavior research methods, 39(3), 600-605.

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 6(2), 461-464.

Sidhu, D. M., & Pexman, P. M. (2018). Lonely sensational icons: semantic neighbourhood density, sensory experience and iconicity. Language, Cognition and Neuroscience, 33(1), 25-31.

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., & Frade, C. S. (2012). The adaptation of the Affective Norms for English words (ANEW) for European Portuguese. Behavior research methods, 44(1), 256-269.

Speed, L. J., & Majid. A. (2017). Dutch modality exclusivity norms: Simulating perceptual modality in space. Behavior research methods. 49(6). 2204-2218.

To, M. P. S., Baddeley, R. J., Troscianko, T., & Tolhurst, D. J. (2010). A general rule for sensory cue summation: evidence from photographic, musical, phonetic and cross-modal stimuli. Proceedings of the Royal Society B: Biological Sciences, 278(1710), 1365-1372.

Van Dantzig, S., Cowell, R. A., Zeelenberg, R., & Pecher, D. (2011). A sharp image or a sharp knife: Norms for the modality-exclusivity of 774 concept-property items. Behavior Research Methods, 43(1), 145-154.

Van Dantzig, S., Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2008). Perceptual processing affects conceptual processing. Cognitive science, 32(3), 579-590.

Vigliocco G., Perniss P., Vinson D. (2014) Language as a multimodal phenomenon: implications for language learning, processing and evolution. Philosophical Transactions Royal Society B 369: 20130292. http://dx.doi.org/10.1098/rstb.2013.0292

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. Psychonomic bulletin & review, 11(1), 192-196.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. Psychonomic bulletin & review, 14(5), 779-804.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychonomic bulletin & review, 25(1), 35-57.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. Behavior research methods, instruments, & computers, 20(1), 6-10.

Winter, B. (2016). Taste and smell words form an affectively loaded and emotionally flexible part of the English lexicon. Language, Cognition and Neuroscience, 31(8), 975-988.

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic?. Interaction Studies, 18(3), 443-464.

Zdrazilova, L., & Pexman, P. M. (2013). Grasping the invisible: Semantic processing of abstract words. Psychonomic bulletin & review, 20(6), 1312-1318.