Behavior Research Methods (2024) 56:126–147
https://doi.org/10.3758/s13428-022-02031-y

The Flickr frequency norms: What 17 years of images tagged online tell us about lexical processing

Marco A. Petilli[1] · Fritz Günther[2] · Marco Marelli[1,3]

Accepted: 19 November 2022 / Published online: 12 December 2022

Abstract
Word frequency is one of the best predictors of language processing. Typically, word frequency norms are entirely based on natural-language text data, thus representing what the literature typically refers to as purely *linguistic* experience. This study presents Flickr frequency norms as a novel word frequency measure from a domain-specific corpus inherently tied to extra-linguistic information: words used as image tags on social media. To obtain Flickr frequency measures, we exploited the photo-sharing platform Flickr Image (containing billions of photos) and extracted the number of uploaded images tagged with each of the words considered in the lexicon. Here, we systematically examine the peculiarities of Flickr frequency norms and show that Flickr frequency is a hybrid metrics, lying at the intersection between language and visual experience and with specific biases induced by being based on image-focused social media. Moreover, regression analyses indicate that Flickr frequency captures additional information beyond what is already encoded in existing norms of linguistic, sensorimotor, and affective experience. Therefore, these new norms capture aspects of language usage that are missing from traditional frequency measures: a portion of language usage capturing the interplay between language and vision, which – this study demonstrates – has its own impact on word processing. The Flickr frequency norms are openly available on the Open Science Framework (https://osf.io/2zfs3/).

Keywords Flickr frequency · Word frequency · Visual strength · Concreteness · Imageability · Flickr

Introduction

Word frequencies in a visual context: From general-domain to specific-domain corpora

Word frequency is one of the best predictors of language processing and explains large portions of variance of the participants' responses in many language processing tasks (e.g., Brysbaert et al., 2011, 2012, 2018; Herdağdelen & Marelli, 2017; van Heuven et al., 2014). The common practice in collecting word frequencies is to count word occurrences in extensive text collections (i.e., corpora). So far, source data were mainly selected among samples of written and spoken language, expected to approximate as much as possible the general linguistic experience of speakers, including newspapers, textbooks, novels, and magazines (Baayen et al., 1996; Kucera & Francis, 1967), television subtitles (Brysbaert & New, 2009; van Heuven et al., 2014), and social media (Herdağdelen & Marelli, 2017). Interestingly, these different data sources appear to complement each other in a non-trivial manner: while corpora based on social media (Herdağdelen & Marelli, 2017) better explain variance in participants' responses, the other corpora still provide a significant unique contribution, thus suggesting that each corpus ends up representing a different portion of the linguistic experience, only partially overlapping with what is captured by other corpora.

However, despite the differences between the sources from which these frequencies were collected, they are still entirely based on natural-language text data, thus representing what the literature typically considers as purely *linguistic* experience (Glenberg & Robertson, 2000; Günther et al., 2019; Sahlgren, 2006). Here we systematically examine the peculiarities of word frequency norms from a domain-specific corpus inherently tied to extra-linguistic information: words used as image tags, and thus as labels or descriptions for synchronously visually available referents or situations. Thus, we focus on a qualitatively different language sample

✉ Marco A. Petilli
marco.petilli@unimib.it

[1] University of Milano–Bicocca, Milan, Italy

[2] University of Tübingen, Tübingen, Germany

[3] NeuroMI, Milan Center for Neuroscience, Milan, Italy

from standard textual data (in that it is directly connected to visual information), expected to capture the distribution of word forms as conditioned by the visual environment.

The choice of the visual domain is not casual. First of all, vision, among the various perceptual modalities, is the most relevant for us: almost 74% of English words (analysis based on the item set of 40,000 words included in the Lancaster Sensorimotor Norms and chosen to represent a complete adult vocabulary; Lynott et al., 2020) are visually dominant (i.e., vision is the sensory dimension through which the referred concept is experienced most strongly), thus indicating that the majority of word meanings are grounded in visual experience (Lynott et al., 2020; Winter et al., 2018). Moreover, although vision and language are frequently treated as distinct domains in the literature, they are entangled in our experience in such a way that one ends up influencing the other (Cohn & Schilperoord, 2022). On the one hand, there are several studies which highlight a clear impact of visual experience on language: asymmetries in our perceptual experiences are reflected in our vocabulary (e.g., Winter et al., 2018); and even in purely linguistic contexts, the visual properties of objects affect conceptual processing (e.g., Günther et al., 2020a; Petilli et al., 2021; Zwaan et al., 2002). Likewise, measures pertaining to the visual experience with a word referent – such as ratings of concreteness (i.e., how concrete vs abstract XXX is), imageability (i.e., how easy it is to form an image of XXX) or visual strength (i.e., to what extent do you experience XXX by seeing) – proved to be important predictors of word processing speed (Binder et al., 2005; Bleasdale, 1987; Connell & Lynott, 2012, 2014; De Groot, 1989; Lynott et al., 2020; Vergallito et al., 2020). On the other hand, other studies show that language experience affects how we perceive the world. Verbal labels do not simply refer to object representations but rather actively modulate them, affecting how we organize and process corresponding visual representations: Lupyan (Lupyan, 2008, 2012a, 2012b; Lupyan et al., 2020) suggests that, in visual experience, the two factors (i.e., linguistic and perceptual) cannot be really disentangled. Thus, the combination between lexical information and perceptual experience as well as their mutual interaction represent a psychologically plausible mechanism of how we can acquire and organize conceptual representations (see for example, Günther et al., 2020b), supporting the psychological validity of employing a source corpus inherently linking linguistic and visual experience.

What is Flickr?

As this corpus, we here employ image tags from the Flickr Image online photo-sharing platform (www.flickr.com), one of the Internet's largest repositories of images, where amateur and professional photographers can share their photographs with the online community. Flickr is one of the first classic 'web 2.0 sites' (Cox, 2008), a term used to refer to the second generation of websites that started around 2004 and is characterized by the growth of social media and the change in the role of users from passive consumers of information to active creators and sharers of online content. Flickr has a relatively long history as a social media. It was created in 2004 (at the same time that digital cameras started outselling analogue cameras; Weinberger, 2007), and its popularity as a photo-sharing platform has grown very rapidly, reaching its peak in 2013–2015 with around 3.5 million new images uploaded daily and more than 112 million members (https://blog.flickr.net/en/2015/06/10/thank-you-flickr-community/). Today, Flickr counts tens of billions of tagged photos uploaded from 63 different countries (http://expandedramblings.com/index.php/flickr-stats/). Flickr users comprise all types of photographers, including casual hobbyists (often taking mundane photos to share for general social interaction), serious hobbyists (those often taking photos to share with hobby contacts and place more attention on the quality of the photos), semi-professional and professional photographers (expert using photography as part of their job) (Cox, 2008; Stuart, 2019).

Tagging (i.e., assigning textual labels to web objects, i.e., images or videos in Flickr) is the key feature of Flickr: while Flickr was not the first photo-sharing platform, it was one of the first websites to adopt tagging in order to emphasize sharing (Smith, 2007). On Flickr, one or multiple tags can be assigned to images (with a maximum of 75 tags, although most pictures have between 1 and 15 tags; see Bolognesi, 2016b; https://www.flickr.com/photos/mariannabolognesi/7073104431), either by the uploader or other Flickr users (if the uploader allows it), enabling categorizing and retrieving images that match those tags. If a person performs a global search within Flickr and looks for images with the tag "bus", all the images that users tagged with "bus" will be displayed (see Fig. 1 for an example of image search results). When looking at users' motivations for tagging on Flickr, Ames and Naaman (2007) found that the main motivating factors are s*ocial communication* (i.e., adding tags to draw attention to images) and *social organization* (i.e., adding tags to allow other people can search for and retrieve images). Another motivation for tagging is *self-organization* (i.e., adding tags to organize photographs in personal photo collections and easily retrieving them) (Stuart, 2012, 2019). Other studies focused on how frequently various tag categories are used on Flickr (Beaudoin, 2007; Bolognesi, 2016b). These studies organized Flickr tags in 18 categories. They showed that locations (e.g., *California, Amsterdam, beach, field*), participants (e.g., *baby, woman, Elvis*), and associated entities that can be found in the same picture (e.g.,

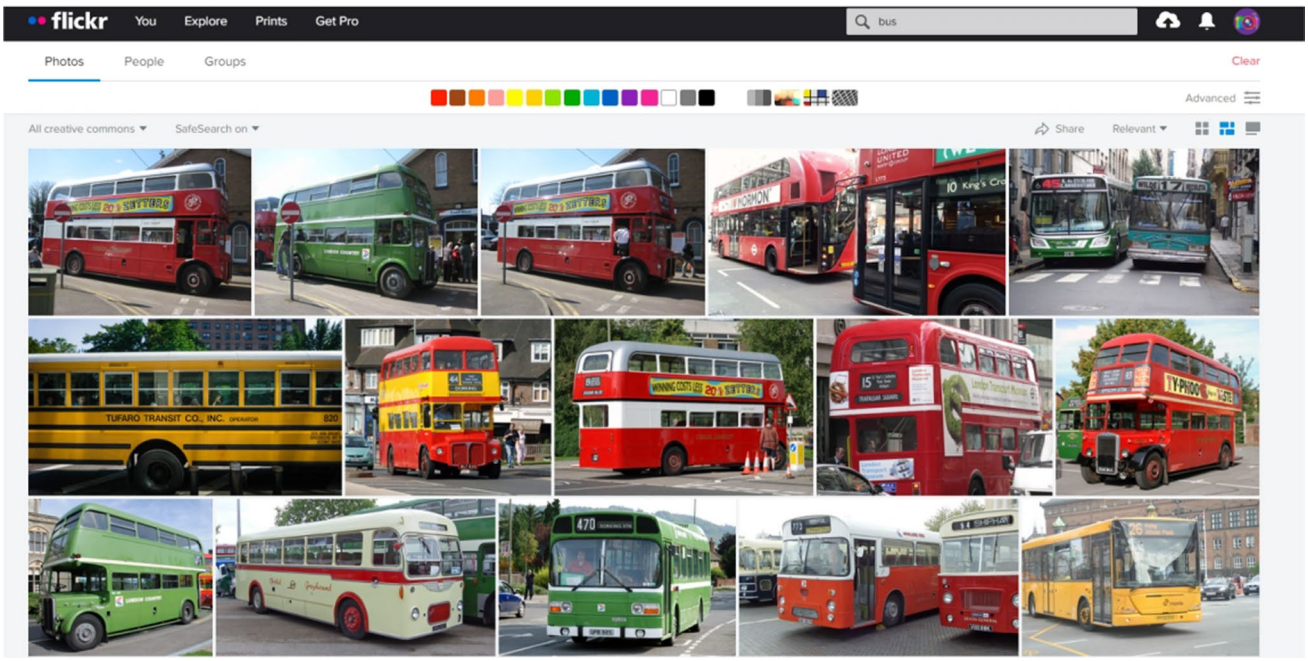

Fig. 1 Example of images results returned by searching for the tag "bus" on www.flickr.com

house, *car*, *rock*, *water*) are the most frequent categories represented in Flickr tags (see also Table 1 for examples of words at different frequency levels extracted from Flickr). In another study focused on how tagging behavior relates to image content, Stuart (2012) examined tags assigned to images on Flickr and classified them according to their relationship with their accompanying image. It was found that the majority of tags "generically identify what an image is of" (e.g., an image of a dog, tagged as "dog" or "animal"). Common tags are also those labels having a specific relationship with the image content, such as tags of place names or events (e.g., an image of the Colosseum in Rome tagged as "Colosseum" or "Rome") or adjectives and descriptive terms identifying what the image is about (e.g., a photograph of people smiling tagged as "happiness"). Also, other types of tags such as self-reference tags (e.g., "my dog") or tags relating to techniques/methods (i.e., photography technique, camera model, film used etc.) are common on Flickr. Tags for which was not possible to determine any relationship with the accompanying image are very infrequent (i.e., less than 1%).

Notably, the use of Flickr in our field of research is not new. For example, Bolognesi (2014), by using a distributional approach, built up a semantic space based on Flickr tag co-occurrences (i.e., Flickr Distributional Space, see also Bolognesi, 2016a, 2016b) and showed that an inherently visual domain – such as the domain of color – is better mapped by the Flickr Distributional Space rather than distributional spaces based on solely textual information (Baroni & Lenci, 2010; Landauer & Dumais, 1997). Moreover, Flickr is a gold mine in computer vision research: with its huge numbers of tagged images, it serves as an excellent resource for training convolutional neural networks (e.g., T. Chen et al., 2014; X. Chen & Gupta, 2015; Das & Clark, 2018). Furthermore,

Table 1 Examples of words at the seven levels of the Zipf scale, based on the Flickr-UK and Flickr-US word frequency

Zipf value	Examples Flickr US	Examples Flickr UK
1	Capitulation, agitate, comorbidity, loudly, complication	Dosage, burgle, noisily, obstructive, quarterly
2	Appreciating, accompaniment, adored, causality, relevance	Inflammable, seedcase, tedious, palatinate, irrelevant
3	Mates, snorkeler, botanist, correct, uncertain	Astrology, skincare, macula, horoscope, dictionary
4	Bucolic, accessories, arsenal, heater, cirque	Cartoon, catenary, milestone, teamwork, teacher
5	Backyard, funeral, peninsula, dessert, cop, shirt	Baptist, cigarette, fighting, galaxy, journal
6	Dinner, windows, model, navy, ship	Aeroplane, butterfly, car, Christmas, school
7	California, Canada, sports, Florida, Arizona	Bus, church, Dublin, railway, Wales

Flickr images constitute a quite large part of ImageNet (Deng et al., 2009), a large-scale database of labeled images adopting the hierarchical category structure of WordNet (Miller, 1998) and designed for use in visual object recognition research (e.g., Krizhevsky et al., 2012), but also adopted to build up prototypical vision-based representations for concepts to be used in psychological research (Anderson et al., 2015; Günther et al., 2022; Petilli et al., 2021).

Peculiarities of Flickr frequency norms

As for any frequency measure, the measure distribution depends substantially on the data source we are considering (Baayen et al., 2016). Corpora based on subtitles provide a clear example of this. Along with other constraints, subtitles have to provide emotionally intense experiences and, accordingly, they were found to include more emotionally arousing words and more extreme valence and dominance words than daily conversational language (Baayen et al., 2016; Heister & Kliegl, 2012).

Given the particularities of the Flickr repository and its tag corpus, what can we expect from a word frequency measure based on such data (i.e., Flickr frequency)? Besides capturing the distribution of word forms in a selective portion of linguistic experience, Flickr frequency (as any frequency measure) is expected to be influenced by the specific peculiarities of the source data it is based on.

First of all, here, the source data is after all a lexical corpus, and therefore, it is expected to have commonalities with standard word frequency measures. Although Flickr tags are not arranged into sentences but produced as isolated labels, image tags share the same lexicon (and, associated to that, the same conceptual space) on which classical corpora are built. Thus, it is expected that words used more often in general-domain corpora would also tend to be used more often in such a specific-domain corpus.

Moreover, Flickr being a social media, tag frequency is expected to present aspects in common with other word frequency measures based on similar platforms such as Twitter and Facebook (Herdağdelen & Marelli, 2017). Indeed, producing tags on Flickr parallels producing words on other social media. Users select the subject they want to communicate with others (i.e., the image on Flickr or the topic on Twitter) and the words to refer to them. As for typical web 2.0 platforms, word production in Flickr is a spontaneous human behavior: people actively produce words and are selective with what they produce, thus constructing a register arguably closer to natural language than traditional corpora (Herdağdelen & Marelli, 2017). Moreover, as for other social media, Flickr users are primarily motivated by social aspects (Stuart, 2012) and are expected to choose mainly a socially acceptable and desirable register and avoid socially inappropriate words. Thus, tags to refer to images are expected to be more positive than negative.

Beyond these commonalities, Flickr frequency is also expected to present its own peculiarities related to the fact that tags are specifically used in Flickr to refer to images (i.e., synchronously present visual stimuli). Although Flickr frequency is based on words (i.e., tags), the language on Flickr is subordinate to visual experience. Indeed, tags are selected to match/fit the photographs. For this reason, Flickr frequency is expected to manifest properties that differentiate it from classic text-based frequency measures.

First of all, tags are expected to refer to or describe something that can be visualized (i.e., that can be turned into a visual format such as an image). An indirect effect of this is that Flickr frequency should end up capturing some distribution of visual information in our experiences (we do not expect high Flickr frequency for extremely abstract things or for things that cannot be visualized). Notably, such visual experience is not necessarily only based on "real-world frequency" but also on "media frequency" from websites, movies, newspapers, books, comics, etc. These two are arguably correlated but do not completely overlap. Indeed, something that can be photographed in real life, in principle, can also be seen on Flickr, but the two types of experience tend to diverge for other visualizable entities: for example, the rate at which the tag for "lion" or "dragon" appears on Flickr reasonably approximates how people see lions or dragons online better than how they see them in real life. Moreover, beyond its validity as a proxy for actual visual experience, this massive repository of images also embeds a reliable portrait of the state of things in the world. Indeed, Menon et al. (2016) showed that images tagged on Flickr can be effectively used to extract robust approximations of animal wildlife population size.

Another critical aspect to consider is the image choice, which can induce certain biases in the measure (exactly as it is the choice of topic to be treated on other platforms such as Twitter or Facebook). The image selection process would depend heavily on the social motivation of uploading images in the hope of drawing others' attention (i.e., social communication) (Ames & Naaman, 2007). Thus, pictures are expected to be chosen according to their degree of visual salience (i.e., people take photos of objects or events that are meaningful, surprising, or in any other way salient to them, and not necessarily of computer keyboards or spoons that, although maybe dull, constitute a larger part of their actual visual experience), their aesthetic (i.e., the preference for beautiful images) and their social acceptability (e.g., sharing positive content over negative content). Likewise, the type of content uploaded is expected to be highly influenced by social trends.

To summarize, we expect Flickr frequency to be a word frequency measure with a hybrid status, lying at the intersection of language and visual experience and with

specific biases induced by being based on image-focused social media. To better understand which latent constructs *Flickr frequency* is tapping into, in the present study we first analyzed its relationships with other linguistic, perceptual, and affective variables. On the linguistic side, we expected to find Flickr frequency to correlate with other measures accounting for the linguistic experience with words. On the perceptual side, we expected Flickr frequency to be related to variables measuring perceptual properties of word-referents – specifically those more informative of visual properties. On the affective side, we were interested in evaluating to what extent Flickr frequency might exhibit a positive valence bias. In a second step, we then tested whether Flickr frequency norms explain additional behavioral variance in experimental tasks, beyond what is captured by previously published norms on both perceptual and linguistic experience with the word stimulus. Following the established practice in the field, we tested whether the Flickr frequency measure predicts word processing time – as measured in various large-scale behavioral datasets – over and above existing linguistic, sensorimotor, and affective measures.

Methods

Flickr frequency norms

Flickr frequencies were obtained from the Flickr photo-sharing platform (www.flickr.com). They were initially extracted for a total of 81,834 words resulting from the combination of entries from the English Lexicon Project (ELP) (including 40,481 American spelling English words) (Balota et al., 2007), the British Lexicon Project (including 28,730 British spelling English words) (Keuleers et al., 2012), and the English Crowdsourcing Project (including 61,851 American spelling English words) (Mandera et al., 2020). We did not apply any transformation or processing to the words as they appear in the three datasets, we used to construct the item set. A Python-based tool was developed to collect Flickr frequency data through the API method *flickr.photos.search* (https://www.flickr.com/services/api/flickr.photos.search.html), which returns the list of public photos[1] tagged with a specific label within a particular time interval and geographical area. As geographical areas, we extracted data separately for images uploaded in the US and the UK by defining bounding boxes delimiting the two areas (taken from https://gist.github.com/graydon/11198540)[2]. As time interval, we extracted data starting from January 1, 2005, to January 1, 2022. Since the Flickr API provides less accurate results when accessing larger sets of photos, the whole time interval was subdivided into five equal-sized sub-windows in case more than 20,000 photos were contained within that window. The subdivision was recursively performed until no API query returned more than 20,000 photos or up to a minimum time interval of 30 days. Results from each API query for each word were then summed together to form two datasets of Flickr frequency: one with the count of images tagged with each word label in the geographical area of the United States (i.e., *Flickr frequency US*) and one with the count of images tagged with each word label in the geographical area of the United Kingdom (i.e., *Flickr frequency UK*). Words never used as tags in Flickr were discarded (i.e., 28,133 words for Flickr frequency US; 34,943 for Flickr frequency UK) [3]. As a result, the Flickr frequency US corpus included 53,699 words, and the Flickr frequency UK dataset included 46,889 words.

Because Flickr tags are not arranged into sentences but are mainly isolated labels, the Flickr corpus includes more nouns, names, and adjectives than classic word frequency corpora. On the other hand, other parts of speech are less represented than in classic textual corpora (see Fig. 2 for the distribution of part of speech in Flickr and Subtlex US - Brysbaert et al., 2012). Misspelled tags are very infrequent (i.e., 0.12% based on an analysis by Stuart (2012) on a sample of 12,832 Flickr tags). The Python script for the extraction of Flickr frequency estimates is openly available on the Open Science Framework (https://osf.io/2zfs3/).

Skewness was used to evaluate the asymmetry of the distribution of Flickr frequency. The skewness of Flickr frequency US is 20.47, while the skewness of Flickr frequency UK is 70.14. It indicates that the Flickr frequency distribution is severely skewed towards the right, so four transformations (i.e., square root, cube root, Laplace [log(FF+1)] and Zipf [log10 (FF per billion words)][4] transformation) were

[1] Video and other types of images – e.g., screenshots – are filtered out from the search (see https://www.flickr.com/services/api/flickr.photos.search.html). Our test runs provide no evidence that the query's results contain any automatic tags in addition to user-selected tags.

[2] Note that the bounding boxes are square areas that include all the regions included between the North-South-West-East extremes of the nations belonging to the US and the UK. Therefore, they end up also including geographical areas that do not belong to US (i.e., a portion of Canada) and UK (i.e., a portion of Ireland).

[3] Following a common practice used for frequency measures over textual corpora (Brysbaert & Diependaele, 2013), control analyses were computed also on the entire dataset assuming frequency = 0 for words never used as tags in Flickr. The general pattern of results was consistent across the different datasets (see Tables S4 in Supplementary Materials).

[4] The Zipf scale goes from 1 to 7. Words with Zipf values lower or equal to 3 are considered low-frequency words; words with Zipf values higher or equal to 4 are considered high-frequency words (van Heuven et al., 2014).

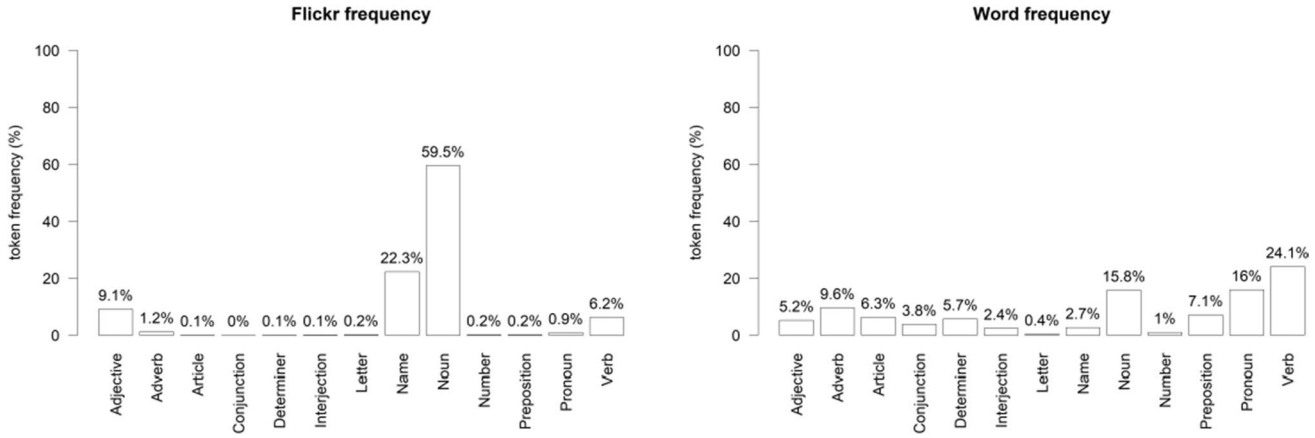

Fig. 2 Token frequency for various part of speech in Flickr US and Subtelx US corpora. The analysis is based on words for which the dominant part of speech is assigned in the Subtlex US dataset by Brysbaert et al. (2012)

used to try to normalize the distributions. In line with other word frequency measures (Brysbaert & Diependaele, 2013; van Heuven et al., 2014), the logarithmic transformations (i.e., Laplace and Zipf) were found to be those that most of all normalized the distribution of Flickr frequency with a skewness of 0.45 for Flickr frequency US and 0.58 for the Flickr frequency UK (see Fig. 3). Zipf-transformed Flickr frequency was chosen for subsequent analysis for its easier interpretability (van Heuven et al., 2014) (see Table 1 for examples of words at the seven levels of the Zipf scale). The correlation between Flickr frequency UK and Flickr frequency US is $r = .783$ ($p < .001$; based on 43,511 words shared between the two norming sets).

Other word-level measures

Word frequency measures

On the linguistic side, we considered as measures of linguistic experience the Zipf-transformed word frequency measures based on conversational corpora obtained from film subtitles:

- SUBTLEX-US corpus (Brysbaert et al., 2012) (including 74,286 words);
- SUBTLEX-UK corpus (van Heuven et al., 2014) (including 160,022 words).

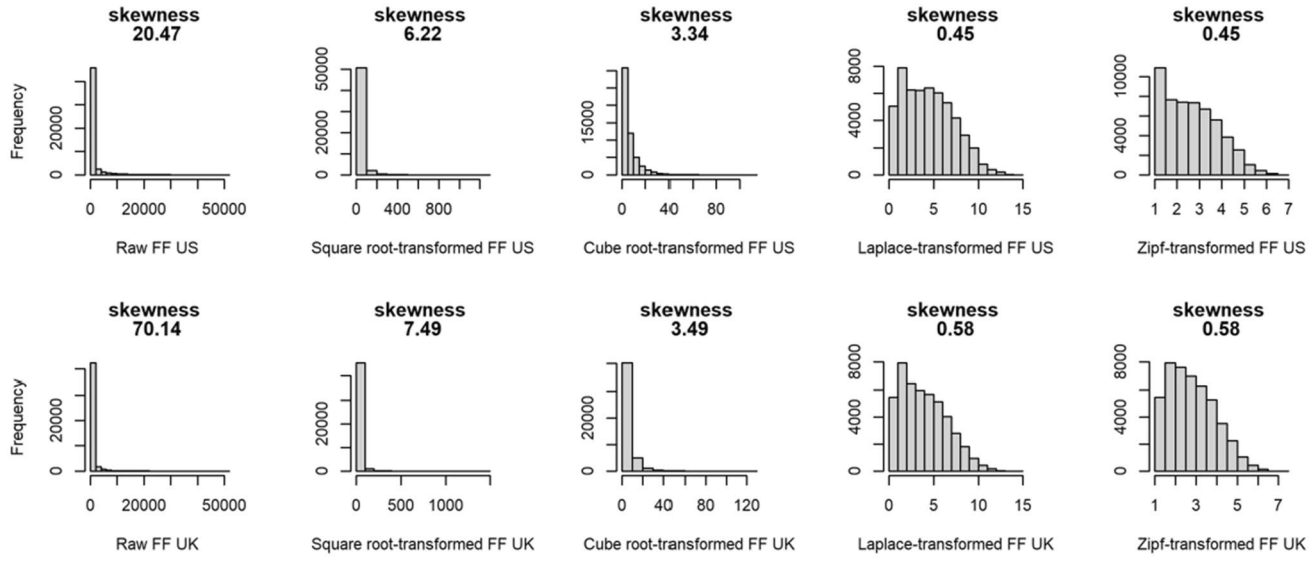

Fig. 3 Histograms and skewness values for raw and transformed (i.e., square root, cube root, Laplace, and Zipf transformations) Flickr frequency estimates

In addition, we considered Zipf-transformed word frequency norms extracted from social media, comprising:

- Twitter-based frequencies (Herdağdelen & Marelli, 2017) (including 37,144 words);
- Facebook-based frequencies on American English and British English datasets (Herdağdelen & Marelli, 2017) (including 37,144 words)
- Reddit-based word frequencies from a subset of the Reddit comments corpus (Hollis, 2020; https://osf.io/j3w6b/) (including 693,672 words).

Other linguistic measures

As additional lexical measures, we considered:

- Word Length, defined as the number of letters in a word;
- Orthographic Levenshtein Distance 20 (OLD20), defined as the mean distance (in terms of the number of letter insertions/deletions/substitutions) between the target and its closest 20 orthographic neighbors (extracted from the English Lexicon Project; Balota et al., 2007);
- Semantic Neighborhood Density, defined as the average degree of similarity between a target stimulus word and all other words in its semantic neighborhood (as derived from a global co-occurrence model) using a cut-off of 3.5 SDs (Danguecan & Buchanan, 2016) computed on a standard distributional semantic space (the best-performing word2vec cbow space by Baroni et al., 2014);
- Semantic Neighborhood Size, defined as the number of words in the semantic neighborhood (following the same definition as Semantic Neighborhood Density);
- Semantic Neighborhood Distance, defined as the average degree of similarity between a target word and its 20 closest semantic neighbors (see, for example, Marelli & Baroni, 2015; Vecchi et al., 2011).

As subjective measures of linguistic experience, we further considered:

- Age of Acquisition (i.e., the age at which a word is typically learned), extracted from the ratings collected by Kuperman et al., 2012 (including 30,121 words);
- Word Prevalence (i.e., the number of people who know the word), obtained in a rating study by Brysbaert et al., 2019 (including 61,858 words).

Sensorimotor measures

On the perceptual side, we extracted measures (all of them subjective rating measures) pertaining to the visual experience with a word's referent:

- concreteness ratings (i.e., how concrete vs abstract a word is) from the large-scale database by Brysbaert et al. (2014) (including 37,058 words);
- imageability ratings (i.e., how well a word gives rise to a mental image) from the Glasgow Norms by Scott et al. (2019) (including 5,553 words);
- visual strength ratings (i.e., how strongly a concept is experienced by seeing) from the Lancaster Sensorimotor Norms by Lynott et al. (2020) (including 39,707 words).

In addition to these vision-related ratings, we extracted perceptual ratings for all the other perceptual modalities (from the database by Lynott et al., 2020; including 39,707 words):

- auditory strength (i.e., how strongly a concept is experienced by hearing);
- gustatory strength (i.e., how strongly a concept is experienced by tasting);
- haptic strength (i.e., how strongly a concept is experienced by feeling through touch);
- interoceptive strength (i.e., how strongly a concept is experienced by sensations inside the body);
- olfactory strength (i.e., how strongly a concept is experienced by smelling).

We also considered different general-level operationalization of perceptual strength, aggregating the scores from the five perceptual modalities (from the database by Lynott et al., 2020, including 39,707 words):

- maximum perceptual strength (i.e., perceptual strength in the dominant modality);
- perceptual Minkowski-3 distance (i.e., perceptual strength in all the five dimensions with the attenuated influence of weaker dimensions)
- perceptual exclusivity (i.e., the extent to which a concept is experienced through a single perceptual modality).

Similarly, from the same database by Lynott et al. (2020) (61,858 words), we considered also the following action strength ratings:

- foot strength (i.e., how strongly a concept is experienced by executing an action with the foot/leg);
- hand strength (i.e., how strongly a concept is experienced by executing an action with the hand/arm);
- head strength (i.e., how strongly a concept is experienced by executing an action with the head excluding mouth);
- mouth strength (i.e., how strongly a concept is experienced by executing an action with the mouth/throat);
- torso strength (i.e., how strongly a concept is experienced by executing an action with the torso).

Finally, we considered corresponding general-level operationalization of action strength (from the database by Lynott et al., 2020, including 39,707 words):

- maximum action strength (i.e., action strength in the dominant effector);
- action Minkowski-3 distance (i.e., aggregated action strength in all effectors with the attenuated influence of weaker effector);
- action exclusivity (i.e., the extent to which a concept is experienced through a single effector).

Affective measures

Finally, we considered affective norms (again, subjective rating values) from the database by Warriner et al., (2013) (including 13,915 words):

- valence (i.e., the pleasantness of a concept);
- arousal (i.e., the intensity of emotion activated by a concept);
- dominance (i.e., the degree of control exerted by a concept).

Behavioral datasets

In order to empirically evaluate our Flickr frequency measures, we employed metrics from five large-scale behavioral datasets as dependent measures of word processing:

- one dataset containing speeded naming data from the English Lexicon Project (ELP-NM) megastudy (including 40,481 words) (Balota et al., 2007), collected from American-English speakers. In the naming paradigm, participants are presented with word stimuli and are instructed to read them out aloud as fast and as accurately as possible.
- Two datasets containing lexical decision data, one from the English Lexicon Project (ELP-LD) megastudy (Balota et al., 2007), collected from American-English speakers and the other one from the British Lexicon Project (BLP-LD) (including 28,730 words) (Keuleers et al., 2012), collected from British-English speakers. In lexical decision, participants are presented with letter string stimuli (real words and pseudowords) and have to decide for each of them whether it is an existing English word as fast and accurately as possible.
- One dataset containing recognition time data for 61,851 English words from the English Crowdsourcing Project (ECP-RC) (Mandera et al., 2020), collected mainly from American-English speakers. The word recognition task is very similar to the lexical decision task, with the differences that (a) judgements are not required under speeded conditions and (b) that participants are explicitly instructed only to indicate which words they knew, and not to guess in the cases in which they are unfamiliar with the presented string of letters (Mandera et al., 2020).
- One dataset containing concrete/abstract semantic decision data for 10,000 words from the Calgary semantic decision project (CAL-SD) (Pexman et al., 2017), collected from Canadian-English speakers. In the concrete/abstract semantic decision task, participants are presented with word stimuli and have to decide for each of them whether it refers to something concrete or abstract.

Statistical analyses

All statistical analyses were conducted in RStudio (RStudio Team, 2020). We analyzed our data using Pearson correlation and factor analyses via the *psych* (Revelle, 2021) R package; For factor analyses, we also used the *FactorAssumptions* (Storopoli, 2022) R package; multiple linear regression analyses were also conducted via the *lm.beta* (Behrendt, 2014), *relaimpo* (Grömping, 2007), *car* (Fox & Weisberg, 2019), *rwa* (Chan, 2020) R packages.

Each analysis reported in the manuscript is computed on datasets resulting from the combination of words from the Flickr frequency databases (UK or US, separately) and the databases of the other variables involved in the analysis. For example, correlation analyses between Flickr frequency and word frequencies based on textual corpora were computed for a dataset resulting from the combination of words from the word frequency databases (listed in the paragraph " Other word-level measures") and the words from either the Flickr frequency US database or the Flickr frequency UK database. This was done to provide results from analyses conducted each time on the largest possible dataset. Additionally, in order to provide also results from analyses conducted systematically on the same item sets, in supplementary materials, we also report: 1) results for statistical analyses conducted on a database resulting from the combination of words from all the control variable databases (listed in the paragraph "Other word-level measures"), the words from the Flickr frequency US database, and the words listed in the ELP; 2) results for statistical analyses conducted on a database resulting from the combination of words from all the control variable databases (listed in the paragraph " Other word-level measures"), the words from the Flickr frequency UK database, and the words listed in the BLP (see Tables in the document S3 in Supplementary Materials, https://osf.io/2zfs3/). Following a common practice (Brysbaert & Diependaele, 2013), control analyses were computed also on the entire dataset assuming frequency = 0 for words never used as tags in Flickr (see Tables S4 in Supplementary Materials, https://osf.io/2zfs3/). The general pattern of results was consistent across the different datasets.

Results

Which construct is Flickr frequency tapping into?

The effect of lexical-semantic variables on Flickr frequency

In a first step, we conducted a preliminary analysis evaluating how much variance of Flickr frequency is explained by other word-level measures. Thus, we fitted a regression model with Flickr frequency US and another model with Flickr frequency UK as the dependent variables and, as predictors, all the sensorimotor and linguistic measures listed above. Adjusted R^2 for the analysis with Flickr frequency US as a dependent variable was .62, and Adjusted R^2 for the analysis with Flickr frequency UK as a dependent variable was .60. This indicates that around 40% of the variance of Flickr frequency is not already explained by the combination of all the other variables and, therefore, it (partially) measures a different construct/latent variable that is not yet captured by the vast collection of metrics used in the literature.

Correlations between Flickr frequency and word frequencies based on textual corpora

In order to further support this interpretation, we tested the relationships between Zipf-transformed Flickr frequency (US and UK) and the linguistic variables using Pearson correlation coefficients. Correlations with Flickr frequency were computed for a dataset resulting from the combination of words from the word frequency databases (listed in the paragraph " Other word-level measures") and the words from either the Flickr frequency US dataset (N = 29,201 words) or the Flickr frequency UK dataset (N = 27,199 words). The correlations between Flickr frequency and traditional word frequency variables are displayed in Table 2. Flickr frequency US and UK exhibited a comparable pattern of correlation with word frequencies. Specifically, Flickr frequency had medium-to-high correlations with all the norms of word frequency (i.e., SUBTLEX-UK, Twitter frequency, SUBTLEX-US, Facebook frequency UK, Facebook frequency US, Reddit frequency) (i.e., the higher the Flickr frequency, the higher was the Word frequency), suggesting that these measures may largely tap into the same latent construct. However, these labels crucially differ from words from traditional text corpora by being produced in the presence of a visual referent and as the result of a visual evaluation. Indeed, correlations reported between Flickr frequencies and Word Frequencies (ranging from .43 and .56; see Table 1) were high, but not to the point of indicating an equivalence between word frequency and Flickr frequency norms. In fact, correlations between different word frequency norms were much stronger than those observed between Flickr and word frequencies (i.e., range Δr: .286 – .469 for correlations with Flickr frequency US; range Δr: .251 – .506 for correlations with Flickr frequency UK), consistently higher than r = .81 (see Table 3): even the difference between the smallest correlation between word frequency measures (i.e., r = .809) and the highest correlation between Flickr frequencies and word Frequencies (i.e., r = . 523 for Flickr frequency US; r = . 558 for Flickr frequency UK) was highly significant (US: z = 65.78, $p < .001$; UK: z = 57.53, $p < .001$).

Correlations between Flickr frequency and other linguistic variables

In the next step, we tested the relationships between Zipf-transformed Flickr frequency (US and UK) and the other linguistic variables (listed in the paragraph "Other word-level

Table 2 The first two columns report the correlations between Flickr frequency (US and UK) with traditional word frequency norms

	Flickr frequency US	Flickr frequency UK	Subtlex frequency US	Subtlex frequency UK	Twitter frequency	Facebook frequency UK	Facebook frequency US	Reddit frequency
Subtlex frequency US	**.523**	.496		**.85**	**.873**	**.841**	**.871**	**.867**
Subtlex frequency UK	**.52**	**.558**			**.847**	**.88**	**.809**	**.825**
Twitter frequency	**.506**	.479				**.921**	**.94**	**.909**
Facebook frequency UK	.483	**.507**					**.907**	**.851**
Facebook frequency US	.478	.436						**.902**
Reddit frequency	.471	.437						

The remaining columns on the right report correlations between word frequency measures. Correlations are computed on two datasets resulting from the combination of words included in word frequency databases (listed in the paragraph " Other word-level measures") with words from Flickr frequency US and words from Flickr frequency UK. All correlations are statistically significant ($p < .001$). High correlations ($r > .5$) are shown in bold. Medium correlations (i.e., $.3 < r < .5$) are underlined

Correlation analyses with Flickr frequency UK – column 2 -, N = 27,199 words. Other correlation analyses– columns 1, 3–6, N = 29,201 words

measures") using Pearson correlation coefficients. In order to highlight similarities and peculiarities of Flickr frequency with respect to classical textual-based word frequency measures, the correlations between Zipf-transformed text-based word frequencies and the other linguistic variables were also tested. Thus, in a former case (i.e., US analyses), correlations were computed for a dataset resulting from the combination of words from the Flickr frequency US dataset, the word frequency datasets based on US textual corpora (i.e., Subtlex US and Facebook US; here summarized in a unique measure by averaging Zipf Subtlex US and Zipf Facebook US), and the linguistic measures databases (listed in the paragraph " Other word-level measures- Other Linguistic Measures") ($N = 17{,}987$). In a latter case (i.e., UK analyses), correlations were computed for a dataset resulting from the combination of words from the Flickr frequency UK dataset, the word frequency datasets based on UK textual corpora (i.e., Subtlex UK and Facebook UK; here summarized in a unique measure by averaging Zipf Subtlex UK and Zipf Facebook UK), and the linguistic measures databases ($N = 16{,}781$). The correlations are displayed in Table 3.

Flickr frequency values exhibited a comparable pattern of correlation with the other variables, with overall medium correlations with Age of Acquisition (i.e., the higher the Flickr frequency, the lower was the age of acquisition of the word) as well as Word Prevalence (i.e., the higher the Flickr frequency, the larger was the number of people who know the word) (see Table 3). Along the same line, Flickr frequency showed medium negative correlations with word length (i.e., the higher the Flickr frequency, the smaller was the orthographic length of its word-tag), reproducing the well-known inverse relationship between word length and word frequency (Sigurd et al., 2004; Zipf, 1935). Finally, small overall correlations emerged between Flickr frequency and measures of orthographic and semantic density (i.e., OLD20, Semantic Neighborhood Density, Semantic Neighborhood Distance, and Semantic Neighborhood). Word frequency based on textual corpora and Flickr frequency exhibited a comparable pattern of correlations with the other linguistic variables, although correlations with the former tended to be larger (see Table 3). This might be explained by word frequencies from traditional textual corpora reflecting a larger subset of our actual language experience compared to Flickr frequency.

Correlation between Flickr frequency and sensorimotor norms

We tested the correlations between sensorimotor variables with Zipf-transformed Flickr frequency (US and UK). Again, also the correlations between Zipf-transformed text-based word frequencies and sensorimotor variables were also tested. In one case (i.e., US analyses), correlations were computed for a dataset resulting from the combination of words from the Flickr frequency US dataset, the word frequency datasets based on US textual corpora (i.e., Subtlex US and Facebook US; summarized in a unique measure by averaging Zipf Subtlex US and Zipf Facebook US), and the sensorimotor measure databases (listed in the paragraph " Other word-level measures - Sensorimotor Measures") ($N = 4361$). In another case (i.e., UK analyses), correlations were computed for a dataset resulting from the combination of words from the Flickr frequency UK dataset, the word frequency datasets based on UK textual corpora (i.e., Subtlex UK and Facebook UK; summarized in a unique measure by averaging Zipf Subtlex UK and Zipf Facebook UK), and the sensorimotor measure databases ($N = 4299$). The correlations are displayed in Table 4.

Flickr frequency US and Flickr frequency UK exhibited a comparable pattern of correlations, although overall Flickr frequency US was more correlated with perceptual variables than Flickr frequency UK. The highest correlations emerged between Flickr frequency and Imageability (i.e., the higher the Flickr frequency, the more imageable the word), followed by Concreteness (i.e., the higher the Flickr frequency, the more

Table 3 Correlations between Flickr frequency (US and UK) and text-based word frequency (here summarized as Word frequency US = mean Zipf value of Subtlex US and Facebook US - and word frequency UK = mean Zipf value of Subtlex UK and Facebook UK) with the other linguistic variables. High correlations ($\boldsymbol{r} > .5$) are shown in bold. Medium correlations (i.e., $.3 < \boldsymbol{r} < .5$) are underlined. "ns" indicates non-significant correlations ($p > .05$)

	Flickr frequency US	Flickr frequency UK	Word frequency US	Word frequency UK
Age Of Acquisition	- .42	- .4	**- .614**	**- .582**
Length	- .372	- .36	- .451	- .441
Prevalence	.333	.315	**.57**	**.561**
OLD20	- .312	- .316	- .465	- .468
Semantic Neighborhood Density	- .225	- .24	- .391	- .432
Semantic Neighborhood Distance	- .074	- .074	- .186	- .185
Semantic Neighborhood Number	.038	.042	- .027	- *.002* ^{ns}

Correlation analyses with Flickr frequency US and Word frequency US - columns 1, 3 -, $\boldsymbol{N} = 17{,}987$ words. Correlation analyses with Flickr frequency UK and Word frequency UK, - columns 2, 4 - $\boldsymbol{N} = 16{,}781$ words

concrete the word) and Visual Perceptual Strength (i.e., the higher the Flickr frequency, the higher the visual strength). Flickr frequency also exhibited medium correlations with general-level operationalization of perceptual strength (i.e., maximum perceptual strength and perceptual Minkowski-3 distance; the higher the Flickr frequency, the higher the perceptual strength). Smaller correlations emerged between Flickr frequency and all the other perceptual measures. The correlations between Flickr frequency and the various operationalization of action strength were small and consistently below .2. This correlation pattern suggested that Flickr frequency measures are informative of perceptual properties of word referents, specifically, properties related to the *visual* domain. To illustrate the relationship between Flickr frequency and concreteness in Fig. 4, we show the distribution of concreteness as a function of the residuals from Flickr frequency US after partialling out the effect of word frequency measures based on US textual corpora (i.e., the residuals of the linear regression: *lm (FlickrFreqUS ~ SubtlexFreqUS + FacebookFreqUS)*). This metric is aimed at capturing the portion of variance uniquely captured by the Flickr measures, once the information encoded in word occurrences from more traditional corpora is accounted for. As can be seen from the figure, concrete words (in red) tend to be distributed in the upper part of the axes of the residuals, while abstract words (in light blue) tend to be distributed in the lower part of the axes of the residuals. This indicates that the type of information specifically encoded in Flickr frequency seems to capture visual properties of the word referents, such as the availability of its visual representation in the real-world or media environment.

Differently, all the correlations between traditional word frequency measures with sensorimotor variables were small, in most cases not reaching an *r* of .20 (Table 4). Unlike Flickr frequency, the correlation pattern did not indicate word frequencies to capture the visual properties of the word referent. If anything, the highest correlations emerged between general-level operationalization of Action Strength, such as Max Action Strength and Minkowski 3 Action, for which *r was* in some cases slightly higher than .20 (see Table 4).

These results indicate that Flickr frequency does not entirely resemble other word frequency norms. On the opposite, they show evident dissociations in how Flickr frequency and traditional word frequency measures relate to variables in the perceptual domain.

Table 4 Correlations between Flickr frequency (US and UK) and text-based word frequency (here summarized as Word frequency US = mean Zipf value of Subtlex US and Facebook US - and word frequency UK = mean Zipf value of Subtlex UK and Facebook UK) with sensorimotor and affective variables

	Flickr frequency US	Flickr frequency UK	Word frequency US	Word frequency UK
Imageability	**.508**	.475	.036	.049
Visual Perceptual Strength	.451	.422	.04	.055
Concreteness	.445	.414	*-.017* ^{ns}	*-.011* ^{ns}
Max Perceptual Strength	.388	.349	.086	.087
Minkowski 3 Perceptual Distance	.347	.3	.137	.125
Interoceptive Perceptual Strength	-.235	-.251	.162	.106
Haptic Perceptual Strength	.233	.214	.07	.072
Foot Action Strength	.194	.192	.134	.135
Hand Action Strength	.189	.18	.104	.103
Olfactory Perceptual Strength	.177	.143	.058	.068
Mouth Action Strength	-.151	-.179	.111	.083
Minkowski 3 Action	.098	.06	.22	.18
Torso Action Strength	.093	.084	.136	.109
Perceptual Exclusivity	.09	.106	-.142	-.119
Gustatory Perceptual Strength	.08	*.036* ^{ns}	.051	.064
Max Action Strength	.065	*.022* ^{ns}	.188	.147
Auditory Perceptual Strength	-.047	*-.035* ^{ns}	.117	.089
Head Action Strength	.046	*.024* ^{ns}	.168	.141
Action Exclusivity	-.043	-.056	-.092	-.093

High correlations ($r > .5$) are shown in bold. Medium correlations (i.e., $.3 < r < .5$) are underlined. "ns" indicates non-significant correlations ($p > .05$)

Correlation analyses with Flickr frequency US and Word frequency US - columns 1, 3 -, $N = 4361$ words. Correlation analyses with Flickr frequency UK and Word frequency UK, - columns 2, 4 - $N = 4299$ words

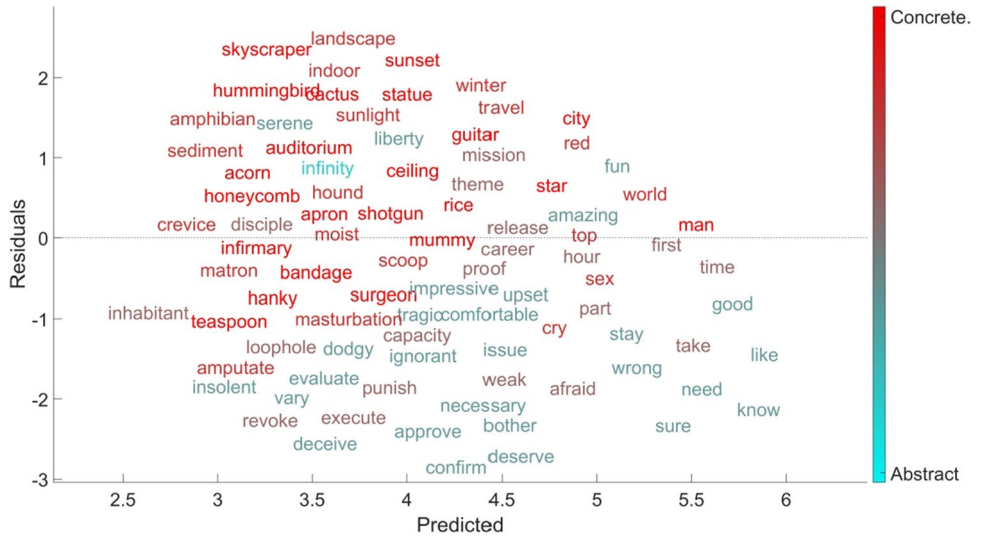

Fig. 4 Scatter plot showing for sample words the distribution of concreteness as a function of Flickr frequency US when word frequency is partialled out. Specifically, it plots the predicted vs residual value from a regression model with Zipf-Flickr frequency US as the dependent variables and, as predictors, Zipf-Subtlex US and Zipf-Facebook US. This metric estimates the portion of variance uniquely captured by the Flickr norms once the information encoded in word occurrences from more traditional corpora is accounted for. Words are colored according to the degree of concreteness (from red for concrete words to light blue for abstract words)

Correlations between Flickr frequency and affective norms

.We tested the correlations between affective variables and Zipf-transformed Flickr frequency (US and UK). As before, the correlations between affective variables and Zipf-transformed text-based word frequencies were also tested. In one case (i.e., US analyses), correlations were computed for a dataset resulting from the combination of words from the Flickr frequency US dataset, the word frequency datasets based on US textual corpora (i.e., Subtlex US and Facebook US; here summarized in a unique measure by averaging Zipf Subtlex US and Zipf Facebook US), and the affective measures databases (listed in the paragraph "Other word-level measures - Affective Measures") ($N = 12{,}134$). In the other case (i.e., UK analyses), correlations were computed for a dataset resulting from the combination of words from the Flickr frequency UK dataset, the word frequency datasets based on UK textual corpora (i.e., Subtlex UK and Facebook UK; here summarized in a unique measure by averaging Zipf Subtlex UK and Zipf Facebook UK), and the affective measures databases ($N = 11{,}666$). The correlations are displayed in Table 5.

Flickr frequency US and Flickr frequency UK exhibited a comparable pattern of correlations with affective variables: specifically, both exhibited medium/small correlations with valence (i.e., the higher the Flickr frequency, the more pleasant is the word), small correlations with Dominance (i.e., the higher the Flickr frequency, the higher is the dominance of the word) and close to zero correlations with arousal. Word frequencies and Flickr frequencies exhibited a similar pattern of correlation with affective variables. If anything, Flickr frequency exhibited a slightly stronger correlation with valence (US: $\Delta r = .100$; UK: $\Delta r = .057$).

Exploratory factor analysis

To further examine the relationships between Flickr frequency and the other variables, we conducted two exploratory factor analyses. Specifically, we considered the items resulting from the combination of words included in all sensorimotor, affective, and linguistic variables (listed in the " Other word-level measures" paragraph) and the words from

Table 5 Correlations between Flickr frequency (US and UK) and text-based word frequency (here summarized as Word frequency US = mean Zipf value of Subtlex US and Facebook US- and word frequency UK = mean Zipf value of Subtlex UK and Facebook UK) with the affective variables. High correlations ($r > .5$) are shown in bold. Medium correlations (i.e., $.3 < r < .5$) are underlined. "ns" indicates non-significant correlations ($p > .05$)

	Flickr frequency US	Flickr frequency UK	Word frequency US	Word frequency UK
Valence	.309	.287	.207	.23
Dominance	.187	.184	.194	.211
Arousal	- .055	- .068	.024	-.001 ^{ns}

Correlation analyses with Flickr frequency US and Word frequency US - columns 1, 3 -, $N = 12{,}134$ words. Correlation analyses with Flickr frequency UK and Word frequency UK, - columns 2, 4 - $N = 11{,}666$ words

either the Flickr frequency US data (3976 items) or the UK data (3933 items).

The two factor analyses yielded the same pattern of results. After communality check and Kaiser-Meyer Olkin analysis (cut-off 0.5; Hair et al., 2018; Kaiser, 1974), seven variables (i.e., Auditory Perceptual Strength, Head Action Strength, Mouth Action Strength, Arousal, Prevalence, Semantic Neighborhood Number) were dropped from both factor analyses. The resulting datasets were suitable for factor analyses, with 30 variables having a KMO index consistent with good sampling adequacy (KMO = 0.80 in the analysis on the US dataset; KMO = 0.79 in the analysis on the UK dataset) and a significant Bartlett Test of Sphericity (all p <.001). Regarding the number of factors, Horn's parallel analysis (Horn, 1965) indicated that an eight-factor solution was the optimal description of the data for both factor analyses (parallel analysis; 74 % explained variance in both factor analyses). The results of the solution with orthogonal rotation (i.e., Varimax) are shown in Table 6A(for the US dataset) and 6B (for the UK dataset).

The loading pattern of Flickr frequency US and UK were comparable. In the analyses, Flickr frequency loaded equally on both factor 1 and factor 2, with minimal loadings (all loadings < .2) on other factors. Together with Flickr frequency, seven variables loaded onto Factor 1 (Twitter frequency, Facebook frequency US, Facebook frequency UK, Subtlex frequency US, Reddit frequency, Subtlex frequency UK, Age Of Acquisition). It is clear from Table 6A and 6B that these items all relate to the latent construct of linguistic experience with words. Besides Flickr frequency, five more variables loaded on factor 2 (Concreteness, Imageability, Visual Perceptual Strength, Interoceptive Perceptual Strength, Haptic Perceptual Strength), all related to perceptual experience mainly in the visual modality (for the relation between the visual and haptic/interoceptive modalities – reflecting that concepts that can be touched/felt can also be seen, see for example Lynott et al., 2020; Vergallito et al., 2020).

These results confirm that Flickr frequency is a hybrid measure where language (in the form of verbal category labels) and visual information collide.

Does Flickr frequency provide additional information about lexical processing over existing norms?

Finally, we tested the contribution of Flickr frequency in explaining lexical processing . Thus, we tested for each behavioral measure of word processing whether Zipf-transformed Flickr frequency measures explain variance in RTs to linguistic stimuli over and above *all* the other linguistic, sensorimotor, and affective variables considered in this study. To this end, we conducted two-step hierarchical regression analyses on response times for the ELP-NM, ELP-LD, ECP-RC, BLP-LD, and CAL-SD [5]. In the first step (baseline model), we fitted a model with logarithmically transformed reaction time measures (Baayen & Milin, 2010) as the dependent variable and, as predictors, all other measures listed in the " Other word-level measures" paragraph. In the second step (test model), we added to the baseline model the Flickr frequency estimates as a predictor and assessed whether this additional parameter improved the baseline model fit. As estimates of Flickr frequency, we used Flickr frequency US for the analyses predicting RTs in the ELP, ECP, and CAL (collected from American or Canadian participants), while Flickr frequency UK was used for the analyses on BLP RTs (collected from British participants). Analyses for the CAL dataset were conducted separately for abstract (CAL-ABS) and concrete (CAL-CON) words. Given the relatively small size of the Glasgow Norms dataset and the affective database by Warriner et al. (2013), regression analyses were separately conducted for “larger-scale" datasets, excluding Imageability and affective variables (i.e., valence, dominance, and arousal), and “smaller-scale” datasets, including also these variables as predictors [6]. Results from all the regression analyses are shown in Table 7.

The comparisons between the baseline and the test models in the large-scale and small-scale datasets exhibited a similar pattern of results. These comparisons indicated that Flickr frequency improved the model fit in the ELP-NT (large scale: $F(1, 17627) = 5.649$, $p = .017$; small scale: $F(1, 3939) = 5.582$, $p = .018$), ELP-LD (large scale: $F(1, 17334) = 52.967$, $p < .001$; $F(1, 3936) = 11.03$, $p < .001$)), BLP-LD (large scale: $F(1,10027) = 45.822$, $p < .001$; small scale: $F(1, 2954) = 3.758$, $p = .053$), ELP-RC (large scale: $F(1, 17666) = 117.45$, $p < .001$; small scale: $F(1, 3939)$ 35.67, $p < .001$), with Flickr frequency, consistently exhibiting a facilitatory effect on RTs (i.e., the higher the Flickr frequency, the faster is the processing time).

[5] Variance inflation factors for Flickr frequencies were less than 3 (ranging from 1.5 to 2.7) in each regression model, so multicollinearity was not a concern (Hair et al., 2018, p. 316).

[6] Supplementary Materials (see Table S1, https://osf.io/2zfs3/) also reports the same analyses conducted on accuracy for which an improvement in the baseline model fit was found in the ELP-LD (large-scale dataset) and CAL-SDabs (large- and small-scale dataset). Note that a baseline model fit improvement was also found in the BLP-LD condition. However, in this condition, the direction of the effect is on the opposite direction than expected (the higher the Flickr frequency the lower the accuracy). An ad hoc analysis using accuracy as the dependent variable and Flickr frequency as a unique predictor shows a significant effect in the expected direction (the higher the Flickr frequency, the higher the accuracy). This suggests that the unexpected result is plausibly due to a suppression effect that emerges when Flickr frequency is added as a predictor together with the other word-level variables.

Table 6 Standardized loadings (pattern matrix) based upon correlation matrix for the eight-factor solution using orthogonal (i.e., Varimax) rotation of the loading matrix. Primary loadings > ± 0.40 are shown in bold

A	MR1	MR2	MR3	MR6	MR4	MR8	MR7	MR5	h2	u2	com
Flickr frequency US	**.50**	**.50**	.08	-.11	-.03	-.08	.20	.03	.56	.437	2.6
Twitter frequency	**.95**	.00	.03	.11	.04	-.11	.07	-.03	.94	.056	1.1
Facebook frequency US	**.95**	-.02	.05	.10	.06	-.10	.06	-.03	.92	.075	1.1
Facebook frequency UK	**.93**	.04	.05	.07	.05	-.13	.09	-.04	.90	.098	1.1
Subtlex frequency US	**.92**	.04	.07	.04	.00	-.10	.03	.00	.86	.137	1.0
Reddit frequency	**.91**	-.03	.03	.06	.04	-.06	.02	-.03	.84	.159	1.0
Subtlex frequency UK	**.89**	.05	.06	-.02	.01	-.06	.14	-.05	.83	.166	1.1
Age Of Acquisition	**-.56**	**-.45**	-.01	-.14	-.04	.18	-.09	.03	.58	.422	2.4
Imageability	.02	**.87**	.06	.01	.02	-.12	.04	.10	.79	.209	1.1
Concreteness	-.05	**.85**	.02	-.04	-.02	-.22	.03	.07	.79	.215	1.2
Max Perceptual Strength	.10	**.79**	-.07	.24	.06	.07	-.06	.06	.72	.285	1.3
Visual Perceptual Strength	.05	**.79**	.07	-.10	-.02	.05	.03	-.02	.64	.358	1.1
Minkowski 3 Perceptual Distance	.13	**.75**	-.01	.30	.35	.09	-.07	.05	.81	.193	1.9
Haptic Perceptual Strength	-.01	**.57**	.24	.14	.27	-.18	.08	-.10	.53	.466	2.4
Interoceptive Perceptual Strength	.14	**-.46**	.21	.32	.25	.15	-.21	.00	.51	.492	3.9
Action Exclusivity	-.06	.10	**-.81**	.27	-.08	-.07	.05	-.02	.76	.237	1.3
Torso Action Strength	.07	.02	**.78**	.25	.05	.01	-.06	.05	.69	.312	1.3
Foot Action Strength	.09	.08	**.78**	.13	-.10	-.02	-.01	-.01	.65	.350	1.1
Hand Action Strength	.02	.38	**.54**	.26	.03	-.14	.12	-.10	.55	.447	2.7
Max Action Strength	.13	.09	-.01	**.95**	.09	.01	.08	.03	.94	.064	1.1
Minkowski 3 Action	.14	.10	.36	**.88**	.10	.01	.07	.04	.95	.051	1.5
Gustatory Perceptual Strength	.02	.15	-.13	.11	**.77**	-.03	.09	.02	.66	.342	1.2
Olfactory Perceptual Strength	.03	.27	-.04	.01	**.74**	-.02	.02	.08	.63	.370	1.3
Perceptual Exclusivity	-.08	.24	-.25	-.05	**-.72**	-.03	-.01	.04	.64	.359	1.5
Length	-.27	-.16	.01	.02	-.01	**.87**	.05	.07	.86	.139	1.3
OLD20	-.31	-.11	-.02	.01	.00	**.82**	.04	.12	.80	.197	1.4
Dominance	.15	.00	-.01	.08	.04	.01	**.82**	-.05	.71	.293	1.1
Valence	.20	.11	-.03	.03	.07	.06	**.82**	.02	.73	.270	1.2
Semantic Neighborhood Density	-.13	.03	-.03	.01	.03	.06	-.04	**.84**	.73	.265	1.1
Semantic Neighborhood Distance	.01	.08	.03	.03	.02	.09	.02	**.81**	.67	.329	1.1

B	MR1	MR2	MR4	MR3	MR6	MR8	MR7	MR5	h2	u2	com
Flickr Frequency UK	**.48**	**.47**	.09	-.14	-.06	-.10	.20	.03	.54	.465	2.7
Twitter Frequency	**.95**	-.01	.03	.11	.04	-.11	.07	-.02	.94	.058	1.1
Facebook Frequency US	**.94**	-.03	.05	.11	.06	-.10	.06	-.03	.92	.078	1.1
Facebook Frequency UK	**.93**	.03	.05	.07	.05	-.12	.09	-.03	.91	.094	1.1
Subtlex Frequency US	**.92**	.03	.06	.04	.00	-.10	.02	.00	.86	.139	1.0
Reddit Frequency	**.91**	-.03	.03	.07	.04	-.06	.02	-.03	.84	.159	1.0
Subtlex Frequency UK	**.90**	.04	.06	-.03	.01	-.06	.15	-.05	.84	.160	1.1
Age Of Acquisition	**-.56**	**-.44**	-.01	-.14	-.05	.18	-.09	.03	.57	.427	2.4
Imageability	.01	**.87**	.06	.01	.02	-.12	.04	.11	.79	.210	1.1
Concreteness	-.06	**.85**	.02	-.04	-.02	-.23	.03	.07	.78	.217	1.2
Max Perceptual Strength	.10	**.79**	-.07	.24	.06	.08	-.06	.06	.72	.284	1.3
Visual Perceptual Strength	.05	**.79**	.06	-.10	-.02	.05	.03	-.01	.64	.360	1.1
Minkowski 3 Perceptual Distance	.12	**.75**	-.01	.30	.35	.10	-.07	.06	.81	.190	1.9
Haptic Perceptual Strength	-.01	**.58**	.24	.14	.27	-.18	.09	-.10	.54	.465	2.4
Interoceptive Perceptual Strength	.14	**-.47**	.21	.33	.25	.15	-.21	.00	.51	.487	3.9
Action Exclusivity	-.06	.10	**-.81**	.27	-.08	-.07	.05	-.02	.76	.239	1.3
Torso Action Strength	.07	.03	**.78**	.25	.05	.02	-.06	.05	.69	.306	1.3

Table 6 (continued)

Foot Action Strength	.09	.08	**.78**	.13	-.10	-.02	-.01	-.01	.65	.352	1.1
Hand Action Strength	.02	.38	**.54**	.26	.03	-.14	.12	-.10	.55	.448	2.7
Max Action Strength	.13	.09	-.01	**.95**	.09	.01	.08	.03	.93	.068	1.1
Minkowski 3 Action	.14	.10	.36	**.88**	.10	.01	.07	.04	.95	.050	1.5
Gustatory Perceptual Strength	.02	.14	-.13	.11	**.77**	-.03	.09	.02	.66	.342	1.2
Olfactory Perceptual Strength	.03	.26	-.04	.01	**.74**	-.02	.02	.08	.63	.374	1.3
Perceptual Exclusivity	-.08	.24	-.25	-.05	**-.72**	-.03	-.01	.04	.64	.358	1.5
Length	-.27	-.16	.02	.02	-.02	**.87**	.05	.07	.86	.142	1.3
OLD20	-.31	-.11	-.01	.01	.00	**.82**	.04	.12	.80	.197	1.4
Dominance	.15	.00	-.01	.08	.04	.02	**.83**	-.05	.72	.280	1.1
Valence	.19	.10	-.03	.03	.07	.06	**.81**	.03	.71	.289	1.2
Semantic Neighborhood Density	-.13	.04	-.03	.02	.03	.06	-.04	**.84**	.73	.273	1.1
Semantic Neighborhood Distance	.02	.08	.04	.03	.02	.09	.02	**.81**	.68	.317	1.1

The column "h2" contains the component communalities (i.e., the amount of variance in each index variable explained by the factors). The column "u2" contains the factor uniquenesses (i.e., the amount of variance not accounted for by the components—or 1–h2). The column "com" reports Hoffman's index of complexity for each item (i.e., the number of latent components required to account for the observed variables)

Also, the comparisons between the baseline and test models predicting RTs in the abstract condition of the semantic decision task improved the model fit (large scale: F(1,3223) = 39.904, $p < .001$; small scale: F(1,653) = 14.94, $p < .001$). The only exception was the semantic decision task with concrete words, for which Flickr frequency did not improve the baseline model fit (large scale: $F(1,3898) = .0005$, $p = .981$; small scale: $F(1,745) = 1.609$, $p = .205$) [7].

Results from the semantic decision task deserve further consideration. This task is the only one where Flick frequency shares a substantial portion of variance in explaining RT with some other perceptual variables: variables such as Interoceptive Perceptual Strength, Minkowski 3 Perceptual Distance, Concreteness, Visual Perceptual Strength, and Imageability are listed among the top ten variables sharing variance with Flickr frequency in explaining RTs in at least one condition of the CAL (but see also Table S1 in Supplementary Materials for similar results using Accuracy as dependent variable). In all the other tasks, perceptual variables never appear among these lists. Likewise, word frequency variables are in the CAL conditions those that share the smallest portion of variance with Flickr frequency in explaining RTs. Moreover, the CAL-ABS (i.e., analyses on the Calgary semantic database – subset with abstract words) represented an exceptional condition: here variables of lexical and perceptual experiences are dissociated in explaining RTs. On the perceptual side, the two variables contributing more to explaining RTs are concreteness and imageability, for which an inverse effect on RTs is shown (the more concrete and imageable is a word, the slower it is to judge an abstract word as abstract). Instead, the significant effects of word frequency (i.e., Facebook Frequency US in the CAL-ABS large-scale dataset and Facebook Frequency UK in the CAL-ABS small-scale dataset) go in the opposite direction: the higher is the word frequency, the faster it is to judge an abstract word as abstract. Interestingly, here, Flickr frequency, when added last in the model, behaves as a perceptual variable. The direction of the effect of Flickr frequency is consistent with concreteness and imageability (the higher is Flickr frequency, the slower it is to judge an abstract word as abstract). Thus, its effect seems to be dissociated from effects driven by word frequency variables. CAL-ABS seems to be the ideal condition for Flickr frequency to exhibit its unique contribution as a measure of lexical experience derived from a visual context. However, it is important to interpret these results with caution. Indeed, Flickr frequency, when added first in the model, exhibits a negative effect on reaction times (i.e., the higher the Flickr frequency, the faster is the participant in judging an abstract word as abstract), thus leaving open the possibility of a suppression effect caused by other word frequency variables when added together with Flickr frequency in the model.

The CAL-CON (concrete condition) instead is the only condition where Flickr frequency does not improve the baseline model fit. In this condition, effects of lexical frequency and perceptual variables have the same direction and seem to capture entirely the portion of variance otherwise explained by Flickr frequency if considered alone[8]. Note

[7] All results from analyses computed including zero word frequency were consistent with those computed on datasets excluding zero word frequencies. (see Tables in the document S4 in Supplementary Materials).

[8] Note that this is a very expected pattern: for any decision, the decision is easier if you are more familiar with the word – hence the facilitatory frequency effect. However, for concreteness the case is different: Making a "this is concrete" decision (i.e., in the CAL-CON condition) is easier the more concrete a word is. On the other hand, saying that "this is not concrete" (i.e., in the CAL-ABS condition) is harder when the thing is more concrete, and hence "abstract" RTs are slower.

Table 7 Summary of the results from all the regression analyses, including the change in R^2 between test and baseline models (i.e., ΔR^2), Flickr frequency estimate (standardized regression coefficients; i.e. **β**), Flickr frequency t-value (i.e. *t*), Flickr frequency p-value (i.e. *p*), Flickr frequency unique variance (contribution of Flickr frequency when included last in the model; i.e. uni var), Flickr frequency total variance (contribution of Flickr frequency when included first in the model; i.e. tot var) and Flickr frequency relative weight (i.e. rel w) scaled as a percentage of predictable variance (estimated through Relative Weights Analysis, Johnson, 2000); list of top ten control variables sharing variance with Flickr frequency (i.e. contribution shared by the control variable with Flickr frequency when included alone in the model as predictors). See Table S2 in Supplementary Materials (https://osf.io/2zfs3/) for detailed results for each predictor. Note: Number of items for the large-scale datasets: ELP-NT = 17,660; ELP-LD = 17,367; BLP-LD = 10,060; ECP-RC = 17,699; CAL-ABS = 3256; CAL-CON = 3,931. Number of items for the small-scale datasets: ELP-NT = 3976; ELP-LD N = 3973; BLP-LD = 2991; ECP-RC = 3967; CAL-ABS = 690; CAL-CON = 782

	Large-scale datasets (excluding imageability and affective variables)							Small-scale datasets (including imageability and affective variables)						
	R^2	*β*	*t*	*p*	*uni var*	*tot var*	*rel w*	R^2	*β*	*t*	*p*	*uni var*	*tot var*	*rel w*
ELP-NT														
ΔR^2	.0002							.0008						
FLICKR Freq US		-.002	-2.377	.017	2e-04	.1519	3.29		-.004	-2.363	.018	8e-04	.1191	3.49
Var Sharing Variance With FF	Twitter Freq, Facebook Freq UK, Subtlex Freq US Facebook Freq US, Subtlex Freq UK, Reddit Freq, Age Of Acquisition, Length, OLD20, Prevalence							Twitter Freq, Facebook Freq UK, Subtlex Freq UK, Subtlex Freq US, Facebook Freq US, Age Of Acquisition, Reddit Freq, Length, OLD20, Prevalence						
ELP-LD														
ΔR^2	.0012							.0012						
FLICKR Freq US		-.007	-7.278	<.001	.0012	.214	4.53		-.007	-3.594	<.001	.0014	.1855	4.06
Var Sharing Variance With FF	Twitter Freq, Subtlex Freq US, Facebook Freq US, Facebook Freq UK, Subtlex Freq UK, Reddit Freq, Age Of Acquisition, Length, OLD20, Prevalence							Twitter Freq, Subtlex Freq UK, Facebook Freq UK, Facebook Freq US, Subtlex Freq US, Age Of Acquisition, Reddit Freq, Length, OLD20, Prevalence						
BLP-LD														
ΔR^2	.0017							.0006						
FLICKR Freq UK		-.007	-6.769	<.001	.0017	.2275	5.69		-.003	-1.939	*.053*	6e-04	.1674	4.6
Var Sharing Variance With FF	Subtlex Freq UK, Facebook Freq UK, Twitter Freq, Subtlex Freq US, Reddit Freq, Facebook Freq US, Prevalence, Age Of Acquisition, Length, Semantic Neighborhood Density							Subtlex Freq UK, Facebook Freq UK, Twitter Freq, Subtlex Freq US, Facebook Freq US, Reddit Freq, Age Of Acquisition, Prevalence, Length, OLD20						
ECP-RC														
ΔR^2	.0017							.0034						
FLICKR Freq US		-.007	-10.84	<.001	.0017	.2494	4.53		-.006	-5.973	<.001	.0034	.2154	5.06
Var Sharing Variance With FF	Twitter Freq, Subtlex Freq UK, Facebook Freq UK, Subtlex Freq US, Facebook Freq US, Reddit Freq, Prevalence, Age Of Acquisition, Length, OLD20							Twitter Freq, Subtlex Freq UK, Facebook Freq UK, Facebook Freq US, Subtlex Freq US, Reddit Freq, Age Of Acquisition, Prevalence, Length, OLD20						
CAL-SD ABSTRACT														
ΔR^2	.0095							.0171						
FLICKR Freq US		.014	6.317	<.001	.0095	.0029	2.36		.019	3.865	<.001	.0171	7e-04	3.83
Var Sharing Variance With FF	Age Of Acquisition, Interoceptive Perceptual Strength, Length, OLD20, Prevalence, Head Action Strength, Semantic Neighborhood Density, Max Action Strength, Minkowski 3 Perceptual Distance, Minkowski 3 Action							Valence, Prevalence, Max Perceptual Strength, Minkowski 3 Perceptual Distance, Arousal, Dominance, Head Action Strength, Length, Interoceptive Perceptual Strength, Max Action Strength						
CAL-SD CONCRETE														
ΔR^2	1e-07							.0010						
FLICKR Freq US		0	.023	*.982*	<.0001	.1023	3.4		.007	1.268	*.205*	.001	.1156	2.16
Var Sharing Variance With FF	Age Of Acquisition, Concreteness, Twitter Freq, Subtlex Freq US, Facebook Freq US, Reddit Freq, Facebook Freq UK, Subtlex Freq UK, Prevalence, Visual Perceptual Strength							Age Of Acquisition, Imageability, Facebook Freq UK, Twitter Freq, Facebook Freq US, Subtlex Freq UK, Concreteness, Reddit Freq, Subtlex Freq US, Visual Perceptual Strength						

that CAL-CON is a challenging condition for Flickr frequency: in the CAL-CON condition, the dataset includes only concrete concepts, so the variability of Flickr frequency is necessarily more limited compared to the other behavioral datasets in which word frequency and perceptual effects have the same direction (Flickr frequency interquartile range: CAL-CON small-scale dataset = 1.32; other small-scale datasets mean = 1.40; min = 1.39; max = 1.42; CAL-CON large dataset = 1.46; other large-scale datasets mean = 1.72; min = 1.63; max = 1.77), and this may be the reason for the lack of effect of Flickr frequency in this condition. Taken together, these results indicate that Flickr frequency provides additional information concerning behavioral responses that goes beyond what is captured by existing norms of both linguistic and perceptual experience.

Discussion

In the present study, we examined Flickr frequency, a measure operationalized as the number of images uploaded on the Flickr photo-sharing platform that are tagged with a given word label, and hence a frequency metrics expected to capture the distribution of word forms in visual contexts. Correlation analyses support the assumption that Flickr frequency is a measure of lexical experience – relatively speaking, on the linguistic side, it is correlated and has commonalities with the other prominent word frequency measures capturing the same latent construct. At first glance, this might suggest that Flickr frequency is nothing more than another measure of lexical frequency, not so different from word frequency measures already reported in the literature. After all, one could claim that the present norms are based on human-produced labels, so data that are linguistic in nature. However, these labels crucially differ from words from traditional text corpora by being produced in the presence of (or even being elicited by) a visual referent, and consequently present some peculiarities that differentiate them from the whole family of traditional frequency measures. Indeed, on the perceptual side, the highest correlations of Flickr frequency are with various measures pertaining to the visual experience with a word referent, such as Imageability, Concreteness, and Visual Strength – besides general perceptual factors (Max Perceptual Strength, Minkowski 3 Perceptual Distance). It is important to emphasize that existing word frequency norms strongly diverge from Flickr frequency in this respect – the correlations between word frequency and all the other perceptual measures were small and without any preferential pattern toward a specific perceptual modality.

At this point, one may wonder whether Flickr frequency measure reflects a linguistic or a visual construct. Arguably, the results suggest that it does both. Flickr frequency, in fact, describes how we use language to describe the visual world and, conversely, how we organize this perceptual input into different labelled categories. Thus, Flickr frequency is inherently tied to words – which is why we evaluated it against word processing data and not image processing data. However, the role of visual information is not secondary. On Flickr, tags are selected to match/fit the photographs. Thus, words chosen in Flickr are affected by certain constraints of perceptual nature, giving Flickr frequency a hybrid status where language (in the form of verbal category labels) and visual information collide. This aspect is well highlighted by the results of the exploratory factor analysis, showing Flickr frequency to load not only on a linguistic factor clustering all word frequency variables but also on what appears to be a visual factor, based on common loading with visuo-perceptual variables.

Furthermore, we observed that the Flickr frequency measure predicts behavioral data in the form of processing times (naming, lexical decision, word recognition, and, in part, semantic decision) in large datasets of lexical processing, over and above a wide range of existing linguistic, sensorimotor and affective measures (Brysbaert et al., 2014, 2019; Brysbaert & New, 2009; Herdağdelen & Marelli, 2017; Hollis, 2020; Keuleers et al., 2012; Kuperman et al., 2012; Lynott et al., 2020; Scott et al., 2019; van Heuven et al., 2014; Warriner et al., 2013). These results indicate that this newly introduced measure captures additional information beyond what is already encoded in existing norms of linguistic and perceptual experience. Therefore, such information represents some aspects of language usage that traditional frequencies are missing: a portion of language usage capturing the interplay between language and vision, which – this study demonstrates – has its own impact on word processing.

It is worth noting that this combination between lexical information and perceptual experience is not simply specific to our measure but also represents a psychologically plausible mechanism of how conceptual representations are acquired and organized (see, for example, Wolff & Holmes, 2011 and Winter et al., 2018). On the one hand, clear evidence for this link between perceptual experience and linguistic labels is traceable in the cross-language differences of grouping concepts into named categories. An expression of this is known as the notion of the "*words for snow*", whereby languages spoken in warmer climates tend to not distinguish between the terms *snow* and *ice*, and they also less frequently refer to these concepts, thus suggesting that asymmetries in our perceptual experiences are reflected in our perceptual vocabulary. Furthermore, a growing number of studies show that experience with language organizes visual information. In this regard, Lupyan (2012b) reviewed a series of studies demonstrating that verbal labels do not simply refer to object representations but rather actively modulate them, affecting categorization and memory of objects (e.g., Lupyan, 2008). The most notable studies in this regard

have been conducted in the domain of colors (Brown & Lenneberg, 1954; Kay & Kempton, 1984; Özgen & Davies, 2002; Winawer et al., 2007) (but see also Gilbert et al., 2008 or Goldstein & Davidoff, 2008 for other studies focusing on visual categories other than colors; see also Lupyan et al., 2020 for a review on behavioral and electrophysiological evidence for the influence of language on visual perception). These studies demonstrated that the way people label things influences how they organize and process the corresponding perceptual representations. Consider, for example, the discrimination of different nuances within the blue spectrum. Winawer et al. (2007) showed that Russian speakers – who have distinct terms that inherently discriminate different types of *blue* – are better than English speakers – who have a unique term for the whole spectrum of *blue* – at distinguishing them in non-linguistic tasks. The superior ability to discriminate items identified by distinct vs unique lexical labels strongly support the role of language as an organizer for our visual representation, suggesting that, in visual experience, the two factors (i.e., linguistic and perceptual) cannot be completely disentangled: entities in our visual world are perceived as discrete things partly because we have words that specify the category they belong to. Likewise, Lupyan (2012a, 2012b) proposes to abandon the distinction between verbal and non-verbal representations in favor of a framework in which verbal labels actively modulate the ongoing processing of non-verbal object representations.

As mentioned in the introduction, this Flickr data is not free from biases. As for any frequency measure, the observed distribution depends substantially on the data source we are considering (Baayen et al., 2016). Indeed, the selection of images uploaded online is subjected to critical human filters, especially related to the motivation for uploading photographs on such social media. More specifically, social signaling/attention is one of the main motivating factors on Flickr (Stuart, 2012). The intent of drawing others' attention (i.e., social communication) (Ames & Naaman, 2007) drives the selection of the images and related tags: people do not just upload everything they see but favour images that are meaningful, surprising, or in any other way salient; they will select socially acceptable photos and tags; they will prefer images considered as more aesthetically pleasing. A clue into these biases is provided by Fig. 4. Concrete things that are not pleasant (e.g., "amputate" or "surgeon "), socially not appropriate (e.g., "sex") or simply not salient (e.g., "bandage") to be uploaded on such a social media are under-represented in the Flickr tags relative to their lexical frequency. On the opposite, even abstract words, especially when giving positive connotations to the photos, tend to be used more as a tag on Flickr (e.g.," serene", liberty", "fun"). A further indication of these biases is provided by the positive correlation between Flickr frequency and valence (i.e., people tend to use more positive words as tags): beautiful images are uploaded more often and are expected to be better described by pleasant words as tags.

Notably, however, biases are natural properties of any frequency norm. Indeed, one can make the same argument for word frequencies based on textual corpora: people do not talk about everything they experience, but talk about things they find interesting. In both textual and image-tag corpora, people arbitrarily select the topics and representations (i.e., concepts to talk about or images to show) and word labels to express them. The parallelism between traditional and Flickr frequency measures is even more evident if we consider that image labels (or, more broadly speaking, labels for visual scenes) are actually a subsample of the general language experience. Both traditional and Flickr frequency are based on word counts, although obtained from different corpora – the former from corpora approximating general language experience, and the latter from a corpus approximating language experience selectively used to refer to visual scenes. Given this parallelism, the same peculiarities affecting word Frequency measures will also affect Flickr frequency measure.

Importantly the methodology used in this study allowed us to characterize visual experience with word referents in a direct way. This methodology protects against the concerns in terms of objectiveness and reliability raised by adopting human-based ratings (of visual experience) as predictors (see Petilli et al., 2021). On a theoretical level, subjective ratings assess participants' introspection about their experience, thereby constituting an inherently psychological variable that taps into memory and metacognitive abilities, without offering much insight as to *why* people arrive at their judgments – why do speakers indicate a higher familiarity and visual strength for BUS than for COMORBIDITY? Word frequencies, on the other hand, tap into the actual experience – as approximated by data sources such as large-scale corpora that are taken to reflect a speaker (population)'s experience – itself. Having such a source of data available is highly desirable for psychological studies, as it allows to bypass the loophole of predicting behavioral data (e.g., lexical processing time) from other behavioral data (ratings) (Jones et al., 2015; Westbury, 2016) – which would leave us staying essentially at the same epistemological level of description rather than providing explanations rooted at a more basic level. Instead, word frequency constitutes an independent and, more importantly, primary source of data that is supposed to act as the foundation for ratings as a secondary variable (i.e., a function of this input) (see Günther et al., 2022). On a practical level, the most crucial benefit of this methodology is that it can be, in principle, applied to extract estimates of Flickr frequency for any existing word in any language. The procedure is fast, consisting of the automatic extraction of data about images spontaneously uploaded and tagged online by Flickr users worldwide,

and does not require any further in-lab data collection with human participants. This permits effortless investigation of Flickr frequency effects even on large-scale datasets, including mega-studies, with evident benefit in statistical power (Keuleers & Balota, 2015).

An important aspect that needs to be considered is how to treat words never used as tags in Flickr, which parallels previous issues with words that do not occur in a given text corpus (see Brysbaert & Diependaele, 2013). The present study reports results for datasets excluding words not associated with any tag. However, when we include these words in the analyses (assigning frequency = 0 for words never used as tags in Flickr (Brysbaert & Diependaele, 2013; see Supplementary Material), the results do not change. Moreover, the results are also consistent across different large-scale studies (see analyses in Supplementary Materials), thus providing a solid empirical basis supporting the reliability of Flickr frequency as a measure for psycholinguistic studies.

A final consideration that needs to be made is why these norms are effective from a cognitive perspective. Importantly, by stating that Flickr image tags capture "the distribution of word forms in a selective portion of linguistic experience", we do not mean that the linguistic experience of relevance here is the actual experience of assigning or attending image tags on a website like Flickr. Even if, in principle, this is not wrong, it only describes the face value of the present norms. To make a parallel with other traditional frequency norms (Brysbaert et al., 2012; Herdağdelen & Marelli, 2017; van Heuven et al., 2014) (Brysbaert & New, 2009; Herdağdelen & Marelli, 2017; van Heuven et al., 2014), they perform well in predicting behaviors not merely because reading books, watching movies or reading messages on social media are actual pieces of linguistic experience in the speaker of a language. In fact, even word frequencies from Chinese corpora predict reaction times of Spanish speakers who have no experience at all with such corpora (Bates et al., 2003). Instead, word frequency norms perform well in predicting behavior because they capture a language usage that well approximates people's communication (Brysbaert et al., 2012; Brysbaert & New, 2009; Herdağdelen & Marelli, 2017; van Heuven et al., 2014), which is relevant from a cognitive perspective. The same can be said for our norms: tagging images on Flickr makes use, at least to a certain extent, of the same linguistic word forms as referring to visual scenes in everyday life. Flickr has the advantage of isolating such word forms from the most general linguistic experience and making them available. Thus, these word forms can be used as an approximation of the language people use to refer to visual scenes in everyday life and here, we have used them to test whether the latent variable they capture is relevant to how we process words.

In conclusion, the present paper presents Flickr frequency norms. In addition to their theoretical appeal as a word frequency measure extracted from a source corpus inherently linking linguistic and visual experience, Flickr frequency can predict, on an empirical level, behavioral performance across different studies employing different experimental paradigms. We believe this contribution can encourage the usage of similar resources in psycholinguistics and motivate future works in this relevant research field.

Author contribution M.P., F. G., and M.M. participated in study concept and design. M.P. created the script for the extraction of norms used in the study and performed the statistical analyses. M.P., F. G., and M.M. participated in the interpretation of results. M.P. and F.G. drafted the work. M.P., F. G., and M.M. participated in the critical revision of the manuscript. M.M. supervised the project.

Data availability Supplementary Materials associated with this article, the Python script for the extraction of Flickr frequency estimates, the dataset and the analysis script for the analyses reported here are openly available on the Open Science Framework (https://osf.io/2zfs3/).

References

Ames, M., & Naaman, M. (2007). Why we tag: Motivations for annotation in Mobile and online media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/1240624

Anderson, A. J., Bruni, E., Lopopolo, A., Poesio, M., & Baroni, M. (2015). Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage, 120*, 309–322. https://doi.org/10.1016/j.neuroimage.2015.06.093

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International journal of. Psychological Research, 3*(2), 12–28. https://doi.org/10.21500/20112084.807

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology, 30*(11), 1174–1220. https://doi.org/10.1080/02687038.2016.1147767

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). The CELEX Lexical Database (CD-ROM). *Linguistic Data Consortium.*

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. In *Behavior research methods* (Vol. 39, Issue 3, pp. 445–459). Springer . https://doi.org/10.3758/BF03193014

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (pp. 238–247)*Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1023

Baroni, M., & Lenci, A. (2010). Distributional Memory: A general framework for corpus-based Semantics. *Computational Linguistics, 36*(4), 673–721. https://doi.org/10.1162/COLI_A_00016

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., Herron, D., Lu, C. C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., et al. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review 2003 10:2, 10*(2), 344–380. https://doi.org/10.3758/BF03196494

Beaudoin, J. (2007). Folksonomies: Flickr image tagging: Patterns made visible. *Bulletin of the American Society for Information Science and Technology, 34*(1), 26–29. https://doi.org/10.1002/BULT.2007.1720340108

Behrendt, S. (2014). *Lm.Beta: Add standardized regression coefficients to lm-objects*. https://cran.r-project.org/package=lm.beta

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience, 17*(6), 905–917. https://doi.org/10.1162/0898929054021102

Bleasdale, F. A. (1987). Concreteness-dependent associative priming: Separate lexical organization for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 582.

Bolognesi, M. (2016a). Modeling semantic similarity between metaphor terms of visual vs. linguistic metaphors through Flickr tag distributions. *Frontiers Communication, 0*, 9. https://doi.org/10.3389/FCOMM.2016.00009

Bolognesi, M. (2016b). Flickr® Distributional Tagspace: Evaluating the semantic spaces emerging from flickr® Tag distributions. In *Big data in cognitive science* (pp. 153–182). Psychology Press.

Bolognesi, M. (2014). Distributional semantics meets embodied cognition: Flickr® as a database of semantic features. *Selected Papers from the 4th UK Cognitive Linguistics Conference*, 18–35.

Brown, R. W., & Lenneberg, E. H. (1954). A study in language and cognition. *The Journal of Abnormal and Social Psychology, 49*(3), 454.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58*(5), 412–424. https://doi.org/10.1027/1618-3169/a000123

Brysbaert, M., & Diependaele, K. (2013). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods, 45*(2), 422–430. https://doi.org/10.3758/S13428-012-0270-5

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science, 27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods, 51*(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

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. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods, 44*(4), 991–997. https://doi.org/10.3758/S13428-012-0190-4

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*. https://doi.org/10.3758/s13428-013-0403-5

Chan, M. (2020). Rwa: Perform a relative weights analysis. https://cran.r-project.org/package=rwa

Chen, T., Borth, D., Darrell, T., & Chang, S.-F. (2014). Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *ArXiv Preprint* ArXiv:1410.8586.

Chen, X., & Gupta, A. (2015). Webly supervised learning of convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, 1431–1439.

Cohn, N., & Schilperoord, J. (2022). Reimagining language. *Cognitive Science, 46*(7), e13164. https://doi.org/10.1111/COGS.13174

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition, 125*(3), 452–465. https://doi.org/10.1016/j.cognition.2012.07.010

Connell, L., & Lynott, D. (2014). I see/hear what you mean: Semantic activation in visual word recognition depends on perceptual attention. *Journal of Experimental Psychology: General, 143*(2), 527. https://doi.org/10.1037/a0034626

Cox, A. M. (2008). Flickr: A case study of Web2.0. *Aslib proceedings: New information. Perspectives, 60*(5), 493–516. https://doi.org/10.1108/00012530810908210/FULL/PDF

Danguecan, A. N., & Buchanan, L. (2016). Semantic neighborhood effects for abstract versus concrete words. *Frontiers in Psychology, 7*(JUL), 1034. https://doi.org/10.3389/fpsyg.2016.01034

Das, D., & Clark, A. J. (2018). Sarcasm detection on Flickr using a CNN. *Proceedings of the 2018 international conference on computing and big data*, 56–61.

De Groot, A. M. B. (1989). Representational aspects of word Imageability and word frequency as assessed through word association. *Article in Journal of Experimental Psychology Learning Memory and Cognition, 15*(5), 824–845. https://doi.org/10.1037/0278-7393.15.5.824

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference On*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression* (Third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2008). Support for lateralization of the Whorf effect beyond the realm of color discrimination. *Brain and Language, 105*(2), 91–98.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language, 43*(3), 379–401. https://doi.org/10.1006/JMLA.2000.2714

Goldstein, J., & Davidoff, J. (2008). Categorical perception of animal patterns. *British Journal of Psychology, 99*(2), 229–243.

Grömping, U. (2007). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software, 17*(1), 1–27. https://doi.org/10.18637/JSS.V017.I01

Günther, F., Marelli, M., Tureski, S., & Petilli, M. A. (2022). ViSpa (Vision Spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000392

Günther, F., Petilli, M. A., & Marelli, M. (2020a). Semantic transparency is not invisibility: A computational model of perceptually-grounded conceptual combination in word processing. *Journal of Memory and Language, 112*, 104104. https://doi.org/10.1016/j.jml.2020.104104

Günther, F., Petilli, M. A., Vergallito, A., & Marelli, M. (2020b). Images of the unseen: Extrapolating visual representations for abstract and concrete words in a data-driven computational model. *Psychological Research Psychologische Forschung*. https://doi.org/10.1007/s00426-020-01429-7

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372

Hair, J. F., Anderson, R. E., Babin, B. J., & Black, W. C. (2018). *Multivariate data analysis* (pp. 95–120). Pearson. https://doi.org/10.1002/9781119409137.ch4

Heister, J., & Kliegl, R. (2012). Comparing word frequencies from different German text corpora. *Lexical Resources in Psycholinguistic Research, 3*, 27–44.

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. https://doi.org/10.1111/cogs.12392

Hollis, G. (2020). Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language, 114*, 104146. https://doi.org/10.1016/J.JML.2020.104146

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35*(1), 1–19.

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). *Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015)*. https://doi.org/10.1037/a0039248

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31–36.

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist, 86*(1), 65–79. https://doi.org/10.1525/AA.1984.86.1.02A00050

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Taylor & Francis*. https://doi.org/10.1080/17470218.2015.1051065

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. https://doi.org/10.3758/s13428-011-0118-4

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 1097–1105*. https://doi.org/10.1145/3065386

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. In *Computational analysis of present-day American English*: Brown University Press.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211. https://doi.org/10.1037/0033-295x.104.2.211

Lupyan, G. (2008). The conceptual grouping effect: Categories matter (and named categories matter more). *Cognition, 108*(2), 566–577. https://doi.org/10.1016/J.COGNITION.2008.03.009

Lupyan, G. (2012a). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology, 0*(MAR), 54. https://doi.org/10.3389/FPSYG.2012.00054

Lupyan, G. (2012b). What do words do? Toward a theory of language-augmented thought. In *Psychology of learning and motivation* (Vol. 57, pp. 255–297). Elsevier.

Lupyan, G., Rahman, R. A., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in cognitive sciences, 24*(11), 930–944.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods, 52*(3), 1271–1291. https://doi.org/10.3758/s13428-019-01316-z

Mandera, P., Keuleers, E., & Brysbaert, M. (2020). Recognition times for 62 thousand English words: Data from the English crowdsourcing project. *Behavior Research Methods, 52*(2), 741–760. https://doi.org/10.3758/s13428-019-01272-8

Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review, 122*(3), 485–515. https://doi.org/10.1037/a0039267

Menon, S., Berger-Wolf, T. Y., Kiciman, E., Joppa, L., Stewart, C. V. , Crall, P. J., Holmberg, J., & Van Oast, J. (2016). Animal population estimation using Flickr images. *2nd International Workshop on the Social Web for Environmental and Ecological Monitoring (SWEEM 2017), June*, *25*.

Miller, G. A. (1998). WordNet: An electronic lexical database. *MIT press*.

Özgen, E., & Davies, I. R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology: General, 131*(4), 477–493. https://doi.org/10.1037/0096-3445.131.4.477

Petilli, M. A., Günther, F., Vergallito, A., Ciapparelli, M., & Marelli, M. (2021). Data-driven computational models reveal perceptual simulation in word processing. *Journal of Memory and Language, 117*, 104194. https://doi.org/10.1016/j.jml.2020.104194

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: Concrete/abstract decision data for 10,000 English words. *Behavior Research Methods, 49*(2), 407–417. https://doi.org/10.3758/S13428-016-0720-6

Revelle, W. (2021). Psych: Procedures for psychological, psychometric, and personality research. https://cran.r-project.org/package=psych

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. http://www.rstudio.com/

Sahlgren, M. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces* Institutionen för lingvistik.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods, 51*(3), 1258–1270. https://doi.org/10.3758/S13428-018-1099-3

Sigurd, B., Eeg-Olofsson, M., & Van Weijer, J. (2004). Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica, 58*(1), 37–52. https://doi.org/10.1111/J.0039-3193.2004.00109.X

Smith, G. (2007). *Tagging: People-powered metadata for the social web*.

Storopoli, J. (2022). FactorAssumptions: Set of assumptions for factor and principal component analysis. https://cran.r-project.org/package=FactorAssumptions

Stuart, E. (2012). Motivations to upload and tag images vs. tagging practice: an investigation of the Web 2.0 site Flickr (Doctoral dissertation, University of Wolverhampton).

Stuart, E. (2019). Flickr: Organizing and tagging images online. *Knowledge Organization, 46*(3), 223–235. https://doi.org/10.5771/0943-7444-2019-3-223

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Vecchi, E. M., Baroni, M., & Zamparelli, R. (2011). (linear) maps of the impossible: Capturing semantic anomalies in distributional space. https://doi.org/10.5555/2043121.2043122

Vergallito, A., Petilli, M. A., & Marelli, M. (2020). Perceptual modality norms for 1,121 Italian words: A comparison with concreteness

and imageability scores and an analysis of their impact in word processing tasks. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01337-8

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*. https://doi.org/10.3758/s13428-012-0314-x

Weinberger, D. (2007). *Everything is miscellaneous : The power of the new digital disorder*. Times Books.

Westbury, C. (2016). Pay no attention to that man behind the curtain. *The Mental Lexicon, 11*(3), 350–374. https://doi.org/10.1075/ml.11.3.02wes

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences, 104*(19), 7780–7785. https://doi.org/10.1073/PNAS.0701644104

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition, 179*, 213–220. https://doi.org/10.1016/J.COGNITION.2018.05.008

Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(3), 253–265. https://doi.org/10.1002/WCS.104

Zipf, G. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, Mass.: MIT Press.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science, 13*(2), 168–171. https://doi.org/10.1111/1467-9280.00430