AI-generated estimates of familiarity, concreteness, valence and arousal for over 100,000 Spanish words

Gonzalo Martínez [1], Javier Conde [2], Pedro Reviriego[2] and Marc Brysbaert [3]

[1] Universidad Carlos III de Madrid, Spain

[2] ETSI de Telecomunicación, Universidad Politécnica de Madrid, Spain

[3] Department of Experimental Psychology, Ghent University, Belgium

Keywords: word norms, concreteness, valence, arousal, multiword expressions, large language model, GPT-4, Spanish

Accepted for publication in the *Quarterly Journal of Experimental Psychology*

Correspondence address: Marc Brysbaert

Department of Experimental Psychology

Ghent University

9000 Ghent, Belgium

marc.brysbaert@ugent.be

Abstract

This study investigates whether estimates of familiarity, valence, arousal and concreteness based on artificial intelligence (AI) are useful alternatives to word counts and human ratings in Spanish. We replicate and extend previous findings in English and show that GPT-4o is effective in estimating these word features. Validity checks even suggest that AI-generated estimates sometimes outperform traditional measurements. The ability to generate AI estimates for large numbers of words at low cost simplifies the process of obtaining word features and provides a new resource for researchers working in Spanish. We provide Excel lists of the collected word features, which can be freely used for research and teaching.

Language is a cornerstone of human cognition. It consists of a collection of words that can be combined to enable effective communication. Although language is a universal trait, its development and mastery vary among individuals. This variability has made language research a central focus of both behavioral sciences and humanities, to investigate its origins, acquisition, and influence on human thought and culture.

Because words are the building blocks, much language research involves manipulation or control of word features and researchers require good norms for these features.

Important word characteristics in language research

The word features commonly studied or controlled are:

- Word length
- Word frequency
- Word concreteness
- Word valence
- Arousal

Word length is the easiest characteristic to obtain and is usually defined as the length of the written word in number of letters or the length of the spoken word in number of phonemes.

Word frequency is found by counting words in written and (sometimes) spoken language. Not all frequency estimates are equally good because the quality depends on the size of the available sources and the representativeness of the source for the language (Brysbaert & New, 2009). Currently, a corpus usually contains a few hundred thousand words, and movie subtitles seem to be a particularly good source because they contain everyday social interactions.

Concreteness, valence and arousal are semantic variables, related to the meaning of the word rather than the form. Concreteness measures the extent to which the word refers to concepts that can be experienced through interactions with the physical environment. Valence refers to the affect evoked by the word and ranges from very negative (e.g., violation) to very positive (heaven). Arousal refers to the degree of excitement evoked by the word and ranges from very low (e.g., tranquility) to very high (murder).

Estimates of concreteness, valence and arousal are usually obtained by asking people to rate words on a Likert scale. This requires considerable investment, which means that for most languages norms are limited to a few hundred or a few thousand words.

Are estimates generated by artificial intelligence a new source of information?

Research in recent months suggests that artificial intelligence may be an interesting new source of information about word features. Large language models (LLMs) like the ones used in ChatGPT have been trained on trillions of words and are easy to interact with. Trott (2024a) queried GPT-4 with the instructions typically given to participants and correlated them with human ratings. He obtained correlations of .81 for concreteness, .76 for valence, and .66 for arousal. In a subsequent paper, Trott

(2024b) showed that the GPT-4 estimates were as good as the average of three participants for concreteness and 11 participants for valence.

Very similar results were reported by Martínez et al. (2024) for much larger samples of English words. They obtained a correlation of .89 between GPT-4 estimates of concreteness and human ratings for a sample of 34,246 words, a correlation of .90 for valence (N = 13,914), and a correlation of .74 for arousal (N = 13,914 as well). In addition, they showed that the GPT-4 estimates also work for multiword expressions (hand wash, hand over, hand in hand, …).

Brysbaert et al. (2024) extended the investigation to word familiarity. They found a correlation of .80 between GPT-4 estimates of word familiarity and human ratings (N = 4515). Familiarity estimates also correlated .67 with word frequency based on subtitles. Further analyses showed that the GPT-4 familiarity estimates better predicted response accuracy in lexical decision experiments and were as good as word frequency norms in predicting reaction times when word length was taken into account.

The present studies investigate to what extent the English findings generalize to Spanish. Large language models like GPT-4 were mainly trained on English (Li et al., 2024). As a result, their performance is likely to be better in English than in other languages. Indeed, Martínez et al. (2023b) observed that LLMs performed better on demanding vocabulary tests in English than in Spanish, with GPT-4 among the best performers in both languages (100% correct in English, 95-99% correct in Spanish). So, the question is how useful GPT4 is to estimate word features in Spanish. We start with word familiarity.

Study 1: Word familiarity

In Study 1, we replicated the research of Brysbaert et al. (2024), who queried GPT-4 about the familiarity of words. They examined word familiarity rather than word frequency because the familiarity instruction ("How familiar is the word __?") is used more often in human rating studies than the frequency instruction ("How often have you encountered the word __?"). It also is less rater oriented. To translate the frequency instruction for a large language model, one would have to change it into "How often does an average person encounter the word __?", which is less direct than the familiarity instruction.

An advantage of large language models is that it is easy to interact with them: Questions can be asked in much the same way as with humans and the same answers can be obtained. Questions and answers can be automatized, so that a large number can be collected in a short period of time (Martínez et al., 2023). We only present data from ChatGPT-4o (version gpt-4o-2024-08-06), as this provided the best output (also in Brysbaert et al., 2024, and Martínez et al., 2023b, 2024).

We used a Spanish translation of the English prompt used by Brysbaert et al. (2024) and added the phrase "in Spanish" (en español), to focus the model on Spanish. The prompt was repeated before each word, to prevent instruction dilution. The specific prompt we used was:

> "La familiaridad es una medida de lo familiar que es algo. Una palabra es muy FAMILIAR si la ves/escuchas a menudo y es fácilmente reconocible. Por el contrario, una palabra es muy POCO FAMILIAR si la ves/escuchas rara vez y es relativamente irreconocible. Por favor, indica como de familiar crees que es cada palabra en español en una escala del 1 (MUY POCO FAMILIAR) al 7

(MUY FAMILIAR), donde el punto medio representa una familiaridad moderada. La palabra es: [insertar palabra aquí]. Responde solo un número del 1 al 7. Limita tu respuesta a números."[1]

For all estimates, we set the temperature of the model to 0, so that the same results are obtained each time a word is presented to the model (The temperature parameter in LLMs affects the variability and randomness of generated responses; a temperature of 0 always produces the same, most likely response).

The parameters of GPT-4 can be set in such a way that it gives the estimated probability of the most likely answers (by selecting "log_probs"). We asked for the probabilities of the 5 most likely answers. For the word "cantina" [canteen], this could for instance be a probability of .52 (rounded off) for rating 4 (moderate familiarity), .44 for rating 3, .02 for rating 5, and .01 for ratings 6 and 2. This allowed us to calculate two values: the rating with the highest probability (i.e., 4) and a more precise value found by multiplying the ratings with their probabilities. So, for "cantina" this would give a value of .52*4 + .44*3 + .02*5 + .01*6 + .01*2 = 3.54. The latter estimate gave slightly better results, and we will use it in all analyses reported below.

The first test we used to evaluate the quality of the GPT-4o familiarity estimates, was to correlate them with human familiarity estimates. If they measure the same feature, they should correlate strongly.

There are six useful datasets of human word familiarity estimates in Spanish:

1. Desrochers et al. (2010): provides familiarity ratings for 330 nouns on a 7-point Likert scale (1 = very unfamiliar, 7= very familiar). Participants were Spanish university students.
2. EsPal (Duchon et al., 2013): provides familiarity ratings for 6326 words on a 7-point Likert scale. Participants were Spanish university students.
3. Moreno-Martínez et al. (2014): provides familiarity ratings for 820 words on a 5-point Likert scale (1 = very unfamiliar, 5= very familiar). Participants were Spanish university students.
4. Guasch et al. (2016): provides familiarity ratings for 1400 words on a 7-point Likert scale. Participants were Spanish university students.
5. EmoFinder (Fraga et al., 2018): provides familiarity ratings for 380 words on a 7-point Likert scale. Participants were Spanish university students.
6. Sarli & Justel (2022): provides familiarity ratings for 1034 words on a 9-point Likert scale. Participants were a combination of university students and other adults from Argentina.

Table 1 shows the Pearson correlations between the different familiarity ratings and the GPT estimates (Spearman correlations were very similar). Because the numbers of overlapping words are very different, they have been added to the table.

[1] "Familiarity is a measure of how familiar something is. A word is very FAMILIAR if you see/hear it often and it is easily recognizable. In contrast, a word is very UNFAMILIAR if you rarely see/hear it and it is relatively unrecognizable. Please indicate how familiar you think each word in Spanish is on a scale from 1 (VERY UNFAMILIAR) to 7 (VERY FAMILIAR), with the midpoint representing moderate familiarity. The word is: [insert word here]. Only answer a number from 1 to 7. Please limit your answer to numbers."

	EsPal fam N = 6326	Moreno N = 820	Guasch N = 1400	EmoFinder N = 380	Sarli N = 1034	GPT N = 7725
Desrochers N = 330	**.73** N = 141	**.78** N = 32	**.72** N = 635	**.38** N = 6	**.41** N = 2540	**.88** N = 330
EsPal		**.65** N = 333	**.69** N = 1099	**.66** N = 198	**.61** N = 601	**.68** N = 6326
Moreno			**.72** N = 140	**.78** N = 105	**.69** N = 79	**.77** N = 820
Guasch				**.85** N = 68	**.71** N = 296	**.75** N = 1400
EmoFinder					**.45** N = 5	**.67** N = 380
Sarli						**.69** N = 1034

Table 1: Pearson correlations between the various human familiarity ratings and the GPT-4o estimates. N indicates the number of data pairs available for the correlation.

Table 1 shows two things. First, the available human familiarity ratings are quite fragmented, with partially overlapping stimulus sets and mediocre correlations between studies (around r = .7). Second, the GPT-4o estimates compare well to the human ratings and are available for all stimuli used in the various studies. Correlations of .8 between GPT estimates and human ratings are in line with the English findings of Martínez et al. (2024).

The second validation test we used was to see whether the GPT-4o familiarity estimates correlated with other word features in the same way as human familiarity ratings do. The familiarity ratings of the different human rating studies were standardized on a scale of 1-7 and averaged for overlapping words. This gave a total of 7725 words with human ratings, which were correlated with the GPT-4o familiarity estimates, word length, and word frequency. Two measures were used for the latter: (1) EsPal written word frequency (Duchon et al., 2013), and (2) the SUBTLEX-ESP subtitle frequencies of Cuetos et al. (2012). The EsPal written frequencies are currently the measure of choice for research in the Spanish language. Brysbaert and New (2009) argued that subtitle-based word frequencies are better predictors of word processing than frequency norms based on written material (fiction and nonfiction). Words with frequencies not found in the EsPal or SUBTLEX databases were considered missing observations. Table 2 shows the results.

	EsPal freq N = 7679	SUBTLEX N = 7264	FAM_rating N = 7725	GPT N = 7725
Length N = 7725	-.19	-.32	-.13	-.16
EsPal		.70	.47	.68
SUBTLEX			.52	.74
FAM_rating				.72

Table 2: Pearson correlations between the GPT-4o familiarity estimates and other word characteristics for the total set of words included in the human rating studies (N = 7725).

Table 2 shows that the GPT-4 familiarity estimates correlate more with the word frequency norms than the human familiarity ratings. All familiarity estimates correlate less with word length than the frequency measures (in particular the SUBTLEX frequencies). Similar findings were observed by Martínez et al. (2024) for English. Table 2 also shows the rather low correlation (r = .70) between the EsPal written frequency norms and the SUBTLEX frequency norms for the words included in the analysis.

A third criterion we can use to validate the GPT-4o familiarity estimates is to see how much variance they explain in measures of word recognition efficiency. There are five datasets we can use:

1. González-Nosti et al. (2014): Lexical decision times (no accuracy) for 2765 words obtained from Spanish university students.
2. SPALEX (Aguasvivas et al., 2018, 2020): Lexical decision data (accuracy and RT) for 44,853 words obtained from Spanish people (crowdsourcing study).
3. Haro et al. (2024): Lexical decision data (accuracy and RT) for 7492 words obtained from Spanish university students.
4. Davies et al. (2013): Naming times for 2764 words, obtained from Spanish university students (accuracy was at ceiling level).
5. Miguel-Abella et al. (2022): Naming times for 4562 verbs, obtained from Spanish university students (accuracy was at ceiling level).

Table 3 shows the correlations of the dependent variables with word frequency and GPT familiarity estimates. An extra word frequency variable added to the table is Multilex. This combines the subtitle word frequencies of van Paridon and Thompson (2021; 513 million words) with the WorldLex word frequencies (blogs, tweets and news: 31.6 + 29.6 + 16.0 million words) of Gimenes and New (2016). Word frequencies are expressed in Zipf scores (Van Heuven et al., 2014). Research in English and Dutch has shown that the combined Multilex frequency measure is better than others available.

	EsPal_freq	SUBTLEX_freq	Multilex_freq	GPT_fam
Gonzalez LDT speed (RT)	-.54 N = 2765	-.63 N = 2752	**-.65** N = 2765	**-.65** N = 2765
SPALEX LDT accuracy	.60 N = 44,853	.34 N = 19,284	.60 N = 36,771	**.76** N = 44,853
SPALEX LDT speed (RT)	-.69 N = 44,850	-.51 N = 19,284	-.72 N = 36,769	**-.76** N = 44,850
Haro LDT accuracy	.31 N = 7492	.34 N = 7343	.39 N = 7492	**.57** N = 7492
Haro LDT speed (RT)	-.50 N = 7492	-.59 N = 7343	-.62 N = 7492	**-.69** N = 7492
Davies nam speed (RT)	-.34 N = 2764	-.40 N = 2751	-.**42** N = 2764	-.37 N = 2764
Miguel nam speed (RT)	-.42 N = 4562	-.37 N = 3100	**-.44** N = 4380	**-**.43 N = 4562

Table 3: Pearson correlation coefficients between word recognition efficiency (accuracy and speed) and word frequency/familiarity measures. The highest correlation for each study is in bold.

As in English, GPT-4o familiarity estimates are much better than word frequencies for predicting which words are known in lexical decision experiments and which are not. For processing speed, familiarity estimates do not seem to have an advantage over the best frequency measure (Multilex). However, because GPT-4o familiarity estimates correlate less with word length than Multilex frequency norms, there is a small but consistent advantage for GPT-4o familiarity estimates when both variables are included in the analysis, particularly when some nonlinearity of the predictors is allowed for by using restricted cubic splines with 4 nodes (see Table 4).

	Multilex	Multilex + Length	GPT_Fam	GPT_Fam + Length
Gonzalez LDT speed (RT)	.43	.50	.45	**.55**
SPALEX LDT accuracy	.40	.52	.65	**.67**
SPALEX LDT speed (RT)	.54	.56	.58	**.61**
Haro LDT accuracy	.20	.26	.39	**.42**
Haro LDT speed (RT)	.41	.43	.48	**.52**
Davies nam speed (RT)	.18	**.32**	.15	**.32**
Miguel nam speed (RT)	.19	.32	.17	**.34**

Table 4: Percentage of variance explained in multiple regression with Multilex frequency or GPT-4o familiarity when word length (number of letters) is included as an extra predictor. Small deviations

from linearity were allowed by using restricted cubic splines with 4 knots (Harrell, 2024). To compare the numbers of Table 4 to those of Table 3, take the square root of the number in Table 4.

Figure 1 shows the effects of word length and GPT familiarity estimate on the lexical decision times of SPALEX.

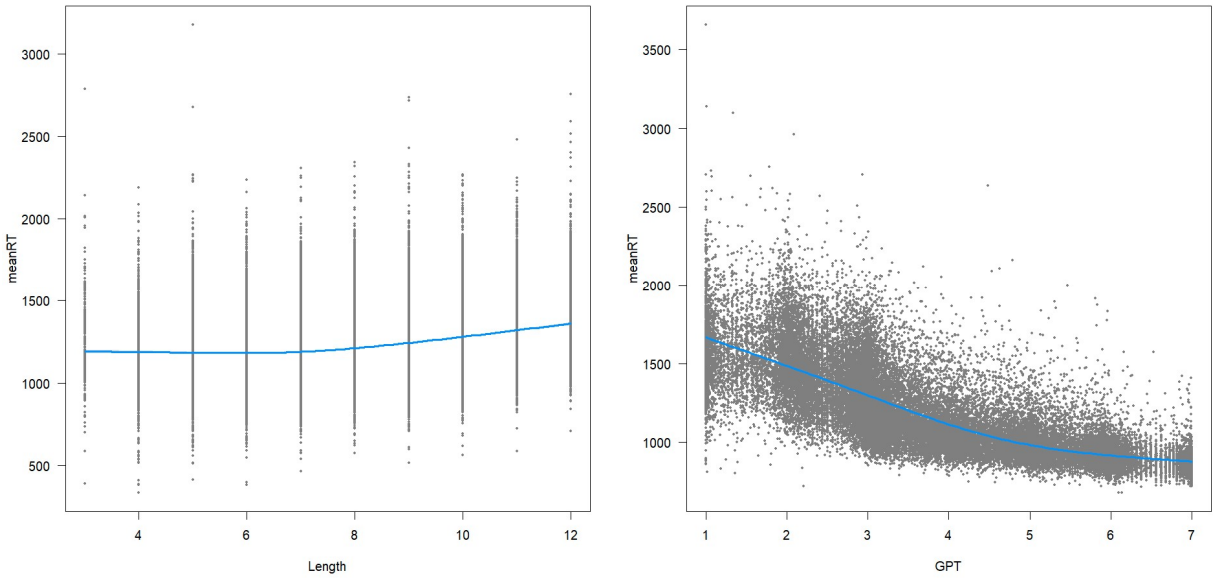

Figure 1: Effects of word length and GPT-4ofamiliarity estimates on mean RTs in the SPALEX database (N = 36,769). Figures made with the visreg package in R (Breheny & Burchett, 2020).

Having found that GPT-4o familiarity estimates work well in Spanish, we felt confident that valence, arousal and concreteness would yield the same results as in English. These features relate more to the meaning of the words than to their form and there is a general consistency between English and Spanish in terms of which words are positive/negative, arousing, or concrete/abstract (Thompson & Lupyan, 2019).

Study 2: Valence

Valence refers to the affect evoked by words. There are seven reasonably large databases of human valence ratings in Spanish. They are:

1. Redondo et al. (2007): includes valence ratings for 1034 words, including a large number of positive and negative words. Spanish students.
2. Guasch et al. (2016): contains valence ratings for 1400 words, obtained from Spanish students.
3. Hinojosa et al. (2016): has valence ratings for 875 words including a large number of positive and negative words. Spanish students.

4. Stadthagen-Gonzalez et al. (2017): has valence ratings for 14,031 words selected from an entire vocabulary list, collected from Spanish students.
5. Sabater et al. (2023): has valence ratings for 1,406 words, collected from Spanish children and adolescents (7, 9, 11, and 13 years old).
6. Sarli & Justel (2022): valence ratings for 1034 words including a large number of positive and negative words. Students from Argentina.
7. Marín-Gutiérrez & Correales (2023): valence ratings for 1917 words, collected from students in Colombia.

To stay in line with the human ratings, a 9-point Likert scale was used for the GPT-4o valence estimates. This was the prompt we used (see Martínez et al., 2024, for a similar prompt in English):

> "¿Podrías por favor calificar cómo la siguiente expresión hace sentir a una persona? Usa una escala del 1 al 9, donde 1 significa muy negativo, malo y 9 significa muy positivo, bueno. Ejemplos de palabras que obtendrían una calificación de 1 son entierro, fusilar y violación. Ejemplos de palabras que obtendrían una calificación de 9 son diversión, sonriente y libre. La expresión es: [insertar expresión aquí]. Responde solo con un número del 1 al 9. Por favor limita tu respuesta a números."

Table 5 shows the Pearson correlations between the different measurements. They show that GPT-4o's estimates score well compared to the other sources. The correlations range from .73 to .95. The correlations with children and students from Colombia are the lowest, indicating that the GPT-4o estimates for the prompt we used align most with adults from Spain. Further research will have to indicate whether it is possible to obtain estimates for children (and other groups of interest) by changing the instructions or by fine-tuning the model to the language used by the group (see also Schepens et al., 2023).

	Guasch	Hinojosa	Stadth	Saba_7	Saba_9	Saba_11	Saba_13	Sarli	Marin	GPT-4o
Redondo N = 1034	.97 (N= 326)	.99 (N = 18)	.98 (N=1033)	.84 (N= 395)	.89 (N= 395)	.92 (N= 395)	.93 (N= 395)	.95 (N=844)	.93 (N= 324)	.92 (N=1033)
Guasch N=1400		.97 (N=134)	.95 (N=1299)	.77 (N=359)	.83 (N=359)	.87 (N=359)	.88 (N=359)	.94 (N=294)	.90 (N=539)	.90 (N=1400)
Hinojosa N = 875			.97 (N= 640)	.82 (N= 196)	.88 (N= 196)	.92 (N= 196)	.93 (N= 196)	.97 (N=45)	.93 (N= 189)	.95 (N=875)
Stadt-G N=14,031				.79 (N= 1373)	.84 (N= 1373)	.88 (N= 1373)	.89 (N= 1373)	.95 (N=960)	.87 (N=1858)	.83 (N=14027)
Saba_7 N= 1406					.88 (N= 1406)	.87 (N= 1406)	.82 (N= 1406)	.83 (N=364)	.78 (N= 512)	.73 (N= 1406)
Saba_9 N= 1406						.91	.88 (N= 1406)	.88 (N=364)	.82 (N= 512)	.77 (N= 1406)
Saba_11 N= 1406							.91 (N= 1406)	.90 (N=364)	.85 (N= 512)	.80 (N= 1406)
Saba_13 N= 1406								.91 (N=364)	.88 (N= 512)	.81 (N= 1406)
Sarli N= 1034									.95 (N=310)	.91 (N=1021)
Marin N=1917										.81 (N=1915)

Table 5: Pearson correlations between the different human rating studies and the GPT-4o estimates of valence. Numbers between brackets are the numbers of data pairs on which the correlations are based.

Study 3: Arousal

Arousal refers to the degree of excitement evoked by a word. All databases with valence ratings also included arousal ratings, except for Marín-Gutiérrez and Correales (2023).

The following prompt was used to query GPT-4o for arousal estimates (see Martínez et al., 2024, for the same prompt in English):

> "¿Podrías calificar cómo hace sentir a una persona la lectura de la siguiente expresión? Usa una escala del 1 al 9, donde 1 significa muy calmado, relajado, y 9 significa muy excitado, energizado. Ejemplos de palabras que recibirían una calificación de 1 son siesta, relajante y suave. Ejemplos de palabras que recibirían una calificación de 9 son asesinar, euforia y violar. La expresión es: [insertar expresión aquí]. Solo responde con un número del 1 al 9. Por favor, limita tu respuesta a números."

Table 6 shows how the GPT estimates correlate with the human ratings. These correlations are lower than the correlations for the valence ratings, also between the human studies. A similar observation was made in English (Martínez et al., 2024). In particular, children's ratings differ greatly from those of adults, except at the extreme ends of very high and low arousal words (see the larger correlations with the selected stimuli from Redondo et al. and Hinojosa et al.) The GPT-4o arousal estimates for the prompt we used resemble those of adults, not children.

	Guasch	Hinojosa	Stadt-G	Saba_7	Saba_9	Saba_11	Saba_13	Sarli	GPT-4o
Redondo N = 1034	.84 (N=326)	.74 (N = 18)	.75 (N=1033)	.08 (N= 395)	.26 (N= 395)	.40 (N= 395)	.51 (N= 395)	.80 (N=844)	.76 (N=1033)
Guasch N=1400		.78 (N=134)	.87 (N=1299)	.03 (N=359)	.29 (N=359)	.49 (N=359)	.58 (N=359)	.80 (N=294)	.84 (N=1400)
Hinojosa N = 875			.71 (N= 640)	.30 (N= 196)	.50 (N= 196)	.60 (N= 196)	.64 (N= 196)	.78 (N=45)	.77 (N=875)
Stadt-G N=14,031				-.14 (N= 1374)	.05 (N= 1374)	.18 (N= 1374)	.24 (N= 1374)	.77 (N=960)	.76 (N= 14027)
Saba_7 N = 1406					.64 (N= 1406)	.56 (N= 1406)	.53 (N= 1406)	-.05 (N=364)	-.04 (N= 1406)
Saba_9 N = 1406						.70	.65 (N= 1406)	.11 (N=364)	.18 (N= 1406)
Saba_11 N = 1406							.76 (N= 1406)	.26 (N=364)	.32 (N= 1406)
Saba_13 N = 1406								.33 (N=364)	.37 (N= 1406)
Sarli N= 1034									.76 (N= 1021)

Table 6: Pearson correlations between the different human rating studies and the GPT-4o estimates of arousal. Numbers between brackets are the numbers of data pairs on which the correlations are based.

Study 4: Concreteness

Concreteness measures the extent to which the word refers to concepts that can be experienced through interactions with the physical environment. For historical reasons, most concreteness ratings are based on a 7-point Likert scale.

There are five databases of human ratings that we can use in Spanish:

1. EsPal (Duchon et al., 2013): includes concreteness ratings for 6372 words, obtained from Spanish students.
2. Guasch et al. (2016): contains concreteness ratings for 1400 words, obtained from Spanish students.
3. Hinojosa et al. (2016): includes concreteness ratings for 875 words that include a large number of positive and negative words. Spanish students.
4. Sarli & Justel (2022): concreteness ratings for 1034 words that include a large number of positive and negative words. Students from Argentina.
5. Thompson & Lupyan (2019) estimated concreteness for Spanish words on the basis of translations of a core set of English words and semantic vectors to generalize the ratings to other words. In English, these ratings correlated .8 with human ratings.

We asked GPT-4o for concreteness information with the following prompt (see Martínez et al., 2024, for the English translation):

> "¿Podrías calificar la concreción de la siguiente expresión en una escala del 1 al 7, donde 1 significa muy abstracto y 7 significa muy concreto? Ejemplos de palabras que recibirían una calificación de 1 son espiritualidad, convencionalismo y creencia. Ejemplos de palabras que recibirían una calificación de 7 son trineo, margarita y pavo real. La expresión es: [insertar expresión aquí]. Solo responde con un número del 1 al 7. Por favor, limita tu respuesta a números."

Table 7 shows the results. They indicate that the GPT-4o estimates are substantially better than the Thompson and Lupyan (2019) estimates. At the same time, the correlations of the GPT-4o estimates with human ratings tend to be lower than the correlations between the four rating studies and lower than in English (where a correlation of r = .89 was found for word stimuli; Martínez et al., 2024, footnote 8). Interestingly, the correlation is highest with ratings obtained in Argentina (Sarli & Justel, 2022).

	Guasch N = 1400	Hinojosa N = 875	Sarli N = 1034	Thompson N = 985,664	GPT N = 127,728
EsPal ratings (N = 6372)	.88 (N = 1103)	.78 (N = 405)	.75 (N = 604)	.67 (N = 6346)	.74 (N = 6346)
Guasch		.82 (N = 134)	.77 (N = 294)	.71 (N = 1400)	.81 (N = 1400)
Hinojosa			.71 (N = 45)	.52 (N = 875)	.63 (N = 875)
Sarli				.82 (N = 1021)	.87 (N = 1021)
Thompson					.65 (N = 75,215)

Table 7: Pearson correlations between the human rating studies, a previous AI-based estimate based on translation, and the GPT-4o estimates of concreteness. Numbers in parentheses are the numbers of data pairs on which the correlations are based.

To see whether the results could be improved by adding extra instructions, we added the instructions of Spreen and Schulz (1966; also used by Paivio et al., 1968, and most later studies) to the prompt. This did not improve the validity of the estimates (there was even a negative trend). The same was true for another more elaborative prompt we tried out.

A look at the deviations between the GPT-4o estimates and the human ratings suggests that the less than optimal correlation is caused by unexpected values both in the human ratings and in the GPT-4o estimates. Words with human ratings below 3 (abstract) and GPT-4o estimates above 5.5 (concrete) are: negación, satanás, poesía, psicoanálisis, fecundidad, agresividad, interesarse, manifestación, afirmación, probabilidad, islam, machista, derecho, administrative, narración, bautismo. Words with human ratings above 5.5 (concrete) and GPT-4o estimates below 3 (abstract) are: puesto, hacer, artilugio, deje, deshacer, cogida, vez, vale, alrededores, tinglado, pedazo, droguero.

Further Research will have to indicate whether the GPT-4o estimates of concreteness are the best possible or whether some further improvements can be made.

Study 5: Familiarity ratings of Spanish-English cognates and false friends

A final question we asked was to what extent the GPT-4o familiarity estimates for Spanish are affected by the fact that the model was trained primarily in English with Spanish accounting only for a small fraction of the training materials.[2]

Language contamination is most likely for words that have similar orthography. There are two types of such words: Cognates and false friends. Cognates are words with similar form and meaning in Spanish and English. Examples are banal, pasta and manual. False friends have a similar form in both languages but different meanings. An example is “injuria,” which means insult in Spanish (not injury), often causing translation errors in less proficient second language users.

[2] The authors thank the action editor for alerting them to this issue.

If the Spanish estimates are influenced by the fact that the words also exist in English, we expect that the familiarity estimates will be higher for Spanish-English cognates than for matched control words, which have little orthographic overlap with their English translation. Expectations are less clear for false friends. Because the form exists in English, one could predict higher familiarity estimates. On the other hand, because the meanings of the words are different in Spanish and English and cause inconsistencies in the mappings, one could also expect lower familiarity ratings for false friends than for matched control words.

We started from a list of identical Spanish-English cognates collected by Barr (2015) and a list of false friends taken from Wiktionary (https://en.wiktionary.org/wiki/Appendix:False_friends_between_English_and_Spanish, retrieved on November 10,2024). For each cognate and false friend, we searched for a control word that was matched in length, Multilex frequency, and part of speech. If the target word was an adjective with different masculine and feminine forms (e.g., sano vs. sana [healthy]), we used the masculine form and searched for a matching adjective of the same type. If the target word was an adjective making no distinction between masculine and feminine (focal, beige, fútil, suave), we searched for control words of the same type (foral, cutre, sagaz, dulce).[3]

We were able to compile a list of 201 pairs of cognates and controls (68 adjectives, 1 adverb, 132 nouns), and a list of 217 pairs of false friends and controls (32 adjectives, 2 prepositions, 4 adverbs, 1 auxiliary verb, 137 nouns, 2 number words, 39 verbs). We added familiarity estimates to each list (which can be downloaded from the osf website).

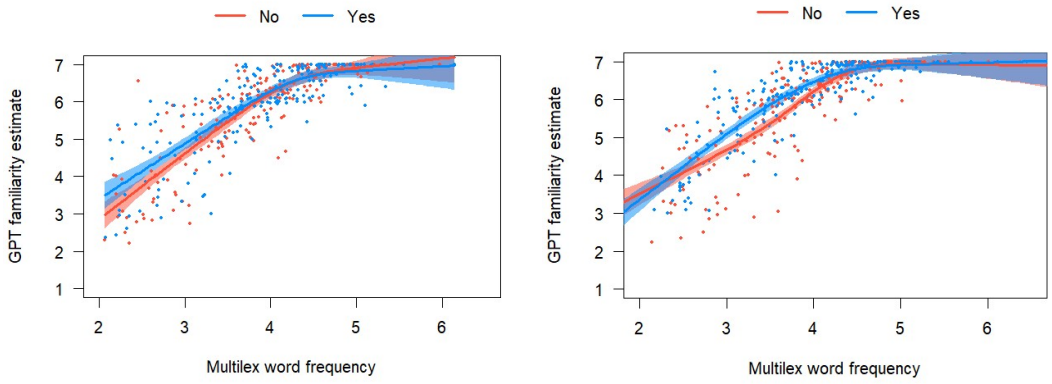

Figure 2: GPT-4o familiarity estimates for words with a large orthographic overlap between Spanish and English (Yes) and matched controls with little orthographic overlap (No). Left part: cognates in Spanish and English (N = 201 data pairs). Right pate: false friends between Spanish and English (N =

[3] Another option would have been to use lemma frequency: the summed frequencies of all inflected forms (feminine singular and plural plus masculine singular and plural). Lemma frequencies usually do not correlate more with lexical decision times than word form frequencies (Brysbaert & New, 2009; Gimenes et al., 2016), which is the reason why they are rarely used. AI-based familiarity estimates may be a better way to combine frequencies of inflected forms because networks are sensitive to correlations between form and meaning without giving them equal weight (as happens in the calculation of lemma frequency).

217 data pairs). As shown previously, familiarity estimates increase with word frequency. Words closely resembling English words got slightly higher GPT familiarity estimates.

Figure 2 shows the results. Contrary to expectations, we found no significant difference in familiarity estimates between cognates and controls ($t < 1$, $p > .5$). The difference between false friends and controls was significant ($t = 4$, $p < .001$) with slightly higher familiarity estimates for false friends than for controls, restricted to words of average frequency.

All in all, it looks like Spanish words with close forms in English get slightly higher familiarity estimates than words with dissimilar translations. However, the difference is minor relative to the full range of familiarity estimates and may also be observed in Spanish speakers with knowledge of English. The small difference is in line with the good performance of Spanish estimates in predicting lexical decision and naming performance in Spanish.

Discussion

Recent studies have shown that GPT-4 provides remarkably good estimates of word familiarity, valence, arousal and concreteness in English (Brysbaert et al., 2024; Martínez et al., 2024; Trott, 2024a, 2024b). The present article repeated the findings in Spanish, with comparable results. GPT-4o estimates of word familiarity, valence, arousal and concreteness turned out to be good proxies for existing measures based on word counting and human judgment. To predict which words will be known by participants, AI-generated familiarity estimates even did better than the best word frequency measure (Table 3).

To some extent, the good performance of a high-performing LLM is not surprising. The requested information is present in the language (rather than outside the language, as in moral reasoning or embodied cognition; Kosaka & Kikkawa, 2024) and the model has been exposed to more language than all participants in human rating studies or the number of words included in corpus analysis. Nevertheless, the results, especially with respect to the familiarity estimates, are surprisingly good given the concerns one may have about the representativeness of the language to which the model has been exposed and possible contamination between languages.

AI-generated estimates are interesting because they are easy to obtain for all words of interest. All you have to do is submit the stimuli to a good language model and use an adequate prompt (current prompts can be used as examples). Another advantage is that the estimates are not limited to words, but are also useful for expressions consisting of multiple words and for ambiguous words that are disambiguated one way or the other (Brysbaert et al., 2024; Martínez et al., 2024; Trott, 2024b). We report the results with ChatGPT-4o because they are currently the best, but other models are likely to catch up soon, including open models that allow more control of the input given (Hussain et al., 2024).

The need for easy feature generation becomes clear when we look at the available human ratings, which rarely cover more than a few hundred or thousand words, as we have seen in the validation studies. In addition, there are many more interesting features than the ones covered now. For example, Connell and Lynott (2012) argued against the use of concreteness norms in language research because such ratings do not take into account the full extent of human sensations and actions. Far better to collect modality-specific ratings by asking “To what extent do you experience [the target word] by seeing?”, “By feeling through touch,” “By hearing,” “By sensations inside your body,” “By smelling,” and “By tasting”.

Similarly, the action strengths associated with words can be normed by asking “To what extent do you experience [the target words] by performing an action with the, “Foot/leg,” “Hand/arm,” “Head excluding mouth,” “Mouth/throat,” or “Torso?” Such detailed sensorimotor norms (Lynott et al., 2020) allow more precise estimates of sensation strength and action strength, which make more refined studies possible than simple concreteness ratings (Khanna & Cortese, 2021; van Hoef et al., 2023). However, this research requires norming on 11 different dimensions.

A similar evolution is taking place regarding valence. Rather than limiting affect to a single dimension, more information is likely to be obtained from more sophisticated measurements such as the degree of happiness, anger, sadness, surprise and fear evoked by the words (Pérez-Sánchez et al., 2021). Positive emotions can be further divided into awe, satisfaction, amusement, excitement, serenity, relief and pleasure (Hinojosa et al., 2024). Again, such detailed research requires the collection of many norms (ideally in different languages).

At the same time, it is important not to glorify AI-generated estimates. These estimates are numbers from a neural network that was primarily exposed to the language most readily available for training (Atari et al., 2024; Dillion et al., 2023; Grieve et al., 2024). This may become a bigger problem in the future if more and more information is generated by LLMs, as we will run the risk of LLMs providing information about the language produced by LLMs.

It is better to use GPT estimates in conjunction with human data, especially when the relative importance of variables is contrasted. Ideally, the human data are collected again, because it cannot be excluded that old data were part of the training material given to the LLM (Trott, 2024a, 2024b). The availability of AI estimates is particularly interesting for variables that need to be controlled. It also helps researchers limit human ratings to words that are likely to be of interest.

Availability

To help researchers in Spanish, we make the GPT-4o estimates available at https://osf.io/frc6a/. There are two files.

The first file is an Excel file with familiarity information for 773,764 words. It contains both the dominant rating given by GPT-4o and the more detailed estimates based on the probabilities of the possible answers. It also contains the source of the word. The following words were included:

1. The words used in Spanish lexical decision tasks: N = 45,263
2. Extra words used in Spanish word naming tasks: N = 979
3. Extra words used in Spanish rating studies: N = 8,014
4. Extra words listed in the Diccionario de la lengua Española de la Real Academia Española: N = 47,059
5. Extra words listed in the Diccionario del español actual (Seco et al., 1999): N = 23,539
6. Extra words from Multilex: N = 648,910.

This selection ensures that most words of interest are covered by the list. It includes all words used in research to date, all lemmas commonly found in dictionaries, and all names and inflected forms observed at least 3 times in the Multilex corpus.

Although the number of words is large, it is unlikely that all words of interest are available. The main reasons for this are that Spanish has many more inflected forms than English and that compound words are more often written as single words (of the type farmhouse, snowman) than as word sequences (like park bench, snow shovel). Given the productivity of word compounding, the number of possible compounds can easily be in the tens of millions. For these words, it is still possible to obtain estimates by querying the GPT-4o model with the prompts we used (or by querying other, equivalent models that become available).

The second file contains valence, arousal, and concreteness estimates for 127,728 Spanish words. These include the first five categories of the familiarity list, plus some Multilex words with high frequencies. They provide researchers with a large list of words to select from (much larger than presently available). Again, estimates for missing words can be queried directly.

The osf website also includes the Multilex frequency list. It contains frequency values for 714,965 Spanish words. Frequencies are given as Zipf values, which is the best standardized measure of word frequency (van Heuven et al., 2014). It goes from 0 to over 7, with 0 representing one occurrence per billion words, and 7 representing 10,000 occurrences per million words. Low-frequency words are words with a Zipf-value below 3 (1 per million words); high frequency words are words with a Zipf-value above 4 (10 per million words). All words observed at least 3 times in the corpus are included; for entries with 1 or 2 occurrences, only those that passed the MS Office spell checker are included. As shown in Table 3, Multilex frequencies are better than the EsPal and SUBTLEX frequencies currently used.

Finally, the osf website includes the stimuli and R code used in the validation studies. This allows readers to verify the claims we made and – possibly – take issue with some of the conclusions we reached.

Acknowledgements

This research was supported by the FUN4DATE (PID2022-136684OB-C21/C22) and ENTRUDIT (TED2021-130118B-I00) projects funded by the Spanish Agencia Estatal de Investigacion (AEI) 10.13039/501100011033 and by the OpenAI research access program, which provided access to ChatGPT-4o on a non-commercial basis.

References

Aguasvivas, J. A., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A. (2018). SPALEX: A Spanish lexical decision database from a massive online data collection. Frontiers in Psychology, 9, 2156.

Aguasvivas, J., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A. (2020). How do Spanish speakers read words? Insights from a crowdsourced lexical decision megastudy. Behavior Research Methods, 52, 1867-1882.

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2024, June 20). *Which humans?* OSF. https://osf.io/preprints/psyarxiv/5b26t

Barr, A. (2015). *1001 Spanish Words You Already Know – A Guide To English-Spanish Cognates*. REALFASTSPANISH. https://www.realfastspanish.com/vocabulary/spanish-cognates.

Breheny, P., & Burchett, W. (2020). *Package 'visreg' Version 2.7.0.* http://r.meteo.uni.wroc.pl/web/packages/visreg/visreg.pdf.

Brysbaert, M., Martínez, G., & Reviriego, P. (2024, November 19). Moving beyond word frequency based on tally counting: AI-generated familiarity estimates of words and phrases are an interesting additional index of language knowledge. RESEARCHGATE. https://www.researchgate.net/publication/385939692

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.

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

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2012). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. Psicológica, 33(2), 133-143.

Davies, R., Barbón, A., & Cuetos, F. (2013). Lexical and semantic age-of-acquisition effects on word naming in Spanish. Memory & Cognition, 41, 297-311.

Desrochers, A., Liceras, J. M., Fernández-Fuertes, R., & Thompson, G. L. (2010). Subjective frequency norms for 330 Spanish simple and compound words. Behavior Research Methods, 42(1), 109-117.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? Trends in Cognitive Sciences, 27(7), 597-600.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. Behavior Research Methods, 45, 1246-1258.

Fraga, I., Guasch, M., Haro, J., Padrón, I., & Ferré, P. (2018). EmoFinder: The meeting point for Spanish emotional words. Behavior Research Methods, 50, 84-93.

Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. Behavior Research Methods, 48, 963-972.

Gimenes, M., Brysbaert, M., & New, B. (2016). The processing of singular and plural nouns in English, French, and Dutch: New insights from megastudies. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 70(4), 316-324.

González-Nosti, M., Barbón, A., Rodríguez-Ferreiro, J., & Cuetos, F. (2014). Effects of the psycholinguistic variables on the lexical decision task in Spanish: A study with 2,765 words. Behavior Research Methods, 46, 517-525.

Grieve, J., Bartl, S., Fuoli, M., Grafmiller, J., Huang, W., Jawerbaum, A., ... & Winter, B. (2024, July 12). The Sociolinguistic Foundations of Language Modeling. ARXIV. https://arxiv.org/abs/2407.09241

Guasch, M., Ferré, P., & Fraga, I. (2016). Spanish norms for affective and lexico-semantic variables for 1,400 words. Behavior Research Methods, 48, 1358-1369.

Haro, J., Hinojosa, J. A., & Ferré, P. (2024). The role of individual differences in emotional word recognition: Insights from a large-scale lexical decision study. Behavior Research Methods. Advance publication. https://doi.org/10.3758/s13428-024-02488-z

Harrell, F.E. Jr. (2024). *Package 'rms' Version 6.8-2.* PSU. http://mirror.psu.ac.th/pub/cran/web/packages/rms/rms.pdf.

Hinojosa, J. A., Guasch, M., Montoro, P. R., Albert, J., Fraga, I., & Ferré, P. (2024). The bright side of words: Norms for 9000 Spanish words in seven discrete positive emotions. Behavior Research Methods, 56(5), 4909-4929.

Hinojosa, J. A., Martínez-García, N., Villalba-García, C., Fernández-Folgueiras, U., Sánchez-Carmona, A., Pozo, M. A., & Montoro, P. R. (2016). Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions. Behavior Research Methods, 48, 272-284.

Hussain, Z., Binz, M., Mata, R., & Wulff, D. U. (2024). A tutorial on open-source large language models for behavioral science. Behavior Research Methods. Advance publication. https://doi.org/10.3758/s13428-024-02455-8

Khanna, M. M., & Cortese, M. J. (2021). How well imageability, concreteness, perceptual strength, and action strength predict recognition memory, lexical decision, and reading aloud performance. Memory, 29(5), 622-636.

Kosaka, T., & Kikkawa, A. (2024). *When ChatGPT-4o Is (Less) Human-Like: Preliminary Subjective Rating Tests for Psycholinguistic Research.* RESEARCHGATE. https://www.researchgate.net/profile/Takumi-Kosaka/publication/384056408.

Li, Z., Shi, Y., Liu, Z., Yang, F., Liu, N., & Du, M. (2024, June 16). *Quantifying Multilingual Performance of Large Language Models Across Languages*. ARXIV. https://arxiv.org/abs/2404.11553

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, 1271-1291.

Marín-Gutiérrez, A., & Correales, M. D. (2023). *Latin American Norms of Emotional Valence for 1917 Spanish words*. CSIC. https://digital.csic.es/bitstream/10261/307466/1/Marin-Gutierrez_2023_Preprint_v.01.pdf

Martínez, G., Conde, J., Reviriego, P., Merino-Gómez, E., Hernández, J. A., & Lombardi, F. (2023a, September 28). *How many words does ChatGPT know? The answer is ChatWords*. ARXIV. https://arxiv.org/abs/2309.16777

Martínez, G., Conde, J., Merino-Gómez, E., Bermúdez-Margaretto, B., Hernández, J. A., Reviriego, P., & Brysbaert, M. (2024, August). *Using large language models to estimate features of multi-word*

expressions: Concreteness, valence, arousal. RESEARCHGATE. https://www.researchgate.net/publication/383529753

Martínez, G., Molero, J. D., González, S., Conde, J., Brysbaert, M., & Reviriego, P. (2023b, October 23). Establishing Vocabulary Tests as a Benchmark for Evaluating Large Language Models. ARXIV. https://arxiv.org/abs/2310.14703

Miguel-Abella, R. S., Pérez-Sánchez, M. Á., Cuetos, F., Marín, J., & Gonzalez-Nosti, M. (2022). SpaVerb-WN—A megastudy of naming times for 4562 Spanish verbs: Effects of psycholinguistic and motor content variables. Behavior Research Methods, 54(6), 2640-2664.

Moreno-Martínez, F. J., Montoro, P. R., & Rodríguez-Rojo, I. C. (2014). Spanish norms for age of acquisition, concept familiarity, lexical frequency, manipulability, typicality, and other variables for 820 words from 14 living/nonliving concepts. Behavior Research Methods, 46, 1088-1097.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. Journal of Experimental Psychology, 76(1, Pt.2), 1–25. https://doi.org/10.1037/h0025327

Pérez-Sánchez, M. Á., Stadthagen-Gonzalez, H., Guasch, M., Hinojosa, J. A., Fraga, I., Marín, J., & Ferré, P. (2021). EmoPro–Emotional prototypicality for 1286 Spanish words: Relationships with affective and psycholinguistic variables. Behavior Eesearch Methods, 1-19.

Real Academia Española. (2014). *Diccionario de la lengua Española.* https://dle.rae.es/.

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

Sabater, L., Guasch, M., Ferré, P., Fraga, I., & Hinojosa, J. A. (2020). Spanish affective normative data for 1,406 words rated by children and adolescents (SANDchild). Behavior Research Methods, 52, 1939-1950.

Sarli, L., & Justel, N. (2022). Emotional words in Spanish: Adaptation and cross-cultural differences for the affective norms for English words (ANEW) on a sample of Argentinian adults. Behavior Research Methods, 54(4), 1595-1610.

Schepens, J., Marx, N., & Gagl, B. (2024, September 13). Can we utilize Large Language Models (LLMs) to generate useful linguistic corpora? A case study of the word frequency effect in young German readers. OSF. https://osf.io/preprints/psyarxiv/gm9b6

Seco, M., Andrés, O., and Ramos, G. (1999). *Diccionario del español actual, volume 2.* https://www.fbbva.es/diccionario/info/el-diccionario/.

Spreen, O., & Schulz, R. W. (1966). Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. Journal of Verbal Learning and Verbal Behavior, 5(5), 459-468.

Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. Behavior Research Methods, 49, 111-123.

Thompson, B., & Lupyan, G. (2018, July). Automatic estimation of lexical concreteness in 77 languages. In The 40th annual conference of the cognitive science society (cogsci 2018) (pp. 1122-1127). Cognitive Science Society.

Trott, S. (2024a). Can large language models help augment English psycholinguistic datasets? Behavior Research Methods, 56, 6082-6100.

Trott, S. (2024b). Large Language Models and the Wisdom of Small Crowds. Open Mind, 8, 723-738.

Van Heuven, W. J., 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.

van Hoef, R., Connell, L., & Lynott, D. (2023). The effects of sensorimotor and linguistic information on the basic-level advantage. Cognition, 241, 105606.

Van Paridon, J., & Thompson, B. (2021). subs2vec: Word embeddings from subtitles in 55 languages. Behavior Research Methods, 53(2), 629-655.