Measuring the Semantic Priming Effect Across Many Languages

Abstract

Semantic priming has been studied for nearly 50 years across various experimental manipulations and theoretical frameworks. These studies provide insight into the cognitive underpinnings of semantic representations in both healthy and clinical populations; however, they have suffered from several issues including generally low sample sizes and a lack of diversity in linguistic implementations. Here, we tested the size and the variability of the semantic priming effect across nineteen languages by creating a large database of semantic priming values, based on an adaptive sampling procedure. Differences in response latencies between related word-pair conditions and unrelated word-pair conditions (i.e., difference score confidence interval is greater than zero) showed evidence for semantic priming, and a random intercept for language provided support for variability in semantic priming across languages.

# 

Measuring the Semantic Priming Effect Across Many Languages

Semantic priming is a well-studied cognitive phenomenon whereby participants are shown a cue word (e.g., DOG) followed by either a semantically related (e.g., CAT) or unrelated (e.g., BUS) target word1. Semantic priming is defined as the decrease in response latency (i.e., reduced linguistic processing or facilitation) for target words that are semantically related to their cue words in comparison to unrelated cue words1. Semantic priming research spans nearly 50 years of study as a tool to investigate cognitive processes, such as word recognition, and to elucidate the structure and organization of knowledge representation2, often by using results from these studies to develop theoretical and computational models that capture empirical effects3–6. Priming has also been used in studies of attention7,8, studies of bi/multilingual people9,10, on neurodivergent individuals such as those affected by Parkinson’s disease, aphasia, or schizophrenia, and in a large body of neuroscience studies11–13. The purpose of this study is to leverage the power and network of the Psychological Science Accelerator (PSA)14 to create a cross-linguistic normed dataset of semantic priming, paired with other useful psycholinguistic variables (e.g., frequency, familiarity, concreteness). The PSA is a large network of research laboratories committed to large-scale data collection and open scholarship principles.

Experimental psychologists have long understood that the stimuli in research studies are of great importance, and that controlled sets of normed information hold significant value for study control and allow for precision in measurement of effects. Often, stimuli are created in small pilot studies and then reused in many subsequent projects. However, both Lucas15 and Hutchison16 provided evidence that these small pilot data should be carefully interpreted given larger, more reliable datasets. In recent years, researchers have begun to more frequently publish large datasets with experimental stimuli for reuse in future work17. These datasets include lexical frequency18,19, large collections of text (e.g., corpora)20, response latencies,21–23 and subjective ratings from participants on semantic dimensions such as emotion24–26, concreteness27, or familiarity28. Recent advances in computational capability, the growth of large-scale online data collection, and the focus on replication and reproducibility may advance this research area. The importance of normed stimuli for research cannot be overstated. Not only do they provide methodological standardization for studies using the stimuli, but the stimuli themselves can also be studied to gain insight into cognitive architecture and processes, such as attention, memory, perception, and language comprehension or production.

Normed datasets provide a wealth of information for studies on semantic priming. Facilitation in priming is based chiefly on semantic similarity or the related word-pair condition as contrasted to the unrelated word-pair condition. Traditionally, word-pairs were simply grouped into pairs that were face-value similar (e.g., DOG-CAT) and unrelated (e.g., BUS-CAT), which was determined through pilot studies where word-pairs provided the expected statistical results. However, for reproducibility and methodological control, semantic similarity values should be defined before the results are known29. Semantic similarity has various conceptual and computational definitions that all generally describe the shared meaning between two words or texts5. The most common forms of similarity are feature-based similarity (i.e., number of shared features between words)30–32, association strength (i.e., the probability of a first word eliciting a second word when simply shown the first word)33,34, or text co-occurrence (i.e., words are similar because they frequently appear in proximity to one another)35–37. Each of these computational definitions of similarity can be calculated from normed datasets or text corpora to provide a continuous measure of similarity distance from 0 (unrelated) to 1 (perfectly related).

The Semantic Priming Project comprised both a large-scale database collection and a semantic priming study that used defined stimuli to create related word pairs21. This project provided data for lexical decision and naming tasks for 1,661 English words and non-words, along with other psycholinguistic measures for future research. The results of the Semantic Priming Project showed 23 ms to 25 ms decreases in word response latencies (i.e., lexical decision or naming speed) for the related word-pair conditions compared to unrelated word-pair conditions. The proposed study seeks to expand this dataset and address three key limitations of the Semantic Priming Project: reliability of item level effects, small sample sizes per item, and the focus on English words and English-speaking participants.

First, Heyman et al.38 explored the split-half reliability of item-level priming effects from the Semantic Priming Project, finding low reliability for the effects. This result corresponds with Hutchison et al.’s39 study, showing low reliability for priming effects; however, they demonstrated that priming effects can still be predicted at the item-level, albeit with a smaller dataset. Relatedly, for the second limitation, Heyman et al.40 noted that the required sample size necessary for reliable priming effects was much larger than the sample size used in the study, potentially explaining the differences between results as well as demonstrating the need for a larger dataset.

Last, the Semantic Priming Project only contains English data. If semantic priming provides a window into the structure of knowledge, the dominant focus on specific languages, such as English, has limited our understanding of the influence of linguistic variation on representation. Languages differ in script, syllables, morphology, and semantics, as well as the cultural variations that occur across language users. Related concepts that one may consider universal, such as LEFT and RIGHT, are not coded into all languages. Studies with more than one language within the same study often focus on bi/multilingual individuals to elucidate the potential shared structure of knowledge across languages41,42. Therefore, claims about human language are often based on a small set of languages, limiting the generalizability of these claims43. Even with the increase in publication of normed datasets in non-English languages17, conducting cross-linguistic studies on the same concepts is challenging, as large-scale data in this area are sparse.

Although it is challenging, using newer computational techniques44,45 and recently published corpora20,46, a broader coverage dataset in up to 43 languages is possible. Therefore, this study aims to provide data that complements and extends the published data, which would encourage research on methodology, item characteristics, models, cross-linguistic consistency in priming, and other theoretical areas that semantic priming has been applied to previously. The data will address the proposed limitations by increasing sample size to hopefully improve reliability and expanding beyond the English language within the same target stimuli. From this openly shared data, two research questions will be assessed as detailed in Table 1:

1) Is semantic priming a non-zero effect? To assess this research question, we will examine the confidence interval of the semantic priming effect to determine if the lower limit of the confidence interval is greater than zero using an intercept-only regression model estimating across all languages. Therefore, we predict semantic facilitation with reduced response latencies for related word-pair conditions in comparison to unrelated word-pair conditions.

2) Does the semantic priming effect vary across languages when examining the same target stimuli? We will add a random intercept of language to the model estimated in Hypothesis 1 to estimate the variability of priming across languages. We will conclude there is variability between priming effects for languages when the AIC for the random-intercept model is two or more points less than the AIC for the model in Hypothesis 147. To contextualize these results, we will provide a forest plot of the priming effects for languages to demonstrate the pattern of variability. For Hypothesis 2, we do not specify predicted directions for the effects but do expect potential variability in priming effects across languages. It is logical to expect differences in language due to culture, orthography, alphabet, etc., and empirical data suggest meaningful differences between languages48,49.

This research crucially supplements the literature outlined above by focusing on several key components of psycholinguistic research. For sampling, we will use accuracy in parameter estimation to ensure precision in our estimates50,51 to address the known reliability issues in item-level responding38,40 to support Hypothesis 1. The items will be selected using new computational techniques for addressing semantic similarity44,45 with recently available large corpora of movie subtitles20 to appropriately match comparable items across languages. As noted in Buchanan et al.17, research in non-English languages is expanding; however, stimuli matching is still sparse across published databases. By using large corpora, items are matched not only in their similarity levels, but also for their frequency of use. Thus, differences in priming can be attributed to differences in linguistic structure or culture, rather than translation or poor item matching, supporting Hypothesis 2.

**Table 1**.

**Pre-registered Design Table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Question | Hypothesis | Sampling plan (e.g., power analysis) | Analysis Plan | Interpretation given to different outcomes |
| Is semantic priming a non-zero effect? | HA: Response latencies will be faster for related word-pairs in comparison to unrelated word pairs.    H0: Response latencies for related word-pairs will be slower or equal to those for unrelated word-pairs. | We will sample participants on items until they reach a desired accuracy in parameter estimation confidence interval width (SE = 0.09). | We will calculate the mean and 95% confidence interval for the priming effect subtracting related word conditions from unrelated word conditions at the item level by using an intercept-only regression model.    These calculations will be repeated for the data with 2.5 *z*-score outlier trials excluded and 3.0 *z*-score outlier trials excluded. | The results will support HA when the lower limit of the confidence interval is positive and non-zero > 0.0001    The results will be inconclusive when the lower limit of the confidence interval is negative or zero ≤ 0.0001. |
| Does the semantic priming effect vary across languages? | HA: Priming response latencies will be variable between languages (i.e., heterogeneous).    H0: Priming response latencies will not be variable between languages (i.e., homogenous). | We will sample participants on items until they reach a desired accuracy in parameter estimation confidence interval width (SE = 0.09). | We will add a random-intercept of language to the previous intercept-only model to assess overall heterogeneity.    These calculations will be repeated for the data with 2.5 *z*-score outlier trials excluded and 3.0 *z*-score outlier trials excluded. | The results will support HA when the ΔAIC (intercept-only minus random-intercept) is ≥ 2 points.      The results will be inconclusive when the ΔAIC (intercept-only minus random-intercept) is < 2 points. |

**Method**

**Power Analysis**

For our power analysis, we first detail the background on how we estimated sample size, explain accuracy in parameter estimation, provide two simulations based on previous research, and the final proposed sample size. We end this section by specifying why this procedure was superior to previous methods and the requirements for publication.

**Background**

One concern is how to estimate the sample size required for cue-target pairs, as the previous literature indicates variability in their results40. Sample sizes of *N* = 30 per study have often been used in an attempt to at least meet some perceived minimum criteria for the central limit theorem. We focused on the lexical decision task for our procedure, wherein participants are simply asked if a concept presented to them is a word (e.g., CAT) or non-word (e.g., GAT). The dependent variable in this study was response latency, and we used lexical decision data from the English Lexicon Project22 and the Semantic Priming Project21 to estimate the minimum sample size necessary for each item, as previous research has suggested an overall sample size may lead to unreliability in the item-level responses40. The English Lexicon Project contains lexical decision task data for over 40,000 words, while the Semantic Priming Project includes 1,661 target words.

**Accuracy in parameter estimation (AIPE)**

***AIPE description.*** In this approach, one selects a minimum sample size, a stopping rule, and a maximum sample size. A minimum sample size was defined for all items based on data simulation below. For the stopping rule, we focused on finding a confidence interval around a parameter that would be “sufficiently narrow”50–52. These parameters are often tied to the statistical test or effect size for the study, such as correlation or contrast between two groups. In this study, we paired accuracy in parameter estimation with a sequential testing procedure to adequately sample each item, rather than estimate an overall effect size. Therefore, we used the previous lexical decision data to determine our sufficiently narrow confidence by finding a generalized standard error one should expect for well measured items. After the minimum sample size, each item’s standard error was assessed to determine if the item had met the goals for accuracy in parameter estimation as our stopping rule. If so, the item was sampled at a lower probability in relation to other items until all items reach the accuracy goals or a maximum sample size determined by our simulations below.

***Estimates from the English Lexicon Project*.** First, the response latency data for the English Lexicon Project were *z*-scored by participant and session as each participant has a somewhat arbitrary average response latency53. The data was then subset for only real word trials that were correctly answered. The average sample size before data reduction was 32.69 (*SD =* 0.63) participants with an average retention rate of 84% and 27.41 (*SD* = 6.43) participants after exclusions. The retention rates were skewed due to the large number of infrequent words in the English Lexicon Project, and we used the median retention rate of 91% for later sample size estimations. The median standard error for response latencies in the English Lexicon Project was 0.14, and the mean was 0.16. Because the retention rates were variable across items, we also calculated the average standard error for items that retained at least 30 participants at 0.12. This standard error rate represented our potential stopping rule.

The data was then sampled with replacement to determine the sample size that would provide that standard error value. One hundred words within the data were randomly selected, and samples starting at *n* = 5 to *n* = 200 were selected (increasing in units of five). The standard error for each of these samples was then calculated for the simulation, and the percent of samples with standard errors at or less than the estimated population value was then tabulated. In order to achieve 80% of items at or below the proposed standard error, we will need approximately 50 participants per word. This value was used as our minimum sample size for a lexical decision task, and the accuracy standard error level was potentially set at 0.12.

***Estimates from the Semantic Priming Project.*** This same procedure was examined with the Semantic Priming Project’s lexical decision data on real word trials. The priming response latencies are expected to be variable, as this priming strength should be predicted by other psycholinguistic variables, such as word relatedness. Therefore, we aimed to achieve an accurate representation of lexical decision times, from which priming could then be calculated. However, it should be noted that accurately measured response latencies do not necessarily imply “reliable” priming or difference score data54, but larger sample sizes should provide more evidence of the picture of item-level reliability. We used this data paired with the English Lexicon Project to account for the differences in a lexical decision only versus priming focused task. The average standard error in the Semantic Priming Project was less at 0.06, likely for two reasons: the data in the Semantic Priming Project are generally frequent nouns and only 1,661 concepts, as compared to the 40,000 in the English Lexicon Project. The retention rate for the Semantic Priming Project was less skewed than the English Lexicon Project at a median of 97% and mean of 96%. Using the same sampling procedure, we estimated sample sizes of *n* = 5 to *n* = 400 participants increasing by units of 5. In this scenario, we found the maximum sample size of 320 participants for 80% of the items to reach the smaller standard error of 0.06. Therefore, we used 320 as our maximum sample size, and the average of the two standard errors found as our stopping rule, i.e., 0.09.

***Final sample size.*** Given our minimum, maximum, and stopping rule, we then estimated the final sample size per language based on study design characteristics. Participants completed approximately 800 lexical decision trials per session, and each participant only completed 150 of these concepts (75 targets in the related condition, 75 targets in the unrelated condition as cue words were not analyzed) that were the target of this sample size analysis (see below for more details on trial composition). Therefore, the target number of items (*n* = 1000 concepts) was multiplied by the minimum/maximum sample size, and conditions (related word pair versus unrelated word pair) and divided by the total number of usable lexical decision trials per participant times the data retention rate (a conservative estimate of 90%). The final estimate for sample size per language was 741 to 4741 [(1000\*50\*2) / (150\*.90); (1000\*320\*2) / 150\*.90]. The complete code and description of this process are detailed in our supplemental documents.

This sample size estimation represents a major improvement from previous database collection studies, as many have used the traditional *N* = 30 to guess at minimum sample size. Because the variability of the sample size was quite large, we employed a stopping procedure to ensure participant time and effort was maximized, and data collection was optimized. To summarize, the minimum sample size was 50 participants per word and the maximum for the adaptive procedure was 320, which results in 741 to 4741 participants per language based on expected usable trials. Therefore, the total sample size was proposed to be 7410 to 47410 participants for ten languages. After 50 participants who answered a real word item, each concept was examined for standard error, and data collection for that concept was decreased in probability when the standard error reached our average criterion of 0.09. Item probability for selection was also decreased when they reached the maximum proposed sample size (*n* = 320). This process was automated online and checked in a scheduled subroutine.

**Languages**

43 languages were originally identified for possible data collection based on the information available from the OpenSubtitles20 and subs2vec46 project. For the following 30 languages/dialects, we translated stimuli and collected data from at least one participant (starred languages were included in our pre-registered minimum data collection plan): Arabic, Brazilian Portuguese, Czech\*, Danish, Dutch, English\*, Farsi, French, German\*, Greek, Hebrew, Hindi, Hungarian, Italian, Japanese\*, Korean\*, Norwegian, Polish, Portuguese\*, Romanian, Russian\*, Serbian, Simplified Chinese\*, Slovak, Slovenian, Spanish\*, Swedish, Thai, Traditional Chinese, Turkish\*, and Urdu. Table 2 portrays a summary of each language’s data collection for the number of included participants (based on the pre-registered data inclusion rules), the number of excluded participants, proportion of correct answers for included participants[[1]](#footnote-1), and the median completion time for included participants. The complete breakdown of gender, education, age, and stimuli completion can be found in the supplemental materials. The following languages met the minimum data collection requirements and will be analyzed in this manuscript: Czech, Danish, German, Greek, English, Spanish, French, Hungarian, Italian, Japanese, Korean, Polish, Portuguese, Brazilian Portuguese, Romanian, Russian, Serbian, Turkish, and Simplified Chinese. The stimuli list for Portuguese and Brazilian Portuguese overlapped by 90%, and data were combined such that each unique target (unrelated and related trials) obtained the minimum number of participant answers[[2]](#footnote-2). All data are available online, even for those languages that did not meet the pre-registered minimum data collection criterion for analysis. For each language, we also provide data checks and a summary of the number of participants, trials, items, and priming trials during data processing (see Supplemental Materials).

**Table 2.**

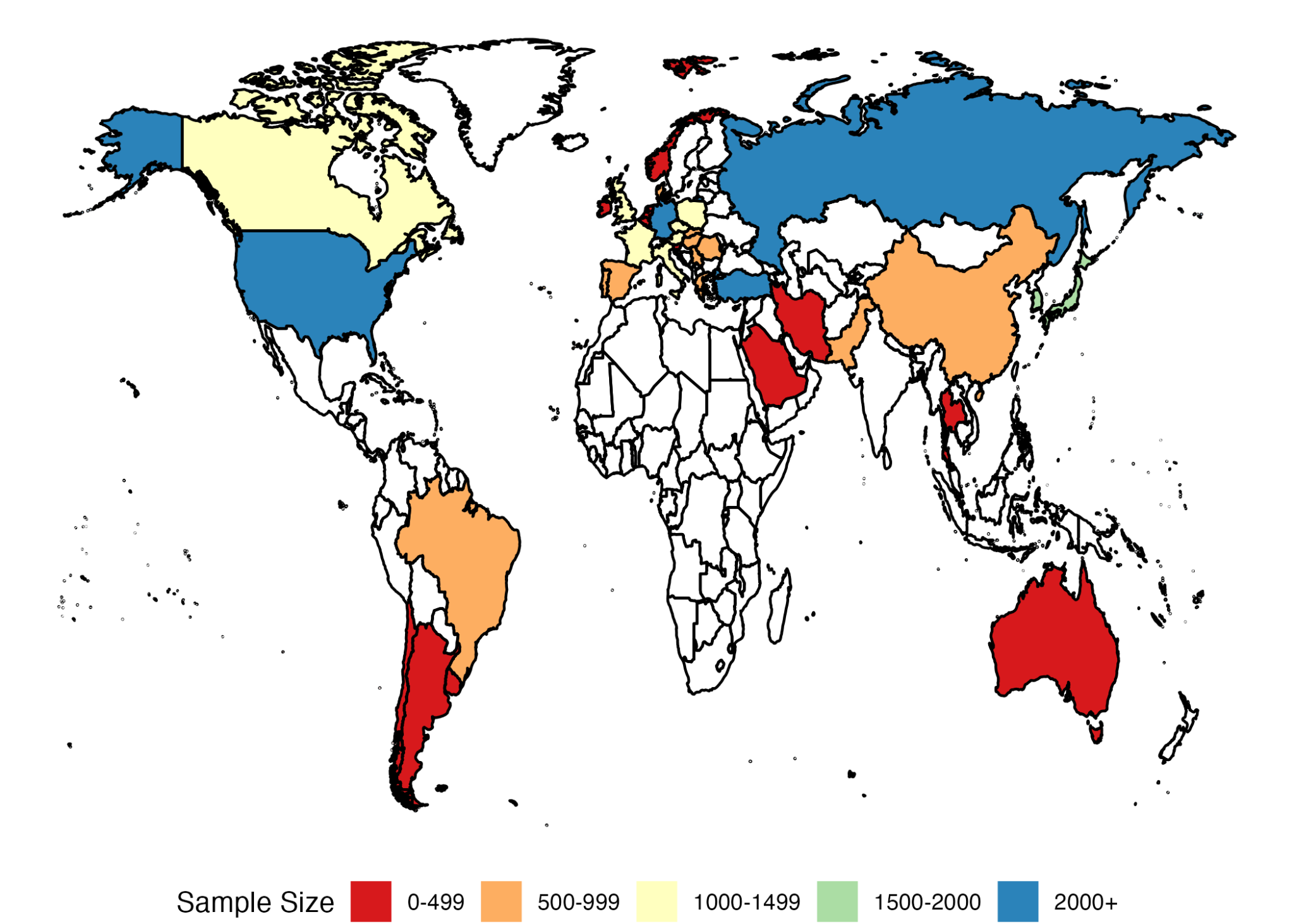
**Language Data Collection Sample Sizes, Accuracy, and Median Response Time**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Language | N Include | N Exclude | Proportion Correct | Median Time |
| Arabic | 133 | 102 | 0.92 | 18.67 |
| Czech | 1,074 | 362 | 0.94 | 19.76 |
| Danish | 829 | 167 | 0.93 | 18.70 |
| German | 2,628 | 469 | 0.94 | 19.02 |
| Greek | 689 | 130 | 0.94 | 18.48 |
| English | 5,122 | 1,607 | 0.92 | 17.64 |
| Spanish | 1,468 | 284 | 0.94 | 18.04 |
| Farsi | 192 | 110 | 0.95 | 17.71 |
| French | 869 | 142 | 0.95 | 17.68 |
| Hebrew | 247 | 74 | 0.92 | 16.63 |
| Hindi | 1 | 2 | 0.82 | 27.39 |
| Hungarian | 718 | 180 | 0.94 | 17.94 |
| Italian | 1,085 | 142 | 0.95 | 18.10 |
| Japanese | 1,165 | 680 | 0.94 | 18.69 |
| Korean | 975 | 601 | 0.91 | 17.59 |
| Dutch | 184 | 25 | 0.93 | 17.60 |
| Norwegian | 85 | 17 | 0.93 | 20.08 |
| Polish | 1,188 | 318 | 0.94 | 19.15 |
| Portuguese | 1,178 | 332 | 0.93 | 18.25 |
| Romanian | 741 | 174 | 0.94 | 19.65 |
| Russian | 1,806 | 956 | 0.94 | 19.68 |
| Slovak | 381 | 391 | 0.94 | 18.68 |
| Slovenian | 31 | 10 | 0.95 | 18.89 |
| Serbian | 681 | 109 | 0.94 | 21.01 |
| Thai | 65 | 20 | 0.95 | 18.34 |
| Turkish | 2,218 | 790 | 0.93 | 17.83 |
| Urdu | 315 | 381 | 0.88 | 22.15 |
| Simplified Chinese | 729 | 291 | 0.93 | 17.75 |
| Traditional Chinese | 174 | 67 | 0.92 | 18.05 |

**Ethics and Research Labs**

We did not collect any identifiable private or personal data as part of the experiment. This project was approved by Harrisburg University of Science and Technology conforming to all relevant ethical guidelines and the Declaration of Helsinki, with special care to conform to the General Data Protection Regulation (GDPR; eugdpr.org). No global exclusion criteria for participating in the study were used, except for a minimum age requirement of 18 years. Please see the analysis section below for other participant and trial level exclusion criteria related to analyses.

133 labs completed ethics documentation for data collection, and 126 labs in 41 geopolitical regions collected data (i.e., *N* >= 1) for the study. Each data collection lab obtained local ethical review (81), relied on the ethical review provided by Harrisburg University (31), or provided evidence of no required ethical review (14). The supplemental materials provide links to the IRB approvals on the Open Science Framework (OSF) and a table of participating labs with their data collection information, which includes languages sampled, geopolitical region of the team, compensation procedure and amount, online versus in person testing, and testing type (individual participants or classroom type settings). This information can be matched to study data using the lab code that is present in the participant and trial level files. Figure 1 demonstrates a visualization of the entire sample using the lab’s geopolitical region during data collection.



**Figure 1**. Binned sample sizes based on research lab geopolitical region demonstrating the full data available for reuse from the project.

**Participants**

35,904 participants opened the link to the study, and 31,645 proceeded through to complete at least one trial of the study trials (i.e., past the practice trials). Of these participants, 26,971 were retained for analysis because they met our three participant level inclusion rules: 1) at least 18 years of age, 2) completed at least 100 trials, and 3) scored at least 80% correct. All exclusion criteria are summarized in the results section for clarity. The pre-registered plan calculated accuracy as in the planned scripts; however, an administrative team discussion revealed that the pre-registered report could be interpreted as . If accuracy is recalculated using this formula, 28,162 participants would be included for analysis. This report will use the stricter criterion of accuracy for analysis, while the analysis using the rescored accuracy can be found in our supplemental materials. The analyses reported below only examine languages that met the minimum data criteria, which includes 32,897 total participants, 29,155 who completed at least one trial, 25,163 who met the strict inclusion rules, and 26,197 that met the rescored version of the inclusion rule for accuracy. Descriptive statistics about participants are provided below in the results for the 25,163 participants who met the strict inclusion rules.

**Materials**

The following details the important facets of the materials. We first explain the types of word-pair conditions in a semantic priming study (i.e., related, unrelated, and non-word). Next, we detail how the related word-pair conditions were created using the OpenSubtitles corpora, new computational modeling techniques, and the selection procedure.

**Word-pair conditions**

In a semantic priming study, there are three types of word-pair conditions. In the related word-pair condition, cue-target pairs are chosen for their similarity or relatedness. Cosine distance is similar to correlation in representing relatedness; however, cosine distance is always positive. Therefore, a cosine distance (when used for similarity purposes) of 1 represents the same numeric vectors (perfect similarity), while a cosine distance of 0 represents no similarity between vectors. To create the unrelated condition, cue-target pairs were shuffled so that the cue word was combined with a target word with which it had a negligible cosine distance similarity (i.e., < .15).

***Deviations*.** For English, cosine similarity for unrelated pairs were shuffled until all but one pair was less than .15. The pair (ONE-TORTURE) that did not achieve this criterion had a cosine of .20, as the word ONE is a high frequency word with high cosine similarity values regardless of the target that could have been selected. For Korean, in the final shuffle of the unrelated word pairs (approximately 100 word pairs), we increased the unrelated cosine criterion to .20 to find the lowest possible pairs, as below .15 was not possible for many pairs due to the smaller dataset size. For Czech, the maximum cosine for unrelated pairs was ~ .16. For Japanese, nearly all pairs were related at very high levels (i.e., *M* = .80 for cosine), which is very unlikely. We shuffled the pairs for the unrelated trials and picked the lowest possible combination for running the study. This model (fastText) was created in the same way as described in the subs2vecpaper46, but these cosine values are improbable. For Serbian, Simplified Chinese, and Traditional Chinese, the same problem occurred that all word pairs were very highly correlated. We followed the same procedure as described for Japanese. Non-words pair conditions were created by using the Wuggy-like algorithm55 for non-logographic languages. For logographic languages, we consulted with at least two native speakers to change one stroke or radical such that the character(s) were a pronounceable word with no meaning by starting from known non-word lists56. Any disagreements between native speakers were resolved by discussion between these speakers. Each cue and target word were first hyphenated using the *sylly* package and LaTeX style hyphenation57. If words were not hyphenated, as they were one syllable or the syllables were not clear, we created bigram character pairs for replacement purposes. The 100,000 most frequent words for each language from the OpenSubtitles data were also hyphenated in this style. From the OpenSubtitles data, we calculated the frequency of each pair of possible hyphenation combinations (e.g., NAPKIN → [\_, NAP], [NAP, KIN], [KIN, \_]) as the transition frequency from Wuggy. For each cue and target, we selected a set of character replacements that: kept or matched closely to the same number of characters as the original word, minimized transition frequency (i.e., the frequency of the replacement was very close to the frequency of the original pair of hyphenated characters), and matched the number of character changes to the number of syllables. At least two native speakers examined each programmatically generated word to ensure they were pronounceable (i.e., phonologically valid) and not pseudo-homophones (i.e., wherein the pronunciation sounds like a real word, KEEP → KEAP)55. In cases of disagreement, the native speakers discussed and resolved these inconsistencies. When they marked a non-word for exclusion, a new non-word was generated until speakers agreed it met the rules for non-words. Native speakers also suggested alternatives, which the lead author checked to ensure that they matched the desired non-word characteristics.

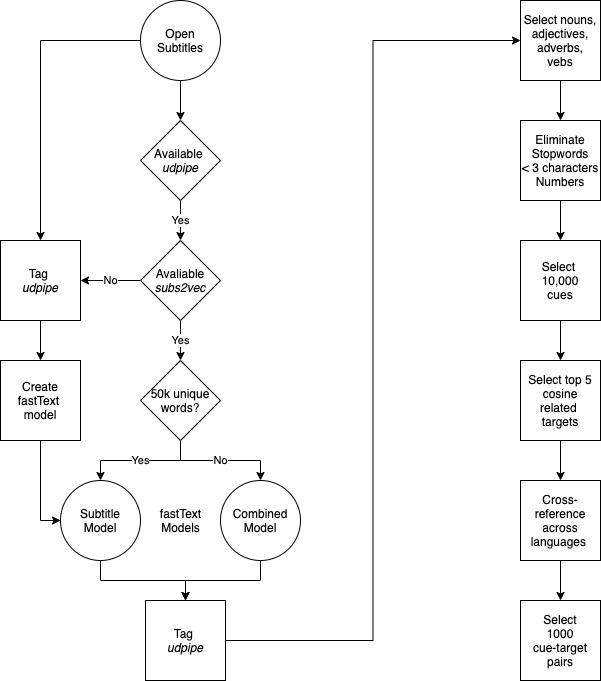
***Deviations*.** A few words were found to be incorrectly marked as non-words or were misspelled in the dataset. These trials were corrected during data collection or post-data collection in the data processing scripts. These deviations and issues are noted in the data processing files found online. Translators often suggested new nonword options from the computationally generated list. Given that translators were native speakers, we used their expertise for this component.

To control the ability of participants to anticipate or guess the answers, we ensured that half the trials should be answered with a word and half with a nonword. Therefore, we used 150 related trials (150 word / 0 nonword; 75 pairs), 150 unrelated trials (150 word / 0 nonword; 75 pairs), 200 word-nonword trials (100 word / 100 nonword, this could have been word-nonword or nonword-word combinations to control for answer chaining; 100 pairs), and 300 nonword-nonword trials (0 word / 300 nonword; 150 pairs). These trials were randomly presented to control the transition probability between word and nonword trials (i.e., random presentation should ensure trials do not present a word-word-nonword-nonword style pattern that allows participants to mindlessly guess the answers). Therefore, the yes-no probability was 50% for words-nonwords across all trials, and the relatedness proportion for pairs was 18.8%. The exact trial proportions for each language can be found online in our data processing summary, as not all participants completed all trials, which can change proportions for each language.

**Similarity calculation**

***Corpora.*** As described in the introduction, the choice of related words based on similarity was key for the study. There are multiple measures of semantic similarity including the cosine similarity between overlapping features32, free association probabilities33,34,58, and local/global coherence values from network models. However, the underlying data for these calculations is inconsistent across languages. Therefore, one solution is to use the data present in the OpenSubtitles datasets20 (i.e., a large collection of movie subtitles) to calculate word frequency and cosine similarity values. These datasets have been used to calculate word frequencies for the SUBTLEX projects, which have validated their use as strong predictors of cognitive related phenomena18,59–66. Cosine similarity was selected over other similarity measures because of the availability of possible languages and models for this project, as described below.

The OpenSubtitles data includes 62 languages or language combinations (e.g., Chinese-English mix). We used the 10,000 most frequent nouns, adjectives, adverbs, and verbs from each potential language without lemmatization (i.e., converting words into their dictionary form RUNS → RUN). The *udpipe* package67 is a natural language processing package that contains more than 100 treebanks to assist in part of speech tagging (i.e., labeling words as noun, verb, etc.), parsing (i.e., separating blocks of text into words and their relationship to other words in a text), and lemmatization. This package was selected for its large coverage of languages with reliable part-of-speech tagging. Cross-referencing the available languages in *udpipe* with the OpenSubtitles data allowed for the possibility of 43 different languages in this project. See Figure 2 for the model selection process.



**Figure 2.** Stimuli selection method flow chart. Circles represent the data or models used in the decision tree. Diamonds represent a decision criterion for the data selected. Squares represent coding processes or data reduction for the final stimuli set.

***Modeling.*** The subs2vec project46 used the OpenSubtitles data to create fastText68 computational representation for 55 languages. fastText is a distributional vector space model, an extension of word2vec44,45**,** wherein each word in a corpus is converted to a vector of numbers that represents the relationship of that word to a number of dimensions. These dimensions can be imagined as a thematic or topic representation of the text. The relationship between these vectors represents the similarity between concepts, as words that have similar or related meanings will appear in similar places and dimensions in a text, and will, therefore, have similar numeric vectors4,5. We used the existing models from subs2vec to extract related word concepts for the most frequent concepts identified using the top cosine distance between word vectors. When the model was not present in subs2vec, we recreated the same model using their parameters on the relevant OpenSubtitles data.

***Cue selection procedure.*** The procedure for stimuli selection can be reviewed in our supplemental materials and is displayed graphically in Figure 3. If the language was available via subs2vec, the provided subtitle frequency counts were examined. If the language has more than 50,000 unique concepts represented in the subtitle data, we used the subtitle model only. If the subtitles do not provide enough linguistic information (i.e., fewer than 50,000 concepts in the corpus), we used the combined Wikipedia and subtitle model46. subs2veccontains models with only the OpenSubtitles data, only Wikipedia for a given language, and a combined model of both. The subtitle data has shown to best represent a language18,59; however, not all subtitle projects contain a large enough corpus for the subtitles to cover the breadth of the possible concepts within that language (e.g., Afrikaans subtitles only represent approximately 18,000 words).

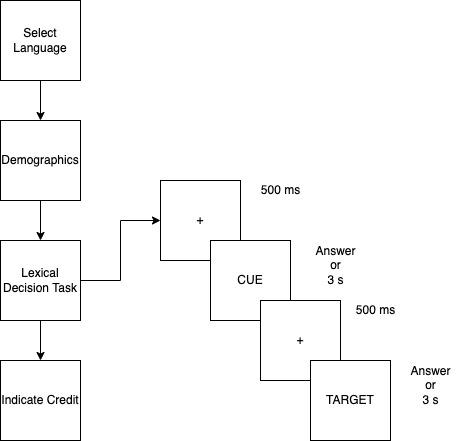
The selected token list was then tagged for part-of-speech using *udpipe,* selecting tokens that were tagged as nouns, adjectives, adverbs, and verbs. From the *udpipe* output, the lemma for each token was selected to control for high similarity between lemma-token forms (e.g., RUN is highly related to RUNS). All stopwords (i.e., commonly used words in a language with little semantic meaning such as THE, AN, OF), words with fewer than three characters for non-logographic languages, and words with numeric characters will be eliminated (i.e., 1 would be eliminated but not ONE). The stopword lists can be found in the stopwords package using the Stopwords ISO dataset69. This procedure covered all but two languages in our list of 43 possible languages. For the final two languages, we used *udpipe* to tag the OpenSubtitles directly and calculate word frequency. Additionally, fastText models using the same parameters as subs2vec were trained for similarity calculation. The 10,000 most frequent concepts were selected at this point.

***Target selection procedure.*** Using the fastText models for each language, we selected the top five cosine distance similarity values for each concept in each language independently, resulting in 50,000 possible cue-target pairs. These were cross-referenced across languages using Google Translate to create a master list of potential cue-target pairings. The related word pairs (*n* = 1000) were selected from this list using each cue or target only once, favoring pairs with translations in most languages. Therefore, the selection procedure was based on the most common cue-target pairs across languages, rather than selecting similar words in one language and then translating. This procedure was programmatic, using Google Translate, which may not produce the most appropriate translation for a word. Therefore, native speakers ensured the accurate translation of word pairs using the PSA’s translation network for the final selected set in a similar manner as described above. They suggested a more common or appropriate word for items they thought were unusual, and in cases of disagreement, group discussion between the two translators took place. In some instances, translation may have indicated that a particular language does not have separate concepts for the cue-target pairing. In this instance, we changed the cue word to a related word for that language from the five selected in the original list. Thus, all targetswere matched across languages, and as many cues as possible while avoiding repetition within a cue-target pair as best possible.

***Deviations*.** We described that we would filter words in the OpenSubtitles for words with at least 3 characters (minus logographic languages). This process was completed, and all cue words are at least 3 characters; however, when we matched cues to high cosine targets, several two letter words were included. Additionally, due to translation suggestions and cross referencing, other two letter words were included. For example, in English, MAKE-GO, DOWN-UP, ENTER-GO were included as potential related cue-target pairs for target selection.

**Procedure**

We describe the important components to the procedure in this section. First, we detail the implementation of the study, focusing on the timing software and adaptive stimuli section, as not all participants see all items. We then discuss the study procedure in order, as shown in Figure 2. First, participants completed a demographic questionnaire, followed by the lexical decision task. We explain how our data compliments the Semantic Priming Project and finally, discuss additional data that we plan to combine with the current dataset.



**Figure 3.** Flow chart of the procedure for the study. Within the lexical decision task, participants are given short breaks after 100 trials (i.e., each answer given). The answer choices for that language will always be displayed on the bottom of the screen during the lexical decision task.

**Implementation**

***Timing software.*** While participants were naïve to the word pairings, the principal investigator knew the pair combinations during data collection and analysis. A small demonstration of the experiment can be found at:<https://psa007.psysciacc.org/> or recreated from our supplemental materials. The study was programmed using lab.js70, which is an online, open-source, study-building software. Precise timing measurement was required for this study, and the lab.js team has documented the accuracy of measurement within their framework71, and previous work has shown no differences between lab and web-based data collection for response latencies72. In addition, SPALEX, a large lexical decision database in Spanish, was collected completely online23. We recommended that research labs suggest Chrome as their browser for participants completing the study due to recommendations from the lab.js team. However, meta-information about the browser and operating system are saved when participants take the experiment to examine for potential implementation differences.

Participants were directed to an online web portal to complete the study, and all data was retained in the online platform with regular backups to the server. Participants were required to complete the study on a computer with a keyboard, rather than on a device with only a touch screen. This requirement allows for tracking of the display of the device which indicates important aspects about screen size, browser, and timing accuracy. In order to enforce this requirement, participants were asked to hit the spacebar to continue the study.

***Adaptive stimuli selection.*** At the start of data collection, all presented items were randomly selected from the larger item pool by equalizing the probability of inclusion equal for all words and non-words (*p* = 1/1000 concepts)[[3]](#footnote-3). After the minimum sample size was collected, each word’s standard error was checked to determine if the sample size for that item had reached our accuracy criteria. If so, the probability of sampling that item was decreased by half. Once a concept has reached the maximum required sample size, the probability of sampling will also be decreased by half. This procedure will allow for random sampling of the items that still need participants without eliminating words from the item pool. Therefore, we ensured that there were always words to randomly select from (i.e., to keep the same procedure and number of trials for all participants) and that the randomization was a sampled mix of words that reach accuracy quickly and words that need more participants (i.e., participants do not only see the unusual words at the end of data collection). Once all words reached the stopping criteria or maximum sample size, the probabilities were equalized. We set minimum, maximum, and a stopping rule for the initial data collection; however, we allowed data collection after these were reached and will post updates to the data using a DOI service to allow researchers to cite the specific dataset they used for their research73 (modeled after the Small World of Words Project33, which is ongoing). All data is included in our dataset, and the analysis section describes how we indicated exclusion criteria. Therefore, data collection was a repeated-measures design in which participants did not see all of the possible stimuli, but did see all the possible conditions (related, unrelated, and non-word pairs). They were blind to the condition each pair was presented in.

***Deviations***. One issue with data collection sites, such as mTurk and Prolific, was the speed of data collection. For example, a researcher can collect thousands of participants in an hour with these services. Our study was designed to collect data more slowly across time, to implement the stimuli randomization and selection algorithm. In order to control for the speed of collection using these sites and any other simultaneous participant runs (i.e., classroom testing), we did the following: multiple versions of the study were programmed and participants were assigned to a random version via Qualtrics randomizer. They were then redirected back to their paid provider. Each version of these languages continued to use the adaptive randomization and selection algorithm. For large paid samples funded by ZPID and Harrisburg University (<https://leibniz-psychology.org/>: Japanese, Russian, Turkish, Czech, and Korean), we created 14 different randomizations that evenly distributed the pairs across the study (with a small overlap as the important trials (word-word) do not evenly distribute). These were static during the data collection process to ensure that we obtained 50+ participants in the paid samples for each word-word trial. After these data collections, the algorithm was turned back on for PSA labs.

Additionally, we decided to run the algorithm/randomization process once every 5 minutes when the data collection for a language started. This update allowed the randomization to be more frequent during the early stages of data collection. As data size increased, we increased the time interval, to account for the time it took for the algorithm code to run, so that each randomization could finish before the next one was scheduled to start. This process also ensured that the .json files of randomized stimuli were not overwritten or corrupted if two processes were running at once.

Some participants may only have a limited number of trials, even though they “completed” the experiment. This event sometimes happened when the algorithm was programmed to run on the data in that folder but accidentally not programmed to write the new stimuli to that folder (and therefore, it gave them test initial trials for six blocks and then two real blocks before we realized this was happening). These tests and inappropriate trials were excluded; please see data checks files for languages and number of trials affected. Other coding related issues included a typo that showed one trial pair twice at the beginning of the study (Japanese, Russian, Turkish, Czech, Korean, and English), instances of garbled non-Latin language items (i.e., symbols shown instead of cyrillic in Russian), and typos in language spellings. These issues were fixed as soon as they were discovered.

Last, when examining data collection progress, we noticed that Korean did not have all matched related-unrelated pairs. This error happened during the shuffle to get low cosine values. 33 new trials were added to ensure each related target had a corresponding unrelated target. In Arabic, it was suggested that we should exclude specific word pairs for their taboo nature and this request was honored; thus, the total number of possible stimuli is lower in that language.

**Study Procedure**

***Demographics.*** Participants were given a language specific link for each research lab. Participants were asked to indicate their gender (i.e., male, female, other, prefer not to say), year of birth, and education level (i.e., none, elementary school, high school, bachelors, masters, doctorate; or their equivalent in the target country of data collection) for demographic variables. They provided their native language in an open text box, and selected left or right as their dominant hand for the mapping of word-nonword answer keys. A flow chart of the procedure is provided in Figure 3.

***Lexical Decision Task.*** Instructions on how to complete a lexical decision task were shown on the next screen, followed by 10 practice trials. Each trial started with a fixation cross (+) in the middle of the screen for 500 ms. The stimulus item was then displayed in the middle of the screen in lowercase san-serif 18-point font (i.e., Arial font, dog). On the bottom of the screen the possible responses were shown as the traditional keys next to the *Shift* key depending on the most common keyboard layout for that language (i.e., Z and / on a QWERTY keyboard or < and - on a QWERTZ keyboard or numbers 1 and 9 for languages that had many keyboard layouts). Response keys were mapped such that the “nonword” response option is on the non-dominant hand side of the keyboard, and the “word” response option is on the dominant hand side74. Participants were asked for the dominant hand at the beginning of the study to determine the response mapping for their study. Participants made their choice for each concept, and during the practice trials, they received feedback if their answer was correct or incorrect. The next stimulus appeared with an intertrial interval of 500 ms (i.e., the time between the offset of the first concept response and onset of the next concept, when the fixation cross was showing). Responses timed out after three seconds and moved on to the next trial. After 10 trials, participants saw the instruction screen again with a reminder that they will now be doing the real task.

After 100 trials, the participants were shown a short break screen with the option to continue by hitting the spacebar after 10 seconds. This break timed out after 60 seconds. After eight blocks of 100 trials (800 word-nonword decisions), the experiment ended with a thank you screen. On this screen, participants were given instructions on how to indicate that they have completed the study to the appropriate lab. Participants were allowed to take the study multiple times as items were randomly selected for inclusion. An estimate for the time required for the study was approximately 30 minutes inclusive of practice trials, reading all instructions, and breaks. This estimate was based on previous studies of lexical decision times22, and the final median completion time was approximately 18 minutes.

***Comparison to the Semantic Priming Project.*** This procedure is a single stream lexical decision task wherein every concept (cue and target) is judged for lexicality (i.e., word/non-word). Many priming studies often present cue words for a short period of time prior to the presentation of target words for lexicality judgment. Evidence from the Semantic Priming Project suggests that the stimulus onset asynchrony (i.e., time between non-judged cue word and target word) does not affect overall priming rates (25 versus 23 ms for 200 ms and 1200 ms). Further, adding the lexicality judgment to each presented concept creates a less obvious link between cue and target to avoid potential conscious expectancy generation effects75,76. Even though they appear sequentially in the task, they are not explicitly paired by being a non-judged cue word followed by a judged target word. Therefore, this procedure varies from the data collected in the Semantic Priming Project; thus, extending their work to different conditions. Lucas15 provides evidence that priming effect sizes are relatively equal across task type (e.g., continuous, masked, paired, and naming), and therefore, we should expect similar results.

***Additional data.*** We then combined available lexical and subject rating data with the priming data, and a tutorial is provided in the supplemental documentation on how to download data and combine with available norms. Lexical measures, such as length, frequency, part of speech, and the number of phonemes (i.e., sounds in a word) are easily created from the concept or the SUBTLEX projects59–65. Subjective measures are concept characteristics that are rated by participants, and we included age of acquisition77–80 (approximate age you learned a concept), imageability81,82 (how easy the concept comes to mind), concreteness83 (how concrete is the concept), valence (how positive versus negative is the concept), arousal (how excited or calm a concept makes a person), dominance (the word denotes something that is weak/subordinate or strong/dominant)24,26, and familiarity (how well a person knows a concept)84. These variables were selected from the list of most published databases for linguistic data17.

**Results**

We first detail the exclusion criteria from the pre-registered plan. Next, the descriptive statistics of the data are provided for participants, trials, items, and priming. The final section covers the hypothesis testing from Table 1. To reduce redundancy, we provide several overview tables of the descriptive results, and all pre-registered descriptives are provided in the supplemental documents linked in the appendix.

**Exclusion Summary**

Data will be excluded for the following reasons in this order (pre-registered plan):

1) Participant level data: the entire participant’s data will be removed from the analyses.

a) Participant did not indicate at least 18 years of age.

b) Participant did not complete at least 100 trials.

c) Participant did not achieve 80% correct.

2) Trial level data: only the individual trials will be removed from the analyses.

a) Timeout trials (i.e., no response given in 3 s window).

b) Incorrectly answered trials.

c) Response latencies shorter than 160 ms.

3) Trial level exclusions dependent on test: trials marked for exclusion that are tested with and without these values in the hypotheses described below.

a) Response latencies over the absolute value of *Z* = 2.5.

b) Response latencies over the absolute value of *Z* = 3.0.

**Descriptive statistics**

**Participant level data**

Participants were generally female (55.49%) or male (37.39)% with the rest either missing data, not wanting to indicate their gender or other. If data was filtered to only participants that were kept for analysis, the data was again predominantly female (60.95%) or male (37.44)%. Participants indicated they had completed high school (42.77%), some college (7.63%), college (30.47%), master’s (9.30%), and the rest other options (Less than High School, Missing, or Doctorate). Participants kept for analysis also followed this pattern: high school (46.02%), some college (8.34%), college (31.97%), master’s (9.61%). The top twenty native languages represented in the data are shown in Table 3. Full language percent tables can be found in the supplemental data. The data indicates that the pattern of native languages are similar in the full data and data used for analysis. The average age for all participants was *M* = 31.38 (*SD* = 14.95) ranging from 18.00 to 104.00. The demographic question asked participants to enter their year of birth, and the high maximum values were likely participants who entered the minimum possible year allowable in the data collection form. The participants kept for analysis showed the same pattern of ages: *M* = 30.43 (*SD* = 14.17) ranging from 18.00 to 104.00.

**Table 3.**

**Native and Browser Languages Indicated by Participants for the Overall and Analyzed Participants**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Native Language | | Browser Language | |
| Language | Overall Percent | Keep Percent | Overall Percent | Keep Percent |
| English | 15.83 | 17.19 | 27.35 | 27.65 |
| Turkish | 8.41 | 8.63 | 8.60 | 8.30 |
| German | 7.80 | 9.39 | 8.53 | 9.72 |
| Missing | 7.76 | 1.65 | 2.85 | 2.61 |
| Russian | 7.61 | 6.99 | 8.10 | 6.99 |
| Spanish | 5.39 | 6.13 | 4.85 | 5.35 |
| Japanese | 5.03 | 4.51 | 5.54 | 4.57 |
| Polish | 4.36 | 4.65 | 4.35 | 4.35 |
| Korean | 4.23 | 3.81 | 4.58 | 3.72 |
| Portuguese | 4.06 | 4.37 | 3.98 | 4.15 |
| Czech | 3.88 | 4.07 | 4.15 | 4.04 |
| Italian | 3.74 | 4.38 | 3.54 | 4.09 |
| French | 2.8 | 3.31 | 2.83 | 3.25 |
| Danish | 2.79 | 3.20 | 2.61 | 2.90 |
| Hungarian | 2.72 | 2.96 | 2.36 | 2.45 |
| Mandarin | 2.58 | 2.68 | NA | NA |
| Greek | 2.35 | 2.73 | 1.60 | 1.73 |
| Serbian | 2.27 | 2.66 | 0.45 | 0.5 |
| Romanian | 1.99 | 2.23 | 0.96 | 1.08 |
| Chinese | 0.62 | 0.57 | 2.43 | 2.24 |

*Note*. Browser language did not distinguish between Mandarin and Cantonese, but instead was marked generally as Chinese.

The majority of participants used a Windows based operating system (76.91%) followed by Mac OS (18.45%), and Linux (1.80%), with some missing data (2.85%). The operating systems were the same for the participants used in analysis: Windows (76.82%), Mac (18.70%), Linux (1.86%), and Missing (2.61%). Web browsers were grouped into the largest categories for reporting. The large majority of participants used Chrome (58.96%), followed by Edge (14.92%), Firefox (8.18%), Safari (8.88%), and Opera (3.09%) A surprise result was that the Russian browser Yandex contributed a small (but not extremely small) percent of the browser usage: (2.37%), followed by other web browsers (3.60%) The results were the same when examining participants who were kept for analysis: Chrome (59.81%), Edge (14.23%), Firefox (8.43%), Safari (9.22%), Opera (2.99%), Yandex (2.03%), and Other (3.29%). The top twenty browser languages represented in the data are shown in Table 3, with full tables of browser languages in the supplemental online data. Generally, this pattern matched the demographics of the study, as well as the targeted languages. More participants had their browser set in English compared to the indicated native language, see Table 3.

**Trial level data**

Each language is saved in separate files, and supplemental files and code are provided to merge trials across concepts and pairings (i.e., CAT [English] → KATZE [German] → GATTA [Italian]). If a participant left the study early (e.g., Internet disconnection, computer crash, closes the study), the data past this point in the study was not recorded, and therefore, the trial level data represented all trials displayed during the experiment. Participants were expected to incorrectly answer trials, and these trials were marked for exclusion. All timeout trials were marked as missing values in the final data. No missing values were imputed.

Trials were marked for exclusion if they were under the minimum response latency of less than 160 ms (i.e., all trials will be presented in the trial level data for openness, but these will be excluded for analysis and calculations listed below). Further, *labjs* automatically codes timeout data with a special marker (i.e., data ended on response or timeout as a column), which excludes trials over 3000 ms as our maximum response latency. However, due to differences in browser/screen refresh rates, some trials were answered with response latencies over 3000 ms when a participant made a key press at the very end of the trial before timeout. Given our pre-registered exclusion rules, these were also marked for exclusion.

The response latencies from each participant’s session were then *z*-scored in line with recommendations from Faust et al.53 We did not collect enough data to note if a person took the experiment multiple times for privacy reasons, but as these are considered different sessions, the recommended *z*-score procedure should control for participant variability at this level. Therefore, repeated participation was not detrimental to data collection. Finally, participants’ overall proportion of correct answers was calculated, and participants who did not correctly answer at least 80% of 100 minimum trials seen were marked for exclusion for item data, priming data, and analysis. The average error in the Semantic Priming Project ranged from 4% to 5%, and this criterion was chosen to include participants who were focused on the task. Please see summary of exclusion criteria for all exclusions, which are marked separately in the provided data files. Additionally, as noted above, accuracy was defined multiple ways by the lead team, and therefore, both criteria are provided.

The study took approximately *M* = 26.40 (*SD* = 303.61) minutes to complete. If a participant’s computer went to sleep during the study, and then they returned to the study (e.g., to close the browser), the last timestamp would include the final time the study was open. Therefore, the median of the study is likely more representative, *Mdn* = 17.88 minutes. The participants used in analysis completed the study in *M* = 24.14 (*SD* = 296.83), *Mdn* = 17.97 minutes. Table 4 includes the number of trials and accuracy for each language and total for all participants and analyzed participants. The mean *Z-*scores for all trials, regardless of item or related/unrelated condition, are presented in the summary files online. The analyses averaged over item statistics is presented below.

**Table 4.**

**Total Numbers of LDT Trials and Accuracy by Word-Nonword Trial**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | All Participants | | Analyzed Participants | | All Participants | | Analyzed Participants | |
| Language | Total Nonword Trials | Total Word Trials | Total Nonword Trials | Total Word Trials | Accuracy Nonword | Accuracy Word | Accuracy Nonword | Accuracy Word |
| Czech | 446,465 | 447,172 | 396,459 | 397,150 | 0.91 | 0.95 | 0.94 | 0.97 |
| Danish | 344,582 | 345,061 | 311,920 | 312,264 | 0.89 | 0.94 | 0.92 | 0.95 |
| English | 2,245,604 | 2,252,266 | 1,961,546 | 1,968,289 | 0.87 | 0.94 | 0.91 | 0.95 |
| French | 349,804 | 350,247 | 331,078 | 331,316 | 0.93 | 0.96 | 0.94 | 0.96 |
| German | 1,090,365 | 1,090,615 | 1,022,547 | 1,022,866 | 0.92 | 0.95 | 0.93 | 0.96 |
| Greek | 280,819 | 281,564 | 264,274 | 264,915 | 0.93 | 0.94 | 0.95 | 0.95 |
| Hungarian | 310,186 | 309,954 | 279,322 | 279,126 | 0.91 | 0.93 | 0.94 | 0.94 |
| Italian | 442,736 | 443,774 | 420,132 | 420,889 | 0.94 | 0.96 | 0.95 | 0.96 |
| Japanese | 445,883 | 444,659 | 379,645 | 378,968 | 0.90 | 0.92 | 0.94 | 0.96 |
| Korean | 388,661 | 390,327 | 321,070 | 322,260 | 0.87 | 0.92 | 0.91 | 0.94 |
| Polish | 492,714 | 492,552 | 448,989 | 448,941 | 0.92 | 0.95 | 0.94 | 0.96 |
| Portuguese (Overall) | 495,485 | 495,373 | 456,065 | 456,166 | 0.89 | 0.95 | 0.91 | 0.96 |
| Romanian | 304,296 | 304,271 | 278,125 | 278,246 | 0.92 | 0.96 | 0.93 | 0.97 |
| Russian | 795,078 | 793,816 | 652,446 | 652,149 | 0.91 | 0.93 | 0.95 | 0.96 |
| Serbian | 285,389 | 285,498 | 262,660 | 262,664 | 0.92 | 0.95 | 0.93 | 0.96 |
| Simplified Chinese | 327,479 | 327,869 | 274,613 | 274,870 | 0.88 | 0.93 | 0.92 | 0.95 |
| Spanish | 586,901 | 586,488 | 556,113 | 555,740 | 0.92 | 0.95 | 0.93 | 0.96 |
| Turkish | 898,853 | 897,783 | 788,613 | 788,008 | 0.91 | 0.94 | 0.94 | 0.95 |
| Overall | 10,531,300 | 10,539,289 | 9,405,617 | 9,414,827 | 0.90 | 0.94 | 0.93 | 0.96 |

**Item level data**

The item data files can be matched with lexical information about all stimuli calculated from the OpenSubtitles20 and subs2vec46 projects using the *semanticprimeR* package (see supplemental for tutorial)85. The descriptive statistics calculated from the trial level data is included separated by language for each item: mean response latency, average standardized response latency, sample size, standard errors of response latencies, and accuracy rate. No data was excluded for being a potential outlier; however, we recommended a cut-off criterion for absolute value z-score outliers at 2.5 and 3.0, and we calculated these same statistics with those subsets of trials excluded. For all real words, the age of acquisition, imageability, concreteness, valence, dominance, arousal, and familiarity values can be merged with the item files. These values do not exist for non-words.

Tables 5 and 6 show the item statistics for average item sample size, average *Z*-scored response time, average *SE* for the *Z*-scored response latencies separated by word type and language. These calculations exclude both participants and trials after first averaging over each item (see exclusions above).

**Table 5.**

**Total Number of Unique Trials and Average Trials Per Item**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | All Trials | | Z < 2.5 | | Z < 3.0 | |
| Language | N Unique NW | N Unique W | M Trials NW | M Trials W | M Trials NW | M Trials W | M Trials NW | M Trials W |
| Brazilian Portuguese | 1,946 | 1,956 | 180.75 | 208.70 | 172.05 | 205.65 | 175.09 | 206.71 |
| Czech | 1,981 | 1,969 | 185.05 | 193.07 | 176.56 | 190.18 | 179.43 | 191.16 |
| Danish | 1,957 | 1,954 | 145.73 | 151.12 | 138.84 | 148.48 | 141.14 | 149.35 |
| English | 1,978 | 2,000 | 889.16 | 932.03 | 851.22 | 915.36 | 863.12 | 920.45 |
| French | 1,976 | 1,936 | 156.07 | 163.90 | 149.51 | 161.36 | 151.66 | 162.17 |
| German | 1,957 | 1,946 | 484.48 | 499.54 | 463.33 | 491.11 | 470.60 | 493.85 |
| Greek | 1,949 | 1,924 | 120.51 | 130.60 | 115.71 | 127.85 | 117.35 | 128.73 |
| Hungarian | 1,936 | 1,924 | 134.59 | 135.65 | 129.57 | 132.80 | 131.25 | 133.73 |
| Italian | 1,992 | 1,991 | 197.80 | 201.52 | 189.60 | 198.37 | 192.38 | 199.40 |
| Japanese | 1,989 | 1,953 | 177.24 | 183.63 | 170.69 | 179.39 | 172.89 | 180.63 |
| Korean | 1,857 | 1,938 | 154.96 | 154.65 | 149.13 | 151.40 | 150.93 | 152.33 |
| Polish | 1,985 | 1,949 | 211.16 | 219.87 | 202.23 | 216.29 | 205.28 | 217.44 |
| Portuguese | 1,965 | 1,956 | 183.61 | 209.07 | 174.44 | 206.09 | 177.64 | 207.10 |
| Romanian | 1,966 | 1,952 | 130.63 | 136.68 | 124.39 | 134.80 | 126.59 | 135.45 |
| Russian | 1,996 | 1,998 | 306.39 | 309.55 | 294.25 | 303.59 | 298.45 | 305.57 |
| Serbian | 1,960 | 1,957 | 123.51 | 128.09 | 117.67 | 126.54 | 120.04 | 127.15 |
| Simplified Chinese | 1,993 | 1,842 | 126.09 | 140.62 | 120.99 | 137.76 | 122.60 | 138.63 |
| Spanish | 1,989 | 1,941 | 259.36 | 273.35 | 247.93 | 269.43 | 251.68 | 270.71 |
| Turkish | 1,866 | 1,929 | 391.22 | 383.96 | 375.84 | 376.19 | 380.81 | 378.57 |
| Overall | 37,238 | 37,015 | 239.97 | 251.59 | 229.74 | 247.20 | 233.16 | 248.60 |

*Note*. N = number, M = mean, NW = nonwords, W = words.

**Table 6.**

**Mean, Standard Error for Nonword and Word Trials by Language**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | All Trials | | | | Z < 2.5 | | | | Z < 3.0 | | | |
| Language | M Z NW | M Z W | SE Z NW | SE Z W | M Z NW | M Z W | SE Z NW | SE Z W | M Z NW | M Z W | SE Z NW | SE Z W |
| Brazilian Portuguese | 0.29 | -0.26 | 0.08 | 0.06 | 0.12 | -0.32 | 0.06 | 0.04 | 0.17 | -0.30 | 0.06 | 0.05 |
| Czech | 0.31 | -0.25 | 0.07 | 0.06 | 0.15 | -0.31 | 0.05 | 0.04 | 0.19 | -0.30 | 0.06 | 0.05 |
| Danish | 0.28 | -0.22 | 0.08 | 0.07 | 0.11 | -0.29 | 0.06 | 0.05 | 0.15 | -0.27 | 0.06 | 0.05 |
| English | 0.26 | -0.20 | 0.03 | 0.03 | 0.09 | -0.28 | 0.02 | 0.02 | 0.13 | -0.26 | 0.03 | 0.02 |
| French | 0.27 | -0.23 | 0.08 | 0.06 | 0.12 | -0.30 | 0.06 | 0.05 | 0.16 | -0.28 | 0.06 | 0.05 |
| German | 0.26 | -0.20 | 0.04 | 0.04 | 0.11 | -0.27 | 0.03 | 0.03 | 0.15 | -0.25 | 0.03 | 0.03 |
| Greek | 0.20 | -0.14 | 0.09 | 0.07 | 0.05 | -0.22 | 0.07 | 0.06 | 0.09 | -0.20 | 0.07 | 0.06 |
| Hungarian | 0.18 | -0.13 | 0.08 | 0.07 | 0.05 | -0.22 | 0.06 | 0.06 | 0.08 | -0.20 | 0.06 | 0.06 |
| Italian | 0.26 | -0.24 | 0.07 | 0.06 | 0.12 | -0.31 | 0.05 | 0.04 | 0.15 | -0.29 | 0.05 | 0.05 |
| Japanese | 0.17 | -0.13 | 0.07 | 0.06 | 0.04 | -0.23 | 0.05 | 0.05 | 0.07 | -0.21 | 0.06 | 0.05 |
| Korean | 0.23 | -0.16 | 0.08 | 0.07 | 0.08 | -0.26 | 0.06 | 0.05 | 0.11 | -0.24 | 0.06 | 0.05 |
| Polish | 0.27 | -0.23 | 0.07 | 0.05 | 0.12 | -0.29 | 0.05 | 0.04 | 0.15 | -0.28 | 0.05 | 0.04 |
| Portuguese | 0.35 | -0.27 | 0.08 | 0.05 | 0.17 | -0.33 | 0.06 | 0.04 | 0.22 | -0.31 | 0.06 | 0.04 |
| Romanian | 0.32 | -0.28 | 0.09 | 0.07 | 0.16 | -0.33 | 0.06 | 0.05 | 0.20 | -0.32 | 0.07 | 0.05 |
| Russian | 0.21 | -0.22 | 0.05 | 0.05 | 0.08 | -0.29 | 0.04 | 0.04 | 0.11 | -0.27 | 0.04 | 0.04 |
| Serbian | 0.36 | -0.33 | 0.09 | 0.06 | 0.22 | -0.37 | 0.07 | 0.05 | 0.27 | -0.36 | 0.07 | 0.06 |
| Simplified Chinese | 0.23 | -0.18 | 0.09 | 0.07 | 0.08 | -0.27 | 0.06 | 0.05 | 0.11 | -0.25 | 0.07 | 0.06 |
| Spanish | 0.29 | -0.25 | 0.06 | 0.05 | 0.13 | -0.31 | 0.05 | 0.04 | 0.17 | -0.30 | 0.05 | 0.04 |
| Turkish | 0.22 | -0.17 | 0.05 | 0.04 | 0.07 | -0.25 | 0.04 | 0.03 | 0.10 | -0.24 | 0.04 | 0.03 |
| Overall | 0.26 | -0.21 | 0.07 | 0.06 | 0.11 | -0.29 | 0.05 | 0.04 | 0.15 | -0.27 | 0.06 | 0.05 |

*Note*. M = mean, SE = standard error, NW = nonwords, W = words.

**Priming level data**

In a separate file, we prepared information about priming results which includes the target word, average response latencies, averaged *Z*-scored response latencies, sample sizes, standard errors, and priming response latency. For each item, priming was defined as the average *Z*-scored response latency when presented in the unrelated minus the related condition. Therefore, the timing for DOG-CAT would be subtracted from BUS-CAT to indicate priming for the word CAT. The similarity scores calculated during stimuli selection are provided for merging, as well as other popular measures of similarity if they are available in that language. For example, semantic feature overlap norms are also available in Italian86, German87, Spanish23, and Dutch88. The overall priming averages by language are shown in Figure 4 as part of Hypothesis 1 and 2.

Item reliability was calculated by randomly splitting priming trials into two halves, calculating *Z*-score priming for each half, and correlating those scores by item. Person level reliability was calculated in a similar fashion by splitting participant related-unrelated trials in half and calculating priming as the average unrelated *Z*-scored response latency minus the related *Z*-scored response latency, and correlating the two priming scores. The Spearman-Brown prophecy formula was applied to these correlations to calculate total reliability. This procedure was repeated 100 times, and average scores were calculated. Table 7 shows the reliability scores for person and items. The average reliability for items was .56 and for people was .08.

**Table 7.**

**Item and Person Reliability for Priming Scores**

|  |  |  |
| --- | --- | --- |
| Language | Item Reliability | Person Reliability |
| Brazilian Portuguese | 0.50 | 0.09 |
| Czech | 0.54 | -0.02 |
| Danish | 0.48 | 0.12 |
| English | 0.72 | 0.06 |
| French | 0.47 | -0.09 |
| German | 0.67 | 0.00 |
| Green | 0.42 | 0.00 |
| Hungarian | 0.47 | 0.04 |
| Italian | 0.55 | 0.11 |
| Japanese | 0.75 | 0.17 |
| Korean | 0.69 | 0.02 |
| Polish | 0.51 | 0.15 |
| Portuguese | 0.44 | 0.13 |
| Romanian | 0.48 | -0.04 |
| Russian | 0.71 | 0.17 |
| Serbian | 0.42 | 0.16 |
| Simplified Chinese | 0.52 | 0.09 |
| Spanish | 0.63 | 0.08 |
| Turkish | 0.66 | 0.01 |

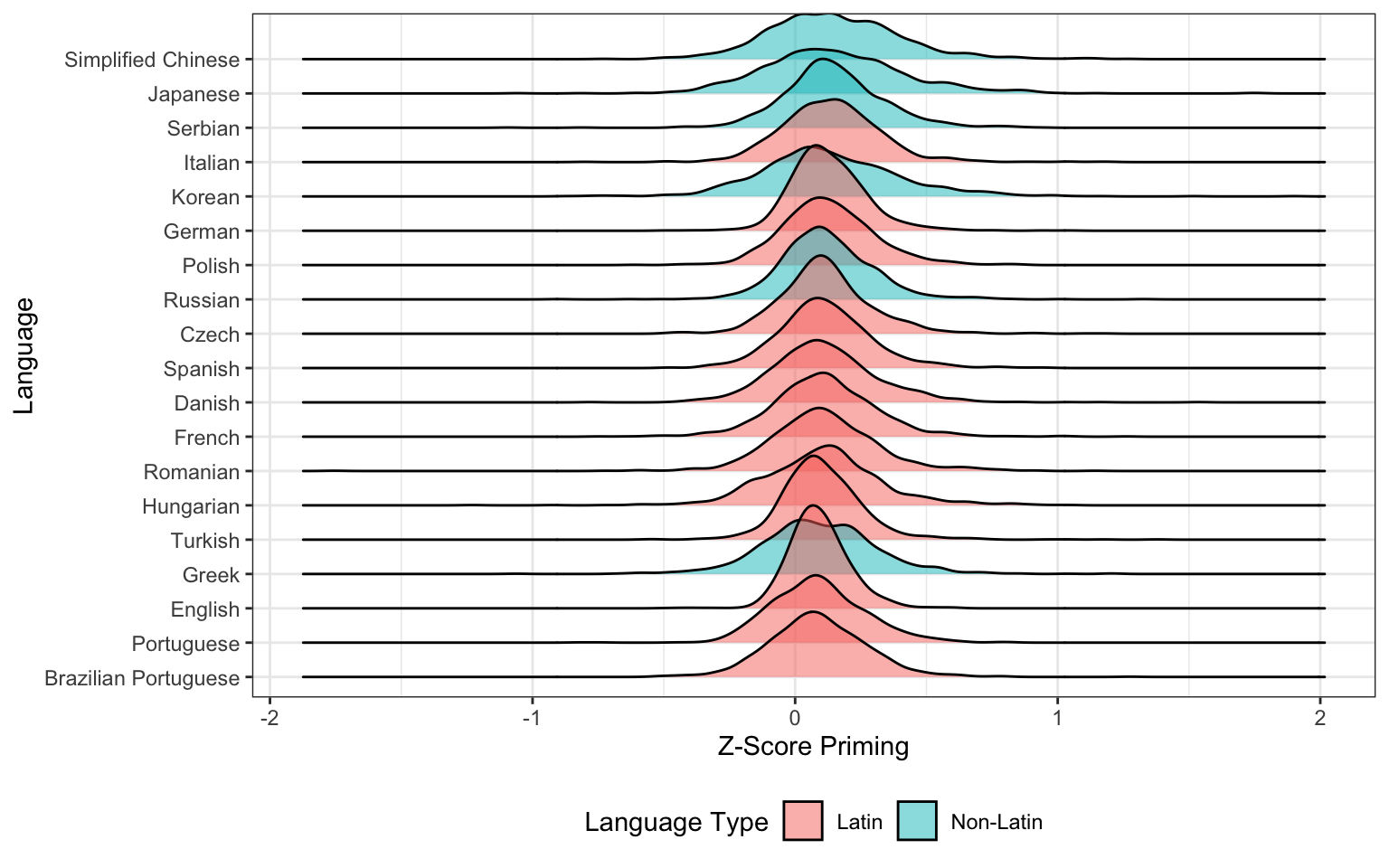
**Hypothesis 1**

Hypothesis 1 predicted finding semantic facilitation wherein the response latencies for related targets would be faster than unrelated targets as shown in Table 1. Hypothesis 1 was tested by fitting an intercept-only regression model using the *z*-scored priming response latency as the dependent variable. The priming response latency was calculated by taking the average of the unrelated pair *z*-scored response latency minus the average related pair response latency within each item by language. Therefore, values that are positive and greater than zero (e.g., > 0.0001) indicate priming because the related pair had a faster response latency than the unrelated pair. The intercept and its 95% confidence interval represent the grand mean of the priming effect across all languages.

**Deviations**

In cases in which the target word is repeated due to language translation, we created pairs of translations (i.e., cue-target-related1, cue-target-unrelated1, cue-target-related2, cue-target-unrelated2) to ensure each pair only gets subtracted once. For example, if SPOON-CHEESE and TREE-CHEESE (unrelated) needed to be paired with MOUSE-CHEESE and CHEDDAR-CHEESE (related), we ensured each version was only combined once: SPOON-CHEESE minus MOUSE-CHEESE and TREE-CHEESE minus CHEDDAR-CHEESE. For Korean, the extra unrelated pairs accidentally implemented were excluded in the priming calculation. When the unrelated target was repeated multiple times with no matching related target (i.e., 1 related target, 3 unrelated targets), we picked the lowest cosine unrelated target pair to be the comparison condition and discarded the rest of the unrelated pairs. This procedure also allowed us to control the slightly higher cosine values found (and noted above) for unrelated pairs in Korean.

Overall priming was *b0* = 0.117, *SE* = 0.001, *95%CI*[0.114, 0.120]. Hypothesis 1 from our pre-registration was that the lower limit of the confidence interval is greater than zero (i.e., a directional comparison). This process was repeated for average priming scores calculated without trials that were marked as 2.5 *Z*-score outliers and 3.0 *Z*-score outliers separately. These results were consistent with overall priming: *b0Z2.5* = 0.104, *SE* = 0.001, *95%CI*[0.101, 0.106], and *b0Z3.0* = 0.107, *SE* = 0.001, *95%CI*[0.104, 0.109]. Figure 4 denotes the distribution of the average item *Z*-score effects, ordered by the size of the overall priming effect for each language. The distributions of the priming scores are very similar with long tails and roughly similar shapes (albeit more variance in some languages).



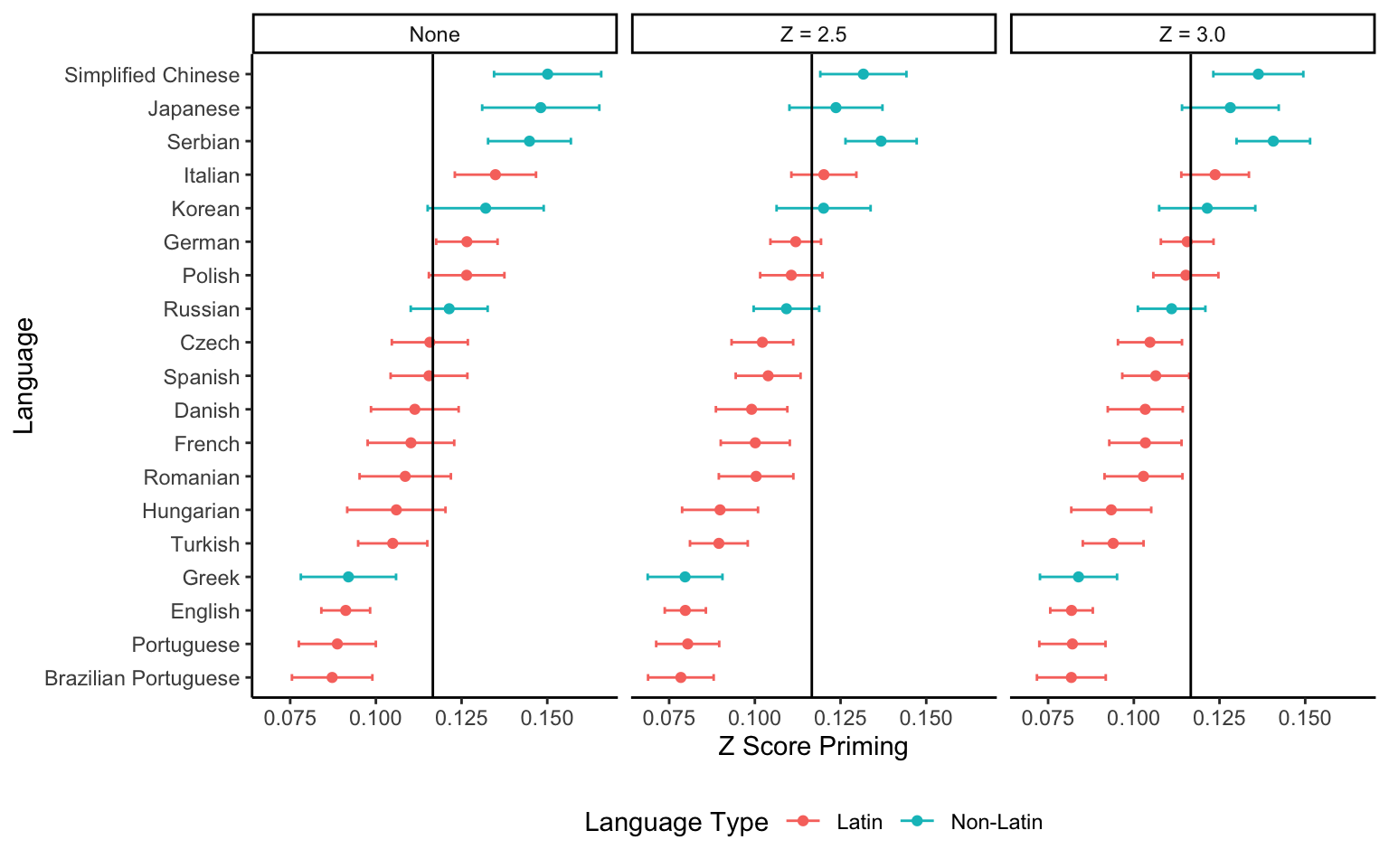
**Figure 4***.* Distribution of average priming effects for languages that met the minimum sample size criteria. Order of languages based on their average priming effect.

**Hypothesis 2**

Hypothesis 2 explored the extent to which these semantic priming effects vary across languages. Therefore, we calculated a random effects model using the *nlme*89 package in *R* wherein the random intercept of language was added to the overall intercept only model for Hypothesis 1. The addition of this parameter improved model fit from *AICHyp1* = -6,613.93 to

*AICHyp2*= -6,711.77, supporting significant heterogeneity as the AIC for the random effects model is two points or more less than the AIC for the intercept-only model47. The standard deviation of the random effect was 0.02, *95% CI*[0.01, 0.03]. The pseudo-*R2* for the model was .00890. The random effect was useful in both *Z*-score 2.5 and 3.0 models: *AICHyp1Z2.5* = -14,469.54 versus *AICHyp2Z2.5* = -14,604.55 and *AICHyp1Z3.0* = -12,977.97 versus *AICHyp2Z3.0* = -13,104.04. The random effect sizes were similar to the overall model: Z2.5 = 0.02, *95% CI*[0.01, 0.02], Z3.0 = 0.02, *95% CI*[0.01, 0.03].

Figure 5 portrays the forest plot for the average priming effects by language, ordered by the size of the effect without removal of outliers. The global priming average is presented on each facet to show how the priming effect changes based on the removal of outliers. In nearly all languages, the priming effect decreases slightly with the removal of outliers with the exception of Serbian. This figure also shows that the priming effect does vary by language, as supported by the results from Hypothesis 2, but that the effect is likely small, given pseudo-*R2* was < .01.



**Figure 5**. Forest plot of average priming effects for each language ordered by priming average when no outliers are removed. The vertical line indicates the global average priming effect with no outliers removed.

**Discussion**

This study represents the largest cross-linguistic study on semantic priming to date with data collection in 30 languages on a set of coordinated stimuli. Using computational models of word embeddings and expanded linguistic corpora, we selected a stimulus set that covered semantic similarity across languages, rather than in a single language to be translated into others. Using a single stream lexical decision task, more than 21 million trials were collected using an adaptive stimulus presentation algorithm that shifted data collection toward uncertainty after a minimum number of trials. Data collection minimums were completed for 19 languages/dialects with more than 700 participants in each language with coverage in both Latin and non-Latin based scripts. Given the large proportion of published linguistics that is still WEIRD91, we provide a diversity of data that can be reused to examine new hypotheses, control stimuli in new studies, and create cross linguistic comparisons for previously found results.

In the 19 analyzed languages, we demonstrated consistent non-zero priming effects ranging from *Z* = 0.09 to *Z* = 0.15, and this effect is resilient to the removal of strong priming pairs with high *Z*-scores such as ROMEO-JULIET, GOLDEN-SILVER, MENTAL-EMOTIONAL, and BLIND-DEAF (i.e., highest positive *Z* priming scores across all languages, translated into their English counterparts). The *Z*-score removal also eliminates strong negative pairs, such as RESCUE-SAVE, FASHIONABLE-ELEGANT, and POSITION-STATUS. The English dataset provided one of the lowest priming averages, *Z* = 0.09, even with an average cosine relatedness of 0.55 (*SD* = 0.11, min = 0.22, max = 0.90). For comparison, the Semantic Priming Project21’s results demonstrated higher priming values when SOAs were short (200ms; *Z* = 0.21 for first associates, *Z* = 0.14 for other associates), but comparable values for longer SOAs (1200ms; *Z* = 0.16 for first associates, *Z* = 0.10 for other associates). Given that participants made lexical decisions on target words in our study, the results should most closely match the longer SOA conditions, and our results generally align with results for other associates. The results also demonstrate higher item reliability estimates than some estimates previously shown (-.06-.4392, .0440, .17-.3338) and more in line with other findings (.66 standardized LDT39). The person reliability estimates are considerably lower than previous examinations of the Semantic Priming Project for first associates (.21-.27) but somewhat similar to results for other associates (.07-.0893). The large sample sizes for this project likely boosted reliability results for item level reliability, as the largest samples show some of the strongest reliability coefficients.

Our secondary hypothesis examined the (in)consistency of priming effects across languages supporting small but non-zero differences in levels of priming across languages. One key takeaway from Figure 4 is the relatively similar distributions found for all languages with what appears to be slightly more platykurtic distributions for non-Latin languages. While Portuguese and Simplified Chinese show clearly non-overlapping confidence intervals in Figure 5, it is a bit surprising how all means are within previous (English) *Z*-score estimates for priming and how remarkably comparable the results are for each analyzed language. Given the potential differences in translation, script, processing, culture, and more, this result points to a universal cognitive mechanism for semantic priming, albeit this study was not designed to answer the long standing automatic versus controlled processing debate2,75. With the wealth of data provided in this project, researchers may be able to begin to discern what variables influence differences found in the strength of priming effects at the language level, rather than within individual multilingual populations.

The limitations of this research include the necessity of picking a single design for semantic priming, but expand the available data to a new study type (i.e., the Semantic Priming Project and others have used short or masked priming, this study used single stream lexical decision). The study design does provide abundant data for all types of word processing analyses, but did not specifically target a single underlying cognitive mechanism for the explanation of priming effects. A few individual demographic variables are present to explore potential reasons for participant variability, while other studies provide more individual differences measures such as reading and vocabulary measures. This limitation did allow the study to be conducted easily in many geopolitical reasons, as institutional review boards vary widely in their approval of studies that collect identifying measures, especially with overseas data management (i.e., they would rather the data be collected and stored locally). Not all languages translated finished initial data collection; however, the data is available for use until minimum data requirements can be collected for updated analyses, and ideally, new low resource languages would be added to new publications of the dataset.

In summary, our results support semantic priming (and its variability) across languages (and cultural contexts, as multiple languages were collected in different geopolitical regions) using a controlled stimuli set that matched target words. Future directions can explore how to explain the variability found within individuals, items, and languages, using the provided *semanticprimeR* package to merge datasets across other psycholinguistic variables. This study demonstrates the effectiveness of large-scale team collaboration to answer cross-linguistic questions, as well as provide resources for future reuse that are more “complete” (i.e., less missing values) than individual lab contributions17. Although linguistics is largely still WEIRD, big team projects can begin to tackle the sampling bias generalizability problems within the field43,91,94 using grassroots networks like the Psychological Science Accelerator14 and the upcoming ManyLanguages community95.

**Protocol Registration**

Our pre-registration for this report can be found at<https://osf.io/u5bp6> (updated 5/31/2022).

**Data Availability**

All raw and processed data will be available for download from GitHub: <https://github.com/SemanticPriming/SPAML> or Zenodo: <https://zenodo.org/records/10888833>.

**Code Availability**

All code used for study creation and delivery, data processing, and analyses are available on OSF (https://osf.io/wrpj4/) and GitHub (https://github.com/SemanticPriming/SPAML).

References

1. Meyer, D. E. & Schvaneveldt, R. W. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* **90**, 227–234 (1971).

2. McNamara, T. P. *Semantic Priming*. (Psychology Press, 2005). doi:10.4324/9780203338001.

3. Mandera, P., Keuleers, E. & Brysbaert, M. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language* **92**, 57–78 (2017).

4. Cree, G. S. & Armstrong, B. C. Computational Models of Semantic Memory. in *The Cambridge Handbook of Psycholinguistics* (eds. Spivey, M., McRae, K. & Joanisse, M.) 259–282 (Cambridge University Press, 2012). doi:10.1017/CBO9781139029377.014.

5. McRae, K. & Jones, M. *Semantic Memory*. (Oxford University Press, 2013). doi:10.1093/oxfordhb/9780195376746.013.0014.

6. Rogers, T. T. Computational Models of Semantic Memory. in *The Cambridge Handbook of Computational Psychology* (ed. Sun, R.) 226–266 (Cambridge University Press, 2001). doi:10.1017/CBO9780511816772.012.

7. Frings, C., Schneider, K. K. & Fox, E. The negative priming paradigm: An update and implications for selective attention. *Psychon Bull Rev* **22**, 1577–1597 (2015).

8. Spruyt, A., De Houwer, J., Everaert, T. & Hermans, D. Unconscious semantic activation depends on feature-specific attention allocation. *Cognition* **122**, 91–95 (2012).

9. McDonough, K. & Trofimovich, P. *Using Priming Methods in Second Language Research*. (Routledge, 2011). doi:10.4324/9780203880944.

10. Singh, L. One World, Two Languages: Cross-Language Semantic Priming in Bilingual Toddlers. *Child Dev* **85**, 755–766 (2014).

11. Kiefer, M. *et al.* Neuro-cognitive mechanisms of conscious and unconscious visual perception: From a plethora of phenomena to general principles. *Advances in Cognitive Psychology* **7**, 55–67 (2011).

12. Steinhauer, K., Royle, P., Drury, J. E. & Fromont, L. A. The priming of priming: Evidence that the N400 reflects context-dependent post-retrieval word integration in working memory. *Neuroscience Letters* **651**, 192–197 (2017).

13. Liu, B., Wu, G., Meng, X. & Dang, J. Correlation between prime duration and semantic priming effect: Evidence from N400 effect. *Neuroscience* **238**, 319–326 (2013).

14. Moshontz, H. *et al.* The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science* **1**, 501–515 (2018).

15. Lucas, M. Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review* **7**, 618–630 (2000).

16. Hutchison, K. A. Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review* **10**, 785–813 (2003).

17. Buchanan, E. M., Valentine, K. D. & Maxwell, N. P. LAB: Linguistic Annotated Bibliography – a searchable portal for normed database information. *Behav Res* **51**, 1878–1888 (2019).

18. New, B., Brysbaert, M., Veronis, J. & Pallier, C. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* **28**, 661–677 (2007).

19. Gimenes, M. & New, B. Worldlex: Twitter and blog word frequencies for 66 languages. *Behav Res* **48**, 963–972 (2016).

20. Lison, P. & Tiedemann, J. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. (2016).

21. Hutchison, K. A. *et al.* The semantic priming project. *Behav Res* **45**, 1099–1114 (2013).

22. Balota, D. A. *et al.* The English Lexicon Project. *Behavior Research Methods* **39**, 445–459 (2007).

23. Aguasvivas, J. A. *et al.* SPALEX: A Spanish Lexical Decision Database From a Massive Online Data Collection. *Front. Psychol.* **9**, 2156 (2018).

24. Bradley, M. M. & Lang, P. J. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* **25**, 49–59 (1994).

25. Warriner, A. B., Kuperman, V. & Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav Res* **45**, 1191–1207 (2013).

26. Bradley, M. M. & Lang, P. J. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*. (1999).

27. Brysbaert, M., Warriner, A. B. & Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav Res* **46**, 904–911 (2014).

28. Stadthagen-Gonzalez, H. & Davis, C. J. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods* **38**, 598–605 (2006).

29. Kerr, N. L. HARKing: Hypothesizing After the Results are Known. *Pers Soc Psychol Rev* **2**, 196–217 (1998).

30. Cree, G. S., McRae, K. & McNorgan, C. An Attractor Model of Lexical Conceptual Processing: Simulating Semantic Priming. *Cognitive Science* **23**, 371–414 (1999).

31. Zannino, G. D., Perri, R., Pasqualetti, P., Caltagirone, C. & Carlesimo, G. A. Analysis of the semantic representations of living and nonliving concepts: A normative study. *Cognitive Neuropsychology* **23**, 515–540 (2006).

32. Buchanan, E. M., Valentine, K. D. & Maxwell, N. P. English semantic feature production norms: An extended database of 4436 concepts. *Behav Res* **51**, 1849–1863 (2019).

33. De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M. & Storms, G. The “Small World of Words” English word association norms for over 12,000 cue words. *Behav Res* **51**, 987–1006 (2019).

34. Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* **36**, 402–407 (2004).

35. Landauer, T. K., Foltz, P. W. & Laham, D. An introduction to latent semantic analysis. *Discourse Processes* **25**, 259–284 (1998).

36. Landauer, T. K. & Dumais, S. T. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* **104**, 211–240 (1997).

37. Lund, K. & Burgess, C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* **28**, 203–208 (1996).

38. Heyman, T., Hutchison, K. A. & Storms, G. Uncovering underlying processes of semantic priming by correlating item-level effects. *Psychon Bull Rev* **23**, 540–547 (2016).

39. Hutchison, K. A., Balota, D. A., Cortese, M. J. & Watson, J. M. Predicting Semantic Priming at the Item Level. *Quarterly Journal of Experimental Psychology* **61**, 1036–1066 (2008).

40. Heyman, T., Bruninx, A., Hutchison, K. A. & Storms, G. The (un)reliability of item-level semantic priming effects. *Behav Res* **50**, 2173–2183 (2018).

41. Perea, M., Duñabeitia, J. A. & Carreiras, M. Masked associative/semantic priming effects across languages with highly proficient bilinguals. *Journal of Memory and Language* **58**, 916–930 (2008).

42. Guasch, M., Sánchez-Casas, R., Ferré, P. & García-Albea, J. E. Effects of the degree of meaning similarity on cross-language semantic priming in highly proficient bilinguals. *Journal of Cognitive Psychology* **23**, 942–961 (2011).

43. Levisen, C. Biases we live by: Anglocentrism in linguistics and cognitive sciences. *Language Sciences* **76**, 101173 (2019).

44. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. in *Advances in neural information processing systems* 3111–3119 (2013).

45. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]* (2013).

46. van Paridon, J. & Thompson, B. subs2vec: Word embeddings from subtitles in 55 languages. *Behav Res* **53**, 629–655 (2021).

47. Burnham, K. P. & Anderson, D. R. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* **33**, 261–304 (2004).

48. Tzelgov, J. & Eben-ezra, S. Components of the between-language semantic priming effect. *European Journal of Cognitive Psychology* **4**, 253–272 (1992).

49. Kirsner, K., Smith, M. C., Lockhart, R. S., King, M. L. & Jain, M. The bilingual lexicon: Language-specific units in an integrated network. *Journal of Verbal Learning and Verbal Behavior* **23**, 519–539 (1984).

50. Kelley, K. Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods* **39**, 755–766 (2007).

51. Kelley, K., Darku, F. B. & Chattopadhyay, B. Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods* **23**, 226–243 (2018).

52. Maxwell, S. E., Kelley, K. & Rausch, J. R. Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annu. Rev. Psychol.* **59**, 537–563 (2008).

53. Faust, M. E., Balota, D. A., Spieler, D. H. & Ferraro, F. R. Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin* **125**, 777–799 (1999).

54. Overall, J. E. & Woodward, J. A. Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin* **82**, 85–86 (1975).

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

56. Tse, C.-S. *et al.* The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behav Res* **49**, 1503–1519 (2017).

57. Michalke, M. sylly: Hyphenation and Syllable Counting for Text Analysis. (2020).

58. De Deyne, S., Navarro, D. J. & Storms, G. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behav Res* **45**, 480–498 (2013).

59. Brysbaert, M. & New, B. 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**, 977–990 (2009).

60. van Heuven, W. J. B., Mandera, P., Keuleers, E. & Brysbaert, M. Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology* **67**, 1176–1190 (2014).

61. Keuleers, E., Brysbaert, M. & New, B. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods* **42**, 643–650 (2010).

62. Cai, Q. & Brysbaert, M. SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. *PLoS ONE* **5**, e10729 (2010).

63. Brysbaert, M. *et al.* The Word Frequency Effect: A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. *Experimental Psychology* **58**, 412–424 (2011).

64. Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J. & Carreiras, M. Subtitle-Based Word Frequencies as the Best Estimate of Reading Behavior: The Case of Greek. *Front. Psychology* **1**, (2010).

65. Mandera, P., Keuleers, E., Wodniecka, Z. & Brysbaert, M. Subtlex-pl: subtitle-based word frequency estimates for Polish. *Behav Res* **47**, 471–483 (2015).

66. Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A. & Carreiras, M. EsPal: One-stop shopping for Spanish word properties. *Behav Res* **45**, 1246–1258 (2013).

67. Wijffels, J. *et al.* udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the ‘UDPipe’ ‘NLP’ Toolkit. (2021).

68. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

69. Benoit, K., Muhr, D. & Watanabe, K. stopwords: Multilingual Stopword Lists. (2021).

70. Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J. & Hilbig, B. E. *Lab.Js: A Free, Open, Online Study Builder*. https://osf.io/fqr49 (2019) doi:10.31234/osf.io/fqr49.

71. Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. & Hilbig, B. E. Who said browser-based experiments can’t have proper timing? Implementing accurate presentation and response timing in browser. (2018).

72. Hilbig, B. E. Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behav Res* **48**, 1718–1724 (2016).

73. Buchanan, E. *et al.* SemanticPriming/SPAML: SPAML v1 Data Release. [object Object] https://doi.org/10.5281/ZENODO.10888833 (2024).

74. Proctor, R. W. & Cho, Y. S. Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin* **132**, 416–442 (2006).

75. Neely, J. H., Keefe, D. E. & Ross, K. L. Semantic priming in the lexical decision task: Roles of prospective prime-generated expectancies and retrospective semantic matching. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **15**, 1003–1019 (1989).

76. Shelton, J. R. & Martin, R. C. How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**, 1191–1210 (1992).

77. Johnston, R. A. & Barry, C. Age of acquisition and lexical processing. *Visual Cognition* **13**, 789–845 (2006).

78. Ghyselinck, M., Lewis, M. B. & Brysbaert, M. Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica* **115**, 43–67 (2004).

79. Juhasz, B. J. Age-of-Acquisition Effects in Word and Picture Identification. *Psychological Bulletin* **131**, 684–712 (2005).

80. Brysbaert, M. & Ellis, A. W. Aphasia and age of acquisition: are early-learned words more resilient? *Aphasiology* **30**, 1240–1263 (2016).

81. Richardson, J. T. E. Imageability and concreteness. *Bull. Psychon. Soc.* **7**, 429–431 (1976).

82. Richardson, J. T. E. Concreteness and Imageability. *Quarterly Journal of Experimental Psychology* **27**, 235–249 (1975).

83. Paivio, A., Walsh, M. & Bons, T. Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**, 1196–1204 (1994).

84. Wilson, M. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers* **20**, 6–10 (1988).

85. Buchanan, E. SemanticPriming/semanticprimeR: semanticprimeR package. [object Object] https://doi.org/10.5281/ZENODO.10697999 (2024).

86. Montefinese, M., Ambrosini, E., Fairfield, B. & Mammarella, N. Semantic memory: A feature-based analysis and new norms for Italian. *Behav Res* **45**, 440–461 (2013).

87. Kremer, G. & Baroni, M. A set of semantic norms for German and Italian. *Behav Res* **43**, 97–109 (2011).

88. Ruts, W. *et al.* Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers* **36**, 506–515 (2004).

89. Pinheiro, J., Bates, D., Debroy, S., Sarkar, D. & Team, R. C. nlme: Linear and nonlinear mixed effects models. (2017).

90. Bartoń, K. MuMIn: Multi-Model Inference. (2020).

91. Bochynska, A. *et al.* Reproducible research practices and transparency across linguistics. Preprint at https://doi.org/10.31222/osf.io/rcews (2022).

92. Stolz, J. A., Besner, D. & Carr, T. H. Implications of measures of reliability for theories of priming: Activity in semantic memory is inherently noisy and uncoordinated. *Visual Cognition* **12**, 284–336 (2005).

93. Yap, M. J., Hutchison, K. A. & Tan, L. C. Individual differences in semantic priming performance: Insights from the semantic priming project. in *Big data in cognitive science* 203–226 (Routledge/Taylor & Francis Group, New York, NY, US, 2017).

94. Andringa, S. & Godfroid, A. Sampling Bias and the Problem of Generalizability in Applied Linguistics. *Annual Review of Applied Linguistics* **40**, 134–142 (2020).

95. ManyLanguages. https://many-languages.com/.

**Acknowledgements**

FUNDING TO BE ADDED FILL IN THE AUTHORSHIP SURVEY

**Author contributions**

TO BE ADDED CHECK YOUR TENZING

**Competing interests**

The authors declare no competing interests.

**Links to Supplemental Documents**

Please note: all files are synced to OSF through GitHub. We have also included the folder you can find files in if the GitHub add-on is not working on OSF. Since you cannot link directly to a folder on OSF storage, we also indicated where on OSF to find the folder.

Complete Files

* Open Science Framework: <https://osf.io/wrpj4/>
* GitHub: <https://github.com/SemanticPriming/SPAML>
* Zenodo: <https://zenodo.org/records/10888833>

Ethics

* Ethics Component OSF Link: <https://osf.io/ycn7z/>
* Ethics/Lab Table Summary: <https://osf.io/ty4hp>
  + GitHub: 06\_Analysis > supplemental

Power Analysis

* Power analysis code: <https://osf.io/v2y9e>
  + Github: 02\_Power

Method

* Materials separated by language:
  + OSF: 03\_Materials
  + Github: 03\_Materials
  + The readme explains the stimuli selection and creation procedure: <https://osf.io/mz7p4>
* *labjs* Scripts to recreate the experiment:
  + OSF: 04\_Procedure
  + Github: 04\_Procedure
* Language Table Information: <https://osf.io/y3dk7>
  + GitHub: 06\_Analysis > supplemental
* Deviation Guide: <https://osf.io/mwuv3>
  + GitHub: 06\_Analysis > supplemental
* Translation Information: <https://osf.io/vdme5>
  + Github: 03\_Materials readme

Data

* Zenodo: <https://zenodo.org/records/10888833>
* Data Release: <https://github.com/SemanticPriming/SPAML/releases>
* Data Processing Scripts:
  + OSF: 05\_Data > data\_processing
  + Github: 05\_Data > data\_processing
* Data Processing Checks/Summary: <https://osf.io/zye59>
  + Github: 05\_Data
* Codebooks:
  + OSF: 05\_Data > codebooks
  + Github: 05\_Data > codebooks
  + Codebook full data: <https://osf.io/xz6nk>
  + Codebook item data: <https://osf.io/5u9t6>
  + Codebook participant data: <https://osf.io/9a368>
  + Codebook priming trial level data: <https://osf.io/49nzq>
  + Codebook priming summarized level data: <https://osf.io/sx26p>
    - Summary table of the sample size calculations: <https://osf.io/kv6am>
  + Codebook trial data: <https://osf.io/s2kqd>
* semanticprimeR tutorial: <https://osf.io/yd8u4>

Analyses

* Scripts:
  + OSF: 06\_Analysis
  + Github: 06\_Analysis
  + Method: <https://osf.io/bqpk2>
  + Descriptive Statistics
    - Participants: <https://osf.io/vdgkr>
    - Trials: <https://osf.io/baem5>
    - Items: <https://osf.io/rvt8f>
    - Priming: <https://osf.io/m8kjv>
  + Hypothesis testing: <https://osf.io/rmkag>
  + Supplemental Meta Analysis: <https://osf.io/rke82>
    - Github: 06\_Analysis > supplemental
* Supplemental Tables/Summaries:
  + Note: summary of labs and languages also in this folder, but linked above
  + Github: 06\_Analysis > supplemental
  + Native Language:
    - Overall Native Language Frequency: <https://osf.io/ta6wf>
    - Analysis Participants Native Language Frequency: <https://osf.io/652h8>
    - Rescored Analysis Participants Native Language Frequency: <https://osf.io/b3y6r>
  + Browser Language:
    - Overall Browser Language Frequency: <https://osf.io/93kep>
    - Analysis Participants Browser Language Frequency: <https://osf.io/3yab7>
    - Rescored Analysis Participants Browser Language Frequency: <https://osf.io/adhbe>
  + Lab Reports:
    - Native Language by Lab: <https://osf.io/hnrgk>
    - Operating System by Lab: <https://osf.io/gud6v>
    - Web Browser by Lab: <https://osf.io/egk9w>
    - Language Locale by Lab: <https://osf.io/wt3xn>
  + Language Reports:
    - Native Language by Language: <https://osf.io/5b72x>
    - Operating System by Language: <https://osf.io/9dwqb>
    - Web Browser by Language: <https://osf.io/bn7uv>
    - Language Locale by Language: <https://osf.io/dyh4e>
  + Reliability data files:
    - Item Reliability: <https://osf.io/r4fym>
    - Person Reliability: <https://osf.io/jf28q>

Manuscript

* Pre-registration: <https://osf.io/u5bp6>
* Registered Report: <https://osf.io/preprints/osf/q4fjy>
* Tenzing chart: <https://osf.io/uv27t>
  + Github: 08\_Credit

1. Participant accuracy scores were calculated, and then the average of participant accuracy scores for each language were calculated. [↑](#footnote-ref-1)
2. We present the results combined when discussing trials or global information, but separated when examining item or priming level effects. [↑](#footnote-ref-2)
3. For each trial type combination: word-word unrelated, word-word related, two nonwords, nonword-word, word-nonword. Some combinations have higher probabilities of being selected, as the trial type is more frequent, but the individual trials were 1/1000 probability. [↑](#footnote-ref-3)