# Measuring the Semantic Priming Effect Across Many Languages

Erin M. Buchanan, Analytics, Harrisburg University, Harrisburg, PA, USA, ebuchanan@harrisburgu.edu*, @aggieerin
Kelly M. Cuccolo, Alma College, Alma, MI, USA
Nicholas Coles, Stanford University, CA, USA
Tom Heyman, Methodology and Statistics Unit, Institute of Psychology, Leiden University, The Netherlands
Aishwarya Iyer, Montfort College, KA, India
Neil Lewis Jr., Cornell University, NY, USA
Kim Peters, University of Exeter, UK
Niels van Berkel, Aalborg University, Denmark
Anna E. van 't Veer, Leiden University, Netherlands
Jack E. Taylor, School of Psychology and Neuroscience, University of Glasgow, UK
Maria Montefinese, IRCCS San Camillo Hospital, Venice, Italy
K. D. Valentine, Massachusetts General Hospital, Boston, MA, USA; Harvard Medical School, Boston, MA, USA
Nicholas P. Maxwell, University of Southern Mississippi, Hattiesburg, MS
Belgüzar Nilay Türkan, Pamukkale University, Department of Psychology, Denizli, Turkey
Glenn P. Williams, School of Psychology, Faculty of Health Sciences and Wellbeing, University of Sunderland, UK
Juan C. Oliveros-Chacana, Centro de Investigación en Ciencias Cognitivas, Facultad de Psicología, Universidad de Talca, Chile
Jan Philipp Röer, Department of Psychology and Psychotherapy, Witten/Herdecke University, Witten, Germany
Chiara Fini, Department of Dynamic and Clinical Psychology and Health Studies, Sapienza University of Rome, Italy
Oguz A. Acar, City, University of London, UK
Joseph P. McFall, State University of New York at Fredonia, USA
Ekaterina Pronizius, Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Austria
Jordan W. Suchow, School of Business, Stevens Institute of Technology, USA
Luisa Batalha, Australian Catholic University, Australia
Asil Ali Özdoğru, Department of Psychology, Üsküdar University, İstanbul, Turkey
Hendrik Godbersen, FOM University of Applied Sciences, Essen, Germany
Muhammad Mussaffa Butt, Government College University, Lahore, Pakistan
Jacek Buczny, Department of Experimental and Applied Psychology, VU Amsterdam, The Netherlands
Bastian Jaeger, Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, The Netherlands
Bradley J. Baker, Temple University, Philadelphia, PA, USA
Philip A. Grim II, Harrisburg University, Harrisburg, PA, USA
Zainab A. Alsuhaibani, Imam Mohammad Ibn Saud Islamic University (IMSIU), Saudi Arabia
Martín Martínez, University of Navarra, Spain
John Protzko, Central Connecticut State University, USA
Dermot Lynott, Department of Psychology, Maynooth University, Ireland
Max Korbmacher, Department of Health, Western Norway University of Applied, Norway
Mehmet Peker, Ege University, Department of Psychology, Turkey
Barnaby J.W. Dixson, School of Health and Behavioural Sciences, University of the Sunshine Coast, Australia
Mahmoud M. Elsherif, Department of Psychology, University of Birmingham, Birmingham, UK
Maital Neta, Department of Psychology, University of Nebraska-Lincoln, USA

Flavio Azevedo, Cambridge University, Cambridge, UK
Paulo Roberto dos Santos Ferreira, UFGD, Brazil
Fredrik Sigfrids, Åbo Akademi University, Finland
Tiago J S Lima, Department of Work and Social Psychology, University of Brasília
Sandra J. Geiger, Environmental Psychology, Department of Cognition, Emotion, and Methods, Faculty of Psychology, University of Vienna, Vienna, Austria
Anjali Thapar, Bryn Mawr College, USA
Manuel Perea, University of València, Spain
Raluca D. Szekely-Copîndean, Romanian Academy, Cluj-Napoca, Romania
Thomas Rhys Evans, School of Psychology and Counselling, University of Greenwich, England, UK
Steven Verheyen, Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands
David Moreau, University of Auckland, NZ
Ulrich S. Tran, Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria
Dina Abdel Salam El-Dakhs, Prince Sultan University, Saudi Arabia
Izuchukwu L. G. Ndukaihe, Alex Ekwueme Federal University Ndufu-Alike, Nigeria
Tijana Vesić Pavlović, University of Belgrade, Faculty of Mechanical Engineering, Serbia
Debora I. Burin, Universidad de Buenos Aires, Facultad de Psicologi-a / CONICET, Argentina
Patrícia Arriaga, Iscte-Instituto Universitario de Lisboa, CIS-IUL Lisbon, Portugal
Dauren Kasanov, Department Of Psychology, Ural Federal University, Ekaterinburg, Russian Federation
Jacob J. Keech, School of Health and Behavioural Sciences, University of the Sunshine Coast, QLD, Australia
María Fernández-López, University of València, Spain
Suzanne L. K. Stewart, School of Psychology, University of Chester, UK
David C. Vaidis, Université Paris Cité, France
Théo Besson, Université Paris Cité, France
Carlota Batres, Franklin and Marshall College, USA
Leigh Ann Vaughn, Ithaca College, USA
Magdalena Senderecka, Institute of Philosophy, Jagiellonian University, Krakow, Poland
Claudia Mazzuca, Department of Dynamic and Clinical Psychology and Health Studies, Sapienza University of Rome, Italy
Leticia Micheli, Institute of Psychology, Würzburg University, Würzburg, Germany
Martin R. Vasilev, Bournemouth University, Department of Psychology, UK
Kathleen Schmidt, Southern Illinois University, USA
Cameron Brick, University of Amsterdam, Department of Psychology, Amsterdam, Netherlands
Bruno Schivinski, School of Media and Communication, RMIT University, Australia
Susana Ruiz-Fernandez, FOM University of Applied Sciences, Essen, Germany
Ewa Ilczuk, Institute of Psychology, Jagiellonian University, Krakow, Poland
Carmel A Levitan, Department of Cognitive Science, Occidental College, USA
Emily Higgins, Dublin, Ireland
Gerit Pfuhl, Department of Psychology, UiT The Arctic University of Norway,
Jackson G. Lu, Massachusetts Institute of Technology, USA
Miroslav Sirota, University of Essex, UK
Zoran Pavlović, University of Belgrade, Faculty of Philosophy, Department of Psychology, Serbia, Europe
Ettore Ambrosini, Department of Neuroscience, University of Padova, Italy; Department of General Psychology, University of Padova; Padova Neuroscience Center, University of Padova

Christopher R. Chartier, Ashland University, Department of Psychology, USA
Heather D. Flowe, University of Birmingham, England, UK
Melissa F. Collof, School of Psychology, University of Birmingham, UK
Francesco Foroni, Australian Catholic University, NSW, Australia
Tess M. Atkinson, Central Connecticut State University, CT, USA
Amanda Kaser, Ashland University, OH, USA
Zdenek Meier, Palacky University Olomouc, Olomouc University Social Health Institute, Olomouc, Czech Republic
Nwadiogo Chisom ARINZE, Alex Ekwueme Federal University Ndufu-Alike, Nigeria
Marton A. Varga, ELTE Eotvos Lorand University Budapest, Budapest, Hungary
David Willinger, Department of Psychology and Psychodynamics, Karl Landsteiner University of Health Sciences, Krems an der Donau, Austria
Rumandeep K. Hayre, University of Birmingham, School of Psychology, UK
Miguel A. Vadillo, Universidad Autónoma de Madrid, Spain
Otto Loberg, Department of Psychology, Faculty of Science and Technology, Bournemouth University, UK
Aspasia Eleni Paltoglou, Department of Psychology, Faculty of Health and Education, Manchester Metropolitan University, UK
Gianni Ribeiro, The University of Queensland, Australia
Roxana-Elena Morariu, Babeș-Bolyai University, Cluj-Napoca, Romania
Timo B. Roettger, Universitetet i Oslo, Department of Linguistics and Scandinavian Studies, Norway
Tolga Ergiyen, Izmir University of Economics, Turkey
Maja Becker, CLLE, Université de Toulouse, CNRS, France
Yoann Julliard, Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, LIP/PC2S, 38000 Grenoble, France
Fatima Zakra Sahli, Mohammed V University Rabat, Morocco
Kelly Wolfe, University of Edinburgh, UK
Klara Malinakova, Olomouc University Social Health Institute, Palacky University in Olomouc, Olomouc, Czech Republic
Michal Parzuchowski, Center for Research on Cognition and Behavior, SWPS University of Social Sciences and Humanities in Sopot, Poland
Radka Zidkova, Olomouc University Social Health Institute, Palacky University Olomouc, Olomouc, Czech Republic
Lukas Novak, Olomouc University Social Health Institute, Palacky University in Olomouc, Olomouc, Czech Republic
Sarah E MacPherson, Human Cognitive Neuroscience, Department of Psychology, University of Edinburgh, Edinburgh, UK
Christopher L Aberson, Cal Poly Humboldt, USA
Wolf Vanpaemel, University of Leuven, Belgium
Bernhard Angele, Bournemouth University, UK
Dominique Muller, Univ. Grenoble Alpes; Institut Universitaire de France, France
Elif Gizem Demirag Burak, Koc University, Turkey
Peter Tavel, Olomouc University Social Health Institute, Palacky University in Olomouc, Olomouc, Czech Republic
Günce Yavuz-Ergiyen, Izmir University of Economics, Turkey
Savannah C. Lewis, Ashland University, OH, USA

## 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 will test the size and the variability of the semantic priming effect across ten 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) will allow quantifying evidence for semantic priming, whereas improvements in model fit with the addition of a random intercept for language will provide 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 word[1]. 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 words[1]. 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 representation[2], often by using results from these studies to develop theoretical and computational models that capture empirical effects[3–6]. Priming has also been used in studies of attention[7,8], studies of bi/multilingual people[9,10], on neurodivergent individuals such as those affected by Parkinson's disease, aphasia, or schizophrenia[11–13], and in a large body of neuroscience studies[14–16]. The purpose of this study is to leverage the power and network of the Psychological Science Accelerator (PSA)[17] 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 Lucas[18] and Hutchison[19] 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 work[20]. These datasets include lexical frequency[21,22], large collections of text (e.g., corpora)[23], response latencies,[24–26] and subjective ratings from participants on semantic dimensions such as emotion[27–29], concreteness[30], or familiarity[31]. 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[32–35].

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 known[36]. Semantic similarity has various conceptual and computational definitions that all generally describe the shared meaning between two words or texts[5]. The most common forms of similarity are feature-based similarity (i.e., number of shared features between words)[37–39], association strength (i.e., the probability of a first word eliciting a second word when simply shown the first word)[33,40], or text co-occurrence (i.e., words are similar because they frequently appear in proximity to one another)[41–43]. 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 pairs[24]. 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.[44] 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.'s[45] 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.[46] 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[47,48]. Related concepts that one may consider universal, such as LEFT and RIGHT, are not coded into all languages[49]. 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 languages[50,51]. Therefore, claims about human language are often based on a small set of languages, limiting the generalizability of these claims[52]. Even with the increase in publication of normed datasets in non-English languages[20], 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 techniques[53,54] and recently published corpora[23,55], 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 1[56]. 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 languages[57,58].

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 estimates[59,60] to address the known reliability issues in item-level responding[44,46] to support Hypothesis 1. The items will be selected using new computational techniques for addressing semantic similarity[53,54] with recently available

large corpora of movie subtitles[23] to appropriately match comparable items across languages. As noted in Buchanan et al.[20], 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.

## Method

### Ethics Information

We will 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). Each research lab will obtain local ethical review, rely on the ethical review provided by Harrisburg University, or provide evidence of no required ethical review. The IRB approvals are available on the Open Science Framework (OSF): https://osf.io/wrpj4/. Participants may be compensated for their participation by course credit or payment depending on individual lab resources. Labs will recruit participants via their own local resources. No exclusion criteria for participating in the study will be used, except for a minimum age requirement of 18 years (i.e., adult participants).

### Power Analysis

For our power analysis, we first detail a background on how we plan to estimate 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 is 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 results[46]. 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 will focus 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 is response latency, and we will use lexical decision data from the English Lexicon Project[25] and the Semantic Priming Project[24] 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 responses[46]. 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 will be 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"[59–61]. 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 will pair accuracy in parameter estimation with a sequential testing procedure to adequately sample each item, rather than estimate an overall effect size. Therefore, we will use 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 will be assessed to determine if the item has met the goals for accuracy in parameter estimation as our stopping rule. If so, the item will be 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 latency[62]. 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 are skewed due to the large number of infrequent words in the English Lexicon Project, and we will use 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 are 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 would represent our potentially 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 will be used as our minimum sample size for a lexical decision task, and the accuracy standard error level will potentially be 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 aim to achieve an accurate representation of lexical decision times, from which priming can then be calculated. However, it should be noted that accurately measured response latencies do not necessarily imply "reliable" priming or difference score data[63], 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 is 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 find the maximum sample size of 320 participants for 80% of the items to reach the smaller standard error of 0.06. Therefore, we will use 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 will complete approximately 800 lexical decision trials per session, and each participant only completes 150 of these concepts (75 targets in the related condition, 75 targets in the unrelated condition as cue words are not analyzed) that are 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 is 741 to 4741 [(1000*50*2) / (150*.90); (1000*320*2) / 150*.90]. The complete code and description of this process are detailed at: https://osf.io/rxgkf/.

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 is quite large, we will employ a stopping procedure to ensure participant time and effort is maximized, and data collection is optimized. To summarize, the minimum sample size will be 50 participants per word and the

maximum will be 320, which results in 741 to 4741 participants per language based on expected usable trials. Therefore, the total sample size will range from 7410 to 47410 participants for ten languages. After 50 participants, each concept will be examined for standard error, and data collection for that concept will be decreased in probability when the standard error reaches our average criterion of 0.09. Item probability for selection will also be decreased when they reach the maximum proposed sample size ($n$ = 320). This process will be automated online and checked in a scheduled subroutine.

While 43 languages have been identified for possible data collection, we plan to first publish the data when ten languages have reached the appropriate sample size as outlined above based on recruitment of PSA partner labs. We will complete minimum data collection in English, Spanish, Chinese, Portuguese, German, Korean, Russian, Turkish, Czech, and Japanese. To date, we have recruited more than 100 researchers in 19 potential languages.

## Materials

The following details the important facets of the materials. We will first explain the types of word-pair conditions in a semantic priming study (i.e., related, unrelated, and non-word). Next, we will 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 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 are shuffled so that the cue word is combined with a target word with which it has a negligible cosine distance similarity (i.e., < .15).

Finally, non-words pair conditions are created by using the Wuggy-like algorithm[64] for non-logographic languages. We will consult with at least two native speakers to change one stroke or radical such that the character(s) are a pronounceable word with no meaning by starting from known non-word lists[65]. Any disagreements between native speakers will be resolved with discussion between these speakers. Each cue and target word were first hyphenated using the sylly package and LaTeX style hyphenation[66]. 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 will examine each programmatically generated word to ensure they are pronounceable (i.e., phonologically valid) and not pseudo-homophones (i.e., wherein the pronunciation sounds like a real word, KEEP → KEAP)[64]. In cases of disagreement, the native speakers will discuss and resolve these inconsistencies. When they have marked a non-word for exclusion, a new non-word will be generated until speakers agree it meets the rules for non-words. Native speakers may also suggest alternatives, which the lead author will check to ensure match to desired non-word characteristics.

To control the ability of participants to anticipate or guess the answers, we will ensure that half the trials will be answered with a word and half with a nonword. Therefore, we will use 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 can be 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 will be 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 is 50% for words-nonwords across all trials, and the relatedness proportion for pairs is 18.8%.

**Similarity calculation**

***Corpora.*** As described in the introduction, the choice of related words based on similarity is key for the study. There are multiple measures of semantic similarity including the cosine between overlapping features[39], free association probabilities[33,40,67], and local/global coherence values from network models[35,68]. However, the underlying data for these calculations is inconsistent across languages. Therefore, one solution is to use the data present in the OpenSubtitles datasets[23] (i.e., a large collection of movie subtitles) to calculate word frequency and cosine distance 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 phenomena[21,69–76]. Cosine distance 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 (i.e., Chinese-English mix). We will use 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 package[77] 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 allows for the possibility of 43 different languages in this project. See Figure 1 for the model selection process.

***Modeling.*** The subs2vec project[55] used the OpenSubtitles data to create fastText[78] computational representation for 55 languages. fastText is a distributional vector space model, an extension of word2vec[53,54]**,** 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 vectors[4,5]. We will use the existing models from subs2vec to extract related word concepts for the most frequent concepts identified using the top cosine distance between word vectors.

***Cue selection procedure.*** The procedure for stimuli selection can be viewed at https://osf.io/s9h3z/ and is displayed graphically in Figure 1. If the language is available via subs2vec, the provided subtitle frequency counts will be examined. If the language has more than 50,000 unique concepts represented in the subtitle data, we will use the subtitle model only. If the subtitles do not provide enough linguistic information (i.e., fewer than 50,000 concepts in the corpus), we will use the combined Wikipedia and subtitle model[55]. subs2vec contains 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 language[21,69]; 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 will then be tagged for part-of-speech using udpipe*,* selecting tokens that are 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 dataset[79]. This procedure will cover all but two languages in our list of 43 possible languages. For the final two languages, we will use *udpipe* to tag the OpenSubtitles directly and calculate word frequency. Additionally, fastText model using the same parameters as subs2vec will be trained for similarity calculation. The 10,000 most frequent concepts will be selected at this point.

***Target selection procedure.*** Using the fastText models for each language, we will select the top five cosine distance similarity values for each concept in each language independently, resulting in 50,000 possible cue-target pairs. These will be cross-referenced across languages using Google Translate to create a master list of potential cue-target pairings. The related word pairs ($n$ = 1000) will be selected from this list using each cue or target only once, favoring pairs with translations in most languages. Therefore, the selection procedure will be based on the most common cue-target pairs across languages, rather than selecting similar words in one language and then translating. This procedure is programmatic, using Google Translate, which may not produce the most appropriate translation for a word. Therefore, native speakers will ensure 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 will suggest a more common or appropriate word for items they think are unusual, and in cases of disagreement, group discussion between the two translators will be used. In some instances, translation may indicate that a particular language does not have separate concepts for the cue-target pairing. In this instance, we will change the cue word to a related word for that language from the five selected in the original list. Thus, all targets are matched across languages, and as many cues as possible while avoiding repetition within a cue-target pair.

**Procedure**

We will 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 will complete 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.

**Implementation**

***Timing software.*** While participants will be naïve to the word pairings, the principal investigator will know the pair combinations during data collection and analysis. A small demonstration of the experiment can be found at: https://psa007.psysciacc.org/. The study will be programmed using lab.js[80], which is an online, open-source, study-building software. Precise timing measurement is required for this study, and the lab.js team has documented the accuracy of measurement within their framework[81], and previous work has shown no differences between lab and web-based data collection for response latencies[82]. In addition, SPALEX, a large lexical decision database in Spanish, was collected completely online[26]. We will recommend 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 will be directed to an online web portal to complete the study, and all data will be retained in the online platform with nightly backups to the server. Participants will be 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 will indicate important aspects about screen size, browser, and timing accuracy. In order to enforce this requirement, participants will be asked to hit the spacebar to continue the study.

***Adaptive stimuli selection.*** At the start of data collection, all presented items will be randomly selected from the larger item pool by equalizing the probability of inclusion equal for all words and non-words ($p$ = 1/1000 concepts). After the minimum sample size is collected, each word's standard error will be checked to determine if the sample size for that item has reached our accuracy criteria. If so, the probability of sampling that item will be 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 will ensure that there are always words to randomly select from (i.e., to keep the same procedure and number of trials for all participants) and that the randomization is 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 have reached the stopping criteria or maximum sample size, the probabilities will be equalized. We have set minimum, maximum, and a stopping rule for the initial data collection; however, we will allow data collection after these have been reached and will post updates to the data using a DOI service to allow researchers to cite the specific dataset they used for their research (modeled after the Small World of Words Project[33], which is ongoing). All data will be included in our dataset, and the analysis section describes how we will indicate potential data for exclusion. Therefore, data collection will occur in a repeated-measures design in which participants do not see all of the possible stimuli, but do see all the possible conditions (related, unrelated, and non-word pairs). They are blind to the condition each pair is presented in.

**Study Procedure**

***Demographics.*** Participants will be directed to select their first language, which will then direct them to the appropriate translation of the experiment. Participants will be 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. A flow chart of the procedure is provided in Figure 2.

***Lexical Decision Task.*** Instructions on how to complete a lexical decision task will be shown on the next screen, followed by 10 practice trials. Each trial starts with a fixation cross (+) in the middle of the screen for 500 ms. The stimulus item will then be displayed in the middle of the screen in uppercase san-serif 18-point font (i.e., Arial font, DOG). On the bottom of the screen the possible responses will be 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). Response keys will be 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 side[83]. Participants will be asked for the dominant hand at the beginning of the study to determine the response mapping for their study. Participants will make their choice for each concept, and during the practice trials, they will receive feedback if their answer was correct or incorrect. The next stimulus will appear 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 is showing). Responses will time out after three seconds and move on to the next trial. After 10 trials, participants will see the instruction screen again with a reminder that they will now be doing the real task.

After 100 trials, the participants will be shown a short break screen with the option to continue by hitting the spacebar after 10 seconds. After eight blocks of 100 trials (800 word-nonword decisions), the experiment will end with a thank you screen. On this screen, participants will indicate what type of credit they are receiving for the study (e.g., course credit, payment, no compensation, prize drawing), and they will be given instructions on how to indicate that they have completed the study to the appropriate lab. Participants will be allowed to take the study multiple times as items are randomly selected for inclusion. An estimate for the time required for the study is approximately 30 minutes inclusive of practice

trials, reading all instructions, and breaks. This estimate is based on previous studies of lexical decision times[25], and pilot testing will be used to determine if the number of trials should be reduced to accommodate the 30-minute expected time.

***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 effects[84,85]. 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. Lucas[18] provides evidence that priming effect sizes are relatively equal across task type (continuous, masked, paired, and naming), and therefore, we should expect similar results.

***Additional data.*** We will combine available lexical and subject rating data with the priming data. 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 projects. Subjective measures are concept characteristics that are rated by participants, and we will include age of acquisition[86–89] (approximate age you learned a concept), imageability[90,91] (how easy the concept comes to mind), concreteness[92] (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)[27,29], and familiarity (how well a person knows a

concept)[93]. These variables were selected from the list of most published databases for linguistic data[20].

## Analysis Plan

An example of the data and processing for English can be found at https://osf.io/6jmzk/. Each of the sections described below in the descriptive statistics are available as files for raw and processed data from our OSF page. All data will be archived on our server, and we will use Zenodo (https://zenodo.org/) to release versions of the data with citable DOIs given the planned continuation of the project after the initial PSA support.

### Descriptive statistics

#### Participant level data

We will present descriptive statistics on the participants involved in the study including percentages of gender, education levels, native language, and average age. Information about the device used to complete the study will include percentages of computer operating system, the web browser, and the language locale (i.e., the language the browser defaults to using). Finally, the sample sizes collected by the collaborating labs will be provided. Each of these statistics will be provided for the overall data and the data separated by language.

#### Trial level data

Each language will be saved in a separate file with an item specific trial identification number to allow for matching concepts across languages (i.e., CAT [English] → KATZE [German] → GATTA [Italian]). If a participant leaves the study early (e.g., Internet disconnection, computer crash, closes the study), the data past this point in the study is not recorded, and therefore, the trial level data represents all trials displayed during the experiment. Participants are expected to incorrectly answer trials, and these trials will be marked for exclusion. All timeout trials will be marked as missing values in the final data. No missing values will be imputed.

We will mark for exclusion minimum response latencies of less than 160 ms[94] (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). The response latencies from each participant's session will then be *z*-scored in line with recommendations from Faust et al.[62]. We will not collect enough data to note if a person takes the experiment multiple times for privacy reasons, but as these would be considered different sessions, the recommended *z*-score procedure should control for participant variability at this level. Therefore, repeated participation would not be detrimental to data collection. Finally, participants' overall proportion of correct answers will be calculated, and participants who do not correctly answer at least 80% of 100 minimum trials will be excluded 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.

We will provide descriptive statistics on the average time to complete the study, the number of trials by word type (word, nonword), the accuracy by word type, and average *z*-scored response latencies by word type (overall, excluding $Z > 2.5$, excluding $Z > 3.0$; see below). These values will be provided for overall results and separated by language.

**Item level data**

The item file will contain lexical information about all stimuli calculated from the OpenSubtitles[23] and subs2vec[55] projects (length, frequency, orthographic neighborhood, bigram frequency, orthographic and phonographic Levenshtein distance). The descriptive statistics calculated from the trial level data will then be included: mean response latency, average standardized response latency, sample size, standard errors of response latencies, and accuracy rate. No data will be excluded for being a potential outlier; however, we will recommend a cut-off criterion for absolute value *z*-score outliers at 2.5 and 3.0, and we will calculate 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 will be included. These values do not exist for non-words.

We will provide descriptive statistics on the average sample size, average *z*-scored response latencies, and average *SE* for the *z*-scored response latencies by each word type (word, nonword). These values will be calculated for the overall data set, separated by language, and without each level of *z*-score outlier criterion.

**Priming data**

In a separate file, we will also prepare 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 is 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 will be provided in this file, 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 Italian[95], German[96], Spanish[26], and Dutch[97]**.**

We will provide the average statistics for *z*-score priming, *z*-score unrelated response latency, *z*-score related response latency, sample size for unrelated trials, and sample size for related trials. These values will be calculated overall, by language, and with/without *z*-score level exclusions. Last, we will calculate the participant level priming reliability[98] and item-level priming reliability[44].

**Exclusion summary**

Data will be excluded for the following reasons in this order:

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 $Z = 3.0$.

**Hypothesis 1**

Hypothesis information is presented in Table 1. Hypothesis 1 predicts semantic facilitation with reduced response latencies for related than unrelated words. Hypothesis 1 will be analyzed by calculating an intercept-only regression model using the *z*-scored priming response latency as the dependent variable. The intercept and its 95% confidence interval will represent the grand mean of the priming effect across all languages. The priming response latency is calculated by taking the average of the unrelated pair *z*-scored response latency minus the related pair response latency within each item. 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. We will determine support for Hypothesis 1 if the lower limit of the confidence interval is greater than zero (i.e., a directional comparison). This process will be repeated for average priming scores calculated without trials that were marked as 2.50 *z*-score outliers and 3.00 *z*-score outliers separately. The decision criteria will remain the same, and we will identify any differences in decisions based on outlier statistics (e.g., priming only occurs when X trials are removed).

**Hypothesis 2**

Hypothesis 2 explores the extent to which these semantic priming effects vary across languages. Therefore, we will calculate a random effects model using the nlme[99] package in *R* wherein the random intercept of language will be added to the overall intercept only model for Hypothesis 1. We will report the standard deviation of the random effect, its 95% confidence interval, the AIC change between models, and the pseudo-$R^2$ values for the

effect size of this parameter[100]. Results will support significant heterogeneity when the AIC for the random effects model is two points or more less than the AIC for the intercept-only model[56]. This analysis will be repeated with the 2.50 *z*-score outliers and 3.00 *z*-score outliers excluded. We will include a forest plot of the priming effect and their 95% confidence intervals to visualize the potential heterogeneity in the priming results. Simulations of models within and without variability in the priming effects can be found at https://osf.io/fbhr8/.

**Protocol Registration**

Our preregistration 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 the website devoted to this project with backups provided on OSF and Zenodo.

**Code Availability**

All code used for study creation and delivery, data processing, and analyses will be 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. Copland, D. The basal ganglia and semantic engagement: Potential insights from semantic priming in individuals with subcortical vascular lesions, Parkinson's disease, and cortical lesions. *J Int Neuropsychol Soc* **9**, 1041–1052 (2003).
12. Haverkort, M. Linguistic Representation and Language Use in Aphasia. in *Twenty-First Century Psycholinguistics: Four Cornerstones* 57–68 (Routledge/Taylor & Francis Group, 2005).
13. Tan, E. J., Neill, E. & Rossell, S. L. Assessing the Relationship between Semantic Processing and Thought Disorder Symptoms in Schizophrenia. *J Int Neuropsychol Soc* **21**, 629–638 (2015).
14. 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).
15. 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).
16. 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).
17. 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).
18. Lucas, M. Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review* **7**, 618–630 (2000).
19. Hutchison, K. A. Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review* **10**, 785–813 (2003).
20. 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).

21. New, B., Brysbaert, M., Veronis, J. & Pallier, C. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* **28**, 661–677 (2007).
22. Gimenes, M. & New, B. Worldlex: Twitter and blog word frequencies for 66 languages. *Behav Res* **48**, 963–972 (2016).
23. Lison, P. & Tiedemann, J. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. (2016).
24. Hutchison, K. A. *et al.* The semantic priming project. *Behav Res* **45**, 1099–1114 (2013).
25. Balota, D. A. *et al.* The English Lexicon Project. *Behavior Research Methods* **39**, 445–459 (2007).
26. Aguasvivas, J. A. *et al.* SPALEX: A Spanish Lexical Decision Database From a Massive Online Data Collection. *Front. Psychol.* **9**, 2156 (2018).
27. 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).
28. Warriner, A. B., Kuperman, V. & Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav Res* **45**, 1191–1207 (2013).
29. Bradley, M. M. & Lang, P. J. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. (1999).
30. Brysbaert, M., Warriner, A. B. & Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav Res* **46**, 904–911 (2014).
31. Stadthagen-Gonzalez, H. & Davis, C. J. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods* **38**, 598–605 (2006).
32. De Deyne, S., Navarro, D. J., Perfors, A. & Storms, G. Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General* **145**, 1228–1254 (2016).

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. Vankrunkelsven, H., Verheyen, S., Storms, G. & De Deyne, S. Predicting Lexical Norms: A Comparison between a Word Association Model and Text-Based Word Co-occurrence Models. *Journal of Cognition* **1**, 45 (2018).

35. Vitevitch, M. S., Goldstein, R., Siew, C. S. Q. & Castro, N. Using complex networks to understand the mental lexicon. *Yearbook of the Poznan Linguistic Meeting* **1**, 119–138 (2014).

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

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

38. 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).

39. 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).

40. 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).

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

42. 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).

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

44. 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).

45. 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).

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

47. Evans, N. & Levinson, S. C. The myth of language universals: Language diversity and its importance for cognitive science. *Behav Brain Sci* **32**, 429–448 (2009).

48. Marslen-Wilson, W. D. Access to lexical representations: Cross-linguistic issues. *Language and Cognitive Processes* **16**, 699–708 (2001).

49. Majid, A. & Levinson, S. C. WEIRD languages have misled us, too. *Behav Brain Sci* **33**, 103–103 (2010).

50. 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).

51. 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).

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

53. 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).

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

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

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

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

58. 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).

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

60. 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).

61. 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).

62. 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).

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

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

65. 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).

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

67. 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).

68. Siew, C. S. Q. & Vitevitch, M. S. Spoken word recognition and serial recall of words from components in the phonological network. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **42**, 394–410 (2016).

69. 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).

70. 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).

71. 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).

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

73. 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).

74. 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).
75. Mandera, P., Keuleers, E., Wodniecka, Z. & Brysbaert, M. Subtlex-pl: subtitle-based word frequency estimates for Polish. *Behav Res* **47**, 471–483 (2015).
76. 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).
77. Wijffels, J. *et al. udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. (2021).
78. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).
79. Benoit, K., Muhr, D. & Watanabe, K. *stopwords: Multilingual Stopword Lists*. (2021).
80. 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.
81. 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).
82. Hilbig, B. E. Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behav Res* **48**, 1718–1724 (2016).
83. Proctor, R. W. & Cho, Y. S. Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin* **132**, 416–442 (2006).
84. 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).

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

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

87. 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).

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

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

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

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

92. 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).

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

94. Proctor, R. W. & Schneider, D. W. Hick's law for choice reaction time: A review. *Quarterly Journal of Experimental Psychology* **71**, 1281–1299 (2018).

95. 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).

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

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

98. 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, 2017).

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

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

**Acknowledgements**

The authors received no specific funding for this work. Specific author funders had no role in study design, data collection and analysis, decision to publish or preperation of this manuscript. Krystian Barzykowski and Ewa Ilczuk were supported by a grant from the National Science Centre, Poland (2019/35/B/HS6/00528).

**Author contributions**

- Erin M. Buchanan: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing - Original Draft, Writing - Review & Editing
- Kelly M. Cuccolo: Data Curation, Investigation, Project Administration, Supervision, Writing - Review & Editing
- Nicholas Coles: Project Administration, Writing - Review & Editing
- Tom Heyman: Conceptualization, Methodology, Project Administration, Writing - Review & Editing
- Aishwarya Iyer: Project Administration, Writing - Review & Editing
- Neil Lewis Jr.: Project Administration, Writing - Review & Editing
- Kim Peters: Project Administration, Writing - Review & Editing
- Niels van Berkel: Project Administration, Software, Writing - Review & Editing
- Anna E. van ’t Veer: Project Administration, Writing - Review & Editing
- Jack E. Taylor: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing
- Maria Montefinese: Conceptualization, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing
- K. D. Valentine: Conceptualization, Writing - Original Draft, Writing - Review & Editing
- Nicholas P. Maxwell: Conceptualization, Writing - Review & Editing
- Belgüzar Nilay Türkan: Investigation, Resources, Writing - Review & Editing
- Glenn P. Williams: Investigation, Writing - Review & Editing
- Juan C. Oliveros-Chacana: Investigation, Resources, Writing - Review & Editing
- Jan Philipp Röer: Investigation, Writing - Review & Editing
- Chiara Fini: Investigation, Writing - Review & Editing
- Oguz A. Acar: Investigation, Writing - Review & Editing
- Joseph P. McFall: Investigation, Writing - Review & Editing
- Ekaterina Pronizius: Investigation, Writing - Review & Editing
- Jordan W. Suchow: Investigation, Writing - Review & Editing
- Luisa Batalha: Investigation, Writing - Review & Editing
- Asil Ali Özdoğru: Investigation, Resources, Writing - Review & Editing
- Hendrik Godbersen: Investigation, Resources, Writing - Review & Editing
- Muhammad Mussaffa Butt: Investigation, Resources, Writing - Review & Editing
- Jacek Buczny: Investigation, Resources, Writing - Review & Editing
- Bastian Jaeger: Investigation, Writing - Review & Editing
- Bradley J. Baker: Investigation, Writing - Review & Editing
- Philip A. Grim II: Investigation, Writing - Review & Editing
- Zainab A. Alsuhaibani: Investigation, Resources, Writing - Review & Editing
- Martín Martínez: Investigation, Resources, Writing - Review & Editing
- John Protzko: Investigation, Writing - Review & Editing
- Dermot Lynott: Investigation, Writing - Review & Editing

- Max Korbmacher: Investigation, Writing - Review & Editing
- Mehmet Peker: Investigation, Resources, Writing - Review & Editing
- Barnaby J.W. Dixson: Investigation, Writing - Review & Editing
- Mahmoud M. Elsherif: Investigation, Resources, Writing - Review & Editing
- Maital Neta: Investigation, Resources, Writing - Review & Editing
- Flavio Azevedo: Investigation, Writing - Review & Editing
- Paulo Roberto dos Santos Ferreira: Investigation, Writing - Review & Editing
- Fredrik Sigfrids: Investigation, Resources, Writing - Review & Editing
- Tiago J S Lima: Investigation, Resources, Writing - Review & Editing
- Sandra J. Geiger: Investigation, Writing - Review & Editing
- Anjali Thapar: Investigation, Writing - Review & Editing
- Manuel Perea: Investigation, Resources, Writing - Review & Editing
- Raluca D. Szekely-Copîndean: Investigation, Resources, Writing - Review & Editing
- Thomas Rhys Evans: NA, Investigation, Writing - Review & Editing
- Steven Verheyen: Investigation, Resources, Writing - Review & Editing
- David Moreau: Investigation, Resources, Writing - Review & Editing
- Ulrich S. Tran: Investigation, Writing - Review & Editing
- Dina Abdel Salam El-Dakhs: Investigation, Resources, Writing - Review & Editing
- Izuchukwu L. G. Ndukaihe: Investigation, Writing - Review & Editing
- Tijana Vesić Pavlović: Investigation, Resources, Writing - Review & Editing
- Debora I. Burin: Investigation, Resources, Writing - Review & Editing
- Patrí-cia Arriaga: Investigation, Resources, Writing - Review & Editing
- Dauren Kasanov: Investigation, Resources, Writing - Review & Editing
- Jacob J. Keech: Investigation, Writing - Review & Editing
- María Fernández-López: Investigation, Resources, Writing - Review & Editing
- Suzanne L. K. Stewart: Investigation, Writing - Review & Editing
- David C. Vaidis: Investigation, Resources, Writing - Review & Editing
- Théo Besson: Investigation, Resources, Writing - Review & Editing
- Carlota Batres: Investigation, Writing - Review & Editing
- Leigh Ann Vaughn: Investigation, Writing - Review & Editing
- Magdalena Senderecka: Investigation, Resources, Writing - Review & Editing
- Claudia Mazzuca: Investigation, Writing - Review & Editing
- Leticia Micheli: Investigation, Resources, Writing - Review & Editing
- Martin R. Vasilev: Investigation, Resources, Writing - Review & Editing
- Kathleen Schmidt: Investigation, Writing - Review & Editing
- Cameron Brick: Investigation, Writing - Review & Editing
- Bruno Schivinski: Investigation, Writing - Review & Editing
- Susana Ruiz-Fernandez: Investigation, Resources, Writing - Review & Editing
- Ewa Ilczuk: Investigation, Writing - Review & Editing
- Carmel A Levitan: Investigation, Writing - Review & Editing
- Emily Higgins: Investigation, Writing - Review & Editing
- Gerit Pfuhl: Investigation, Resources, Writing - Review & Editing
- Jackson G. Lu: Investigation, Writing - Review & Editing
- Miroslav Sirota: Investigation, Writing - Review & Editing
- Zoran Pavlović: Investigation, Resources, Writing - Review & Editing
- Ettore Ambrosini: Investigation, Resources, Writing - Review & Editing
- Nienke Böhm: Investigation, Resources, Writing - Review & Editing
- Aslan Karaaslan: Investigation, Resources, Writing - Review & Editing
- Marietta Papadatou-Pastou: Investigation, Resources, Writing - Review & Editing
- Sezin Öner: Investigation, Resources, Writing - Review & Editing
- Ernest Baskin: Investigation, Writing - Review & Editing
- Kate E. Mulgrew: Investigation, Writing - Review & Editing
- José Luis Ulloa: Investigation, Resources, Writing - Review & Editing

- Ewa Szumowska: Investigation, Resources, Writing - Review & Editing
- Patricia Garrido-Vásquez: Investigation, Writing - Review & Editing
- Krystian Barzykowski: Investigation, Resources, Writing - Review & Editing
- Alexandra I. Kosachenko: Investigation, Resources, Writing - Review & Editing
- Chin Wen Cong: Investigation, Resources, Writing - Review & Editing
- Claus Lamm: Investigation, Writing - Review & Editing
- Andrei Dumbravă: Investigation, Resources, Writing - Review & Editing
- Vanessa Era: Investigation, Writing - Review & Editing
- Luis Carlos Pereira Monteiro: Investigation, Writing - Review & Editing
- Peter R. Mallik: Investigation, Writing - Review & Editing
- Chris Isloi: Investigation, Writing - Review & Editing
- Ali H. Al-Hoorie: Investigation, Resources, Writing - Review & Editing
- Natalia Irrazabal: Investigation, Resources, Writing - Review & Editing
- Yuri G. Pavlov: Investigation, Resources, Writing - Review & Editing
- Anna O. Kuzminska: Investigation, Resources, Writing - Review & Editing
- William E. Davis: Investigation, Writing - Review & Editing Sarah E. Fisher: Investigation, Writing - Review & Editing
- Mai Helmy: Investigation, Writing - Review & Editing
- Julia Valeiro Paterlini: Investigation, Resources, Writing - Review & Editing
- Guanxiong Huang: Investigation, Resources, Writing - Review & Editing
- Anna M. Borghi: Investigation, Writing - Review & Editing
- Balazs Aczel: Investigation, Resources, Writing - Review & Editing
- Stefan Stieger: Investigation, Writing - Review & Editing
- S. C. Chen: Investigation, Resources, Writing - Review & Editing
- Laura M. Stevens: Investigation, Writing - Review & Editing
- Christophe Blaison: Investigation, Writing - Review & Editing
- Abigail G. Sanders: Investigation, Writing - Review & Editing
- Robert M. Ross: Investigation, Writing - Review & Editing
- Madeleine P. Ingham: Investigation, Writing - Review & Editing
- Tia C. Bennett: Investigation, Writing - Review & Editing
- Jason Geller: Formal Analysis, Validation, Writing - Review & Editing
- Ogeday Çoker: Investigation, Writing - Review & Editing
- Erin Sievers: Investigation, Writing - Review & Editing
- Christopher R. Chartier: Investigation, Writing - Review & Editing
- Heather D. Flowe: Investigation, Resources, Writing - Review & Editing
- Melissa F. Collof: Investigation, Writing - Review & Editing
- Francesco Foroni: Investigation, Writing - Review & Editing
- Tess M. Atkinson: Investigation, Writing - Review & Editing
- Amanda Kaser: Investigation, Writing - Review & Editing
- Zdenek Meier: Investigation, Writing - Review & Editing
- Nwadiogo Chisom ARINZE: Investigation, Writing - Review & Editing
- Marton Aron Varga: Investigation, Resources, Writing - Review & Editing
- David Willinger: Investigation, Resources, Writing - Review & Editing Rumandeep K. Hayre: Investigation, Writing - Review & Editing
- Miguel A. Vadillo: Investigation, Resources, Writing - Review & Editing
- Otto Loberg: Investigation, Writing - Review & Editing
- Aspasia Eleni Paltoglou: Investigation, Resources, Writing - Review & Editing
- Gianni Ribeiro: Investigation, Writing - Review & Editing
- Roxana-Elena Morariu: Investigation, Writing - Review & Editing
- Timo B. Roettger: Investigation, Resources, Writing - Review & Editing
- Tolga Ergiyen: Investigation, Resources, Writing - Review & Editing
- Maja Becker: Investigation, Resources, Writing - Review & Editing
- Yoann Julliard: Investigation, Writing - Review & Editing

- Fatima Zakra Sahli: Investigation, Resources, Writing - Review & Editing
- Kelly Wolfe: Investigation, Writing - Review & Editing
- Klara Malinakova: Investigation, Writing - Review & Editing
- Michal Parzuchowski: Investigation, Resources, Writing - Review & Editing
- Radka Zidkova: Investigation, Writing - Review & Editing
- Lukas Novak: Investigation, Writing - Review & Editing
- Sarah E MacPherson: Investigation, Writing - Review & Editing
- Christopher L Aberson: Investigation, Writing - Review & Editing
- Wolf Vanpaemel: Investigation, Resources, Writing - Review & Editing
- Bernhard Angele: Investigation, Writing - Review & Editing
- Dominique Muller: Investigation, Writing - Review & Editing
- Elif Gizem Demirag Burak: Investigation, Resources, Writing - Review & Editing
- Peter Tavel: Investigation, Writing - Review & Editing
- Günce Yavuz-Ergiyen: Investigation, Resources, Writing - Review & Editing
- Savannah C. Lewis: Project Administration, Resources, Writing - Review & Editing

**Competing interests**

The authors declare no competing interests.

**Figure 1.** 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.

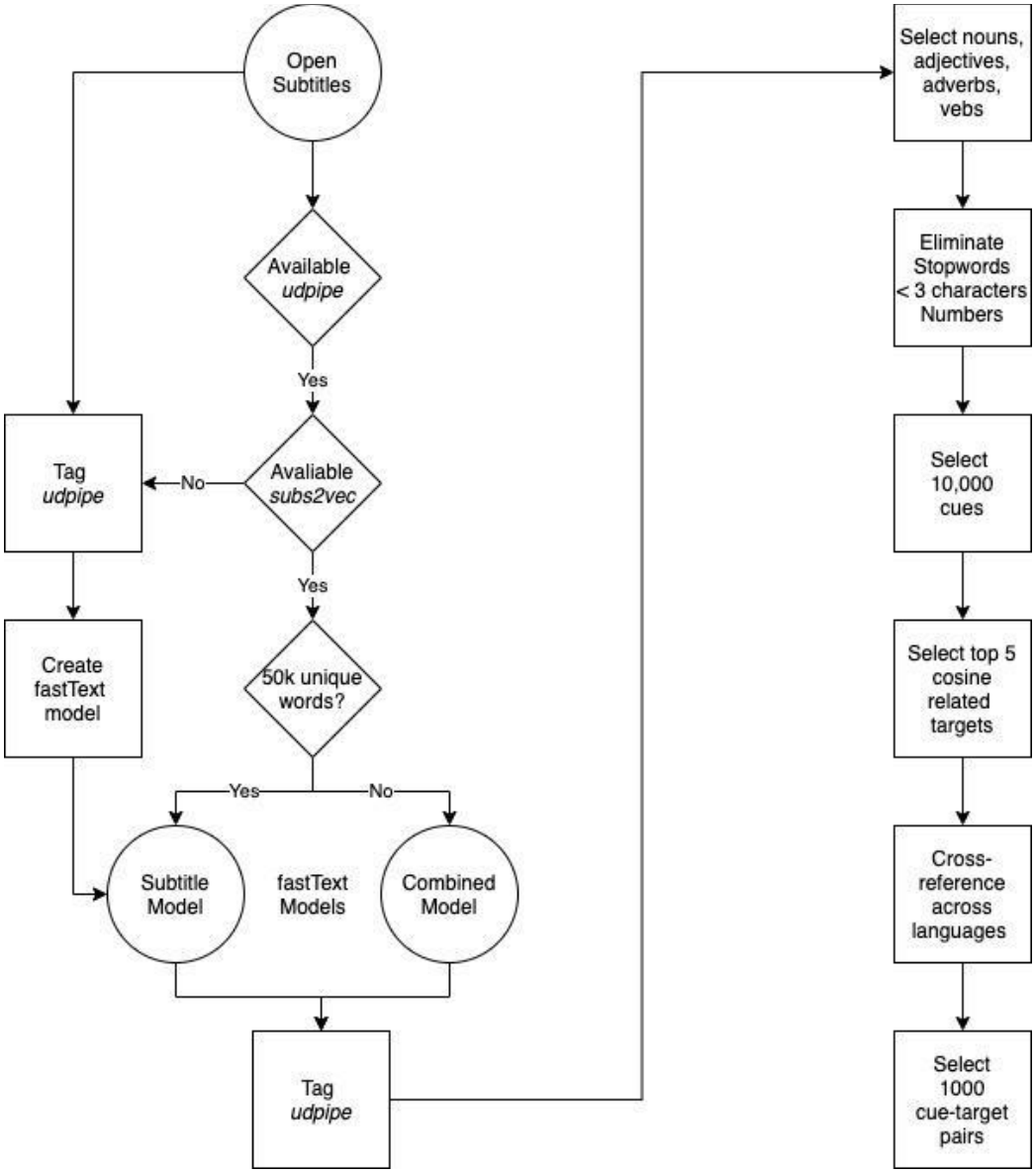

**Figure 2.** Flow chart of the procedure for the study. Within the lexical decision task, participants are given short breaks after 100 trials (i.e., each answer given). The answer choices for that language will always be displayed on the bottom of the screen during the lexical decision task.

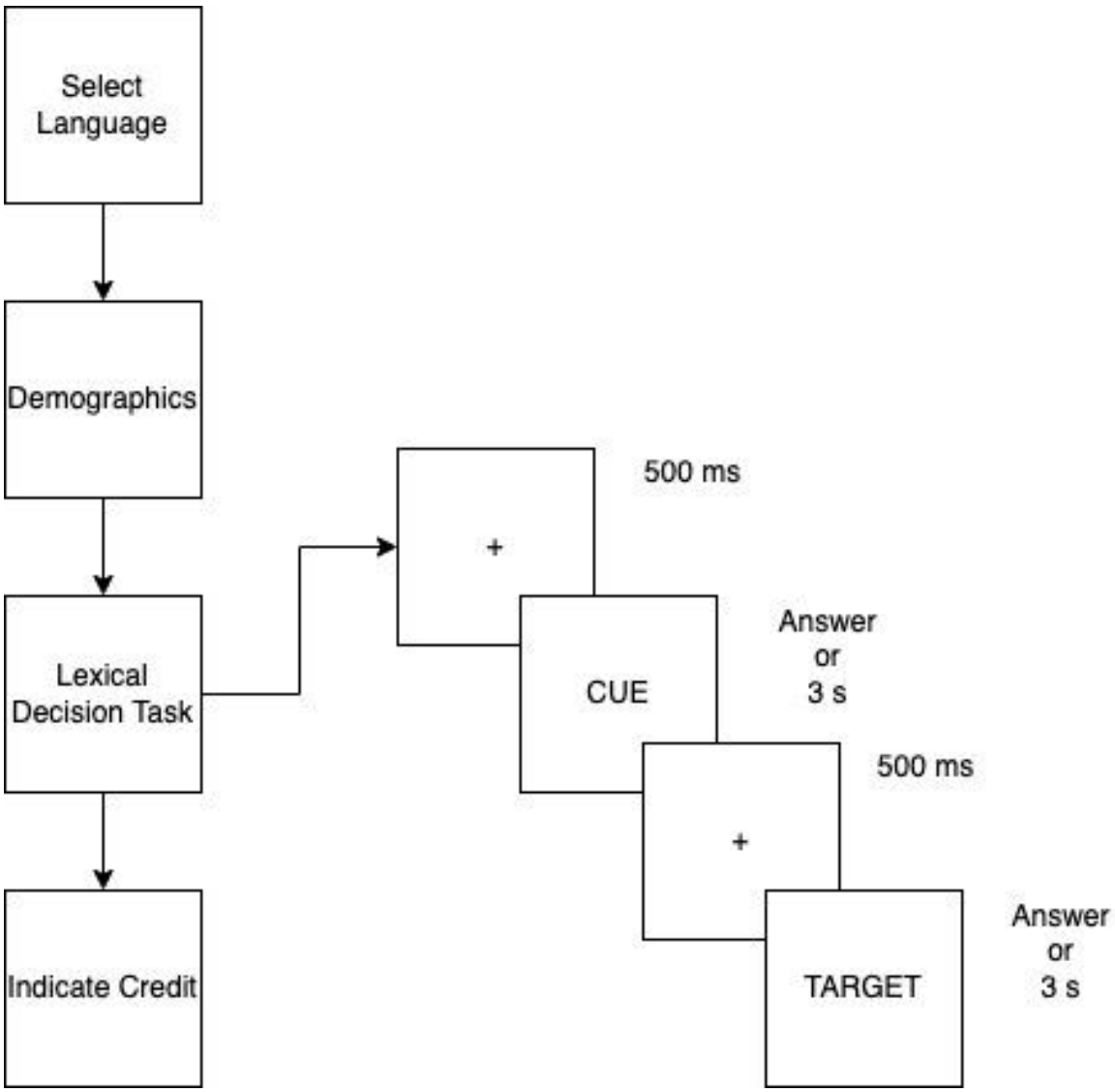

**Table 1.**

**Design Table**

| **Question** | **Hypothesis** | **Sampling plan (e.g., power analysis)** | **Analysis Plan** | **Interpretation given to different outcomes** |
|---|---|---|---|---|
| Is semantic priming a non-zero effect? | $H_A$: Response latencies will be faster for related word-pairs in comparison to unrelated word pairs.<br><br>$H_0$: 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.**<br><br>**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 $H_A$ when the lower limit of the confidence interval is **positive and non-zero > 0.0001**<br><br>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? | **$H_A$: Priming response latencies will be variable between languages (i.e., heterogeneous).**<br><br>**$H_0$: 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.**<br><br>**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 $H_A$ when the **ΔAIC (intercept-only minus random-intercept) is ≥ 2 points.**<br><br>The results will be inconclusive when the **ΔAIC (intercept-only minus random-intercept) is < 2 points.** |