SPAM-L: Semantic Priming Across Multiple Languages

Erin M. Buchanan*, Harrisburg University of Science and Technology

Maria Montefinese, University of Padova and University College London

Felix Henninger, University of Mannheim

Jack Taylor, University of Glasgow

K. D. Valentine, Massachusetts General Hospital

Draft Date: 2019-09-15

Keywords: semantic priming, normed stimuli, cognitive science, psycholinguistics

*Submitting author

All other authors listed credit contributions (as current):

https://github.com/doomlab/SPAML/blob/master/credit_authorship.xlsx

Abstract

Semantic priming has been studied for nearly fifty years across various experimental manipulations and theoretical frameworks. Critically, the understanding of semantic priming relies on reliable, well-studied stimuli with defined similarity values. In the last twenty years, the publication rates of normed stimuli databases and corpora (i.e., large bodies of text) has exponentially increased. Further, newer computational models of concept representation have been detailed using these databases. Using these newer models, we can define similarity between concepts to create reliable stimuli for study in semantic priming. In this proposal, we outline the need for a database of semantic priming values, particularly in non-English languages. We detail the process for creating a large database of priming values, from which new theories and hypotheses can be examined. Further, we describe the novel outputs that this proposal will support including a framework for determining sample size for studies of this nature.

SPAM-L: Semantic Priming Across Multiple Languages

Experimental psychologists have long understood that the stimuli in a research study are of great importance, and that controlled sets of normed information hold great value in creating precisely measured effects. While many areas of psychology use word and picture stimuli that could benefit from selection from normed datasets, this project will focus specifically on cognitive psychology, psycholinguistics, and computational linguistics. Often, stimuli for studies were created in small pilot studies, which were then used in many subsequent projects. Both Lucas (2000) and Hutchison (2003) provided evidence that the interpretation of small pilot data should be carefully considered in the context of larger, more reliable datasets. For example, both noted that it was previously unclear how to interpret semantic priming study results because the measurement of the stimuli similarity was inconsistent. As noted in Buchanan, Valentine, and Maxwell (2019a), we have seen an explosion in publications outlining normed datasets in the last ten to twenty years. Advances in computational ability, the growth of large-scale data collection via the Internet, and the focus on replication and reproducibility have propelled this research forward. The importance of normed stimuli for research cannot be overstated - not only do they allow for control in methodology for studies using the stimuli, the stimuli themselves can be studied and used for understanding memory and cognitive architecture (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019; De Deyne, Navarro, Perfors, & Storms, 2016; Vankrunkelsven, Verheyen, Storms, & De Deyne, 2018; Vitevitch, Goldstein, Siew, & Castro, 2014).

These projects are usually time-consuming to collect, process, clean, and validate - anecdotally, authors of two large datasets in word meaning (Buchanan, Valentine, & Maxwell, 2019b) and word association (De Deyne et al., 2019) have discussed that each took nearly ten years to complete. Mechanical Turk, an online paid participant pool, has eased the data

collection process by allowing researchers to pay users to provide data (Buhrmester, Talaifar, & Gosling, 2018), although recent events have questioned the continued use of this data collection method (Chandler & Paolacci, 2018; Hauser, Paolacci, & Chandler, 2019). Cognitive psychologists have successfully used Mechanical Turk to collect data in word pair norms (Buchanan et al., 2019b), concreteness ratings (Brysbaert, Warriner, & Kuperman, 2014), valence (Dodds, Harris, Kloumann, Bliss, & Danforth, 2011; Warriner, Kuperman, & Brysbaert, 2013), age of acquisition ratings (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), and past tense information (Cohen-Shikora, Balota, Kapuria, & Yap, 2013). Data is often mined from online sources, such as subtitles (Brysbaert & New, Boris, 2009), Twitter (Gimenes & New, 2016; Kloumann, Danforth, Harris, Bliss, & Dodds, 2012) and Google books (Michel et al., 2011). Even with these advances in data collection, the lack of overlap between datasets can make it difficult to methodologically examine multiple variables of interest. Further, conducting cross-language studies on the same data is very difficult, even with the increase in publication of normed data in non-English languages (Buchanan et al., 2019a).

Therefore, the purpose of this study is to leverage the power and network of the Psychological Science Accelerator to provide a cross-language normed dataset of semantic priming, paired with other useful psycholinguistic variables. Semantic priming is a well-studied cognitive phenomenon wherein participants are shown a prime word (e.g., the first word shown, *doctor*) followed by either a related (i.e., *nurse*) or unrelated (i.e., *tree*) target word (Meyer & Schvaneveldt, 1971). Semantic priming is defined as the decrease in response latency (i.e., faster linguistic processing) for target words that were related to their prime words. Participants are often asked to judge the lexicality of a concept (i.e., is this a word?, lexical decision task) or to read the word aloud (naming task). The Semantic Priming Project was a large scale semantic priming study that focused on providing data for lexical decision and naming tasks for 1661

English words (Hutchison et al., 2013). While this study has been cited for using the stimuli, only a few studies have investigated the information provided in the data (Heyman, Hutchison, & Storms, 2016; Mandera, Keuleers, & Brysbaert, 2017; Yap, Hutchison, & Tan, 2017). Potentially, the limitation in using these data is the somewhat bounded overlap between this project and other linguistic variables, as well as the complete lack of a large dataset available in a non-English language. For example, when analyzing the Semantic Priming Data for Buchanan (Buchanan, 2019), several hundred stimuli had to be excluded, even though the variables used were those provided by the Semantic Priming Project and other studies designed to overlap completely.

Semantic priming research spans nearly fifty years of study as a tool to investigate cognitive processes, such as word recognition, and to elucidate the structure and organization of knowledge representation (McNamara, 2005). Results from priming tasks are used to define the mental lexicon, often by developing theoretical and computational models that capture the effects shown empirically (Cree & Armstrong, 2012; Mandera et al., 2017; McRae & Jones, 2013; Rogers, 2008). Priming has also been used in studies of attention (Frings, Schneider, & Fox, 2015; Spruyt, De Houwer, Everaert, & Hermans, 2012), bi/multilinguals (McDonough & Trofimovich, 2011; Singh, 2014), psychological disorders (Copland, 2003; Haverkort, 2017; Tan, Neill, & Rossell, 2015), and in a large body of neuroscience studies (Kiefer et al., 2011; Liu, Wu, Meng, & Dang, 2013; Steinhauer, Royle, Drury, & Fromont, 2017).

Semantic priming has generally been studied at the group level comparing related to unrelated conditions, largely ignoring the variability of items, participants, and languages, and it is likely that these facets are tied to the variability found in semantic priming (Buchanan, 2019; L. L. Jones & Golonka, 2012). Recently, Heyman, Bruninx, Hutchison, and Storms (2018) explored the reliability of priming effects, deeming them mostly unreliable, in contrast to their

previous study on the Semantic Priming Project (Heyman et al., 2016). However, it should be noted that they discuss 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 large scale dataset. Further, Hutchison, Balota, Cortese, and Watson (2008) demonstrated that priming effects can be predicted at the item level, albeit with a smaller dataset.

While behavioral priming has received the most criticism for replication efforts (Cesario, 2014), semantic priming research can also fail to replicate, leading to the methodological questions described above, which focused on the characteristics of the subjects and items used in a study. Large-scale data in this area is sparse, unlike the other published databases found in Buchanan et al. (2019a). Therefore, this study aims to provide data that complements and extends the published data, which would encourage research on methodology, item characteristics, models, cross-language consistency in priming, and other theoretical areas that semantic priming has been applied to previously. The global aims of this project include:

1) Create an online framework to collect semantic priming data, modeled after the success of the Small World of Words project (De Deyne et al., 2019). The online framework would allow data collection from any internet capable computer, thus lowering the burden on research labs to collect data of this nature. The online framework can then be used to deliver updates to the data, even after the conclusion of the initial data collection.
2) Provide a large dataset of response latencies and priming scores for prime and target words in at least five[1] of the nine languages with available frequency data, depending on

[1] The experiment would be designed with all nine languages, but the proposal is for PSA support to recruit for at least five of the languages depending on lab availability. Data collection for any languages that do not reach criterion will continue after publication of the first wave of data with publication at a later date.

recruitment and availability. Further, these prime and target words will be supplemented with variables that are theoretically important for research in cognitive architectures to provide a dataset with less missing data. The dataset provided allows researchers to continue to use these datasets to select carefully controlled stimuli, as well as investigate questions about items, participants, reliability, and language.

Two secondary aims of this project also include:

3) Provide an *R* package and Shiny web interface to use and explore the data modeled after the *LexOPS* package (Taylor, 2019). Stimuli selection remains a primary use for lexical data, and this package would include all the current data from the project (linked to the online framework noted above). Researchers could use the package/Shiny web interface to create stimuli relevant for their future studies or to import the current data for investigation.
4) Encourage the use of lexical datasets as "participants" by engaging in a secondary data analysis challenge focused on using the dataset for hypothesis testing when presenting the data to the larger scientific community.

Method

Participants

Data from the English Lexicon Project (Balota et al., 2007) and the Semantic Priming Project (Hutchison et al., 2013) were used to estimate the minimum sample size necessary for the study. The aim of this study is to provide a large dataset, rather than test a hypothesis, so traditional ways to estimate sample size via power and effect size were not applicable. Therefore, an accuracy in parameter estimation approach was employed using the previous data as a metric. In this approach, one focuses on finding a confidence interval around a parameter that would be "sufficiently narrow" (Kelley, 2007; Kelley, Darku, & Chattopadhyay,

2018; Maxwell, Kelley, & Rausch, 2008). Both the English Lexicon Project and Semantic Priming Project used a lexical decision task, which will be employed in this study. These data were used to estimate the likely standard errors of lexical decision data for individual words. These values were used as the rubric of accurately measured lexical decision response latencies.

Given proposed standard error value, 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 selected, and samples starting at $n = 5$ to $n = 400$ 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. From this calculation, *n* for each target concept was estimated at 100-320 participants. The design of the study, the number of words per session, expected data loss due to incorrect answers, number of target words desired, and number of required conditions were all taken into account and the final estimate for sample size per language is 741 to 4741. The complete code and description of this process is detailed at: https://github.com/doomlab/SPAML/tree/master/parameter_estimation.

This sample size estimation represents a major improvement from previous database collection studies, as many have used the traditional $n >= 30$ as a way to guess at minimum sample size. As indicated, it's often unclear how to exactly estimate a sample size for these types of studies, and this study will detail that procedure to provide guidance for future work. The upper range of estimated participants is high because of the uncertainty in estimating an "accurate" parameter. 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 minimized. The minimum sample size will be 50 participants per concept or 741 total

participants, and the maximum will 320. After 50 participants, each concept will be examined for standard error, and data collection for that concept will be stopped when the standard error reaches and average of the two metrics found in this exploration (0.06, 0.012; see supplemental material) or 0.09. This process will be automated online and checked in a daily subroutine. From the current simulations, this approximates to 100-150 participants per word, and 1482-2223 participants per language total.

Materials

Semantic priming focuses on word-pair relatedness or similarity, and therefore, prime-target pairs are often chosen for their similarity in the related condition. The unrelated condition pairs are then created by shuffling the prime-target pairs so that the prime word is combined with a target word it has no relationship to. Non-words are created by changing one to two letters in a prime or target word to create a nonsense word (*nurse* → *lurse*), with the stipulation that they must be pronounceable and not pseudo-homophones (i.e., wherein the pronunciation sounds like a real word, *keep* → *keap*). Consequently, the choice of related words is key for the study. There are multiple measures of semantic similarity including the cosine between overlapping features (Buchanan et al., 2019b), free association probabilities (De Deyne, Navarro, & Storms, 2013), and local/global coherence values from network models (Siew & Vitevitch, 2016; Vitevitch et al., 2014). However, the underlying data for these calculations is spotty across languages. Therefore, one solution is to use the SUBTLEX projects to calculate lexical co-occurrence as a measure of semantic similarity. The SUBTLEX projects are large corpora (i.e., many words) collected from movie subtitles in nine languages: American and British English (British Brysbaert & New, Boris, 2009; van Heuven, Mandera, Keuleers, & Brysbaert, 2014), Dutch (Keuleers, Brysbaert, & New, 2010), Simplified Chinese (Cai & Brysbaert, 2010), Spanish (Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2012), German (Brysbaert

et al., 2011), Greek (Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010), Polish (Mandera, Keuleers, Wodniecka, & Brysbaert, 2015), Italian (Crepaldi, Amenta, Pawel, Keuleers, & Brysbaert, 2015), and French (New, Boris, Brysbaert, Veronis, & Pallier, 2007).

With the subtitle data, we will take the first 10000 most frequent nouns, adjectives, adverbs, and verbs from each language, and these will be cross-referenced using the *translateR* package (C. Lucas & Tingley, 2014). Next, a distributional space model for each language will be created to identify concepts related to the 10000 most frequent words and to calculate their respective similarity values (Mandera et al., 2017). The top five most related words will be selected, and these will be cross-referenced across languages. Native speakers will be recruited to ensure the accurate translation of word pairs. The related word pairs (n = 1000) will be selected from the list using each concept only once, favoring pairs with translations in most languages. If a selected pair does not exist in a language, translation from a Native speaker will be used to create that pair. Words will also be cross referenced for polysemy (i.e., multiple meanings) and these will be restricted when possible. Lastly, concepts will be examined for their relative statistics on lexical measures (length, part of speech, neighborhood, phonemes/morphemes) and subjective measures (age of acquisition, imageability, concreteness, valence, dominance, arousal, and familiarity) because of their known associations with concept representation. The Wuggy program will be used to create nonwords from the final list (Keuleers & Brysbaert, 2010). A short demonstration of this selection procedure can be found at: https://github.com/doomlab/SPAML/tree/master/parameter_estimation.

Procedure

A small demonstration of the experiment can be found at: https://open-lab.online/code/PSA%20LDT%20Example/?generate=true. The study will be

programmed using lab.js (Henninger, Shevchenko, Mertens, Kieslich, & Hilbig, 2019), which is an online, open-source study creation project. Precise timing measurement is required for this study, and the lab.js team has documented the accuracy of measurement within their framework (Henninger, Shevchenko, Mertens, Kieslich, & Hilbig, 2018), and previous work has shown no differences between lab and web-based data collection for response latencies (Hilbig, 2016). In addition, SPALEX, a large lexical decision database in Spanish was collected completely online (Aguasvivas et al., 2018). We will recommend that research labs use Chrome as their browser, however, meta-information about the browser and operating system are saved when participants take the experiment to control for implementation differences. Participants will be directed to an online web portal to take the study, and all data will be retained in the online platform with nightly backups to GitHub. They will be asked to indicate their gender (male, female, other, prefer not to say), year of birth for age, and education level (none, elementary school, high school, bachelors, masters, doctorate) for demographic variables. To continue in the study, they will select their primary language, which will direct them to the appropriate stimuli set.

Participants will be required to complete the study on a computer, rather than a mobile or tablet device. 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. 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 concept will then be displayed in the middle of the screen in uppercase San-Serif font (i.e., NURSE). On the bottom of the screen the answer choices will be shown as the traditional keys next to the *shift* key depending on the common keyboard layout for that language (i.e., Z and /

on a QWERTY keyboard or > and - on a QWERTZ keyboard)[2]. These choices will be reversed in half of the subjects, which will be randomly selected at the start of the study to counterbalance word/nonword selection. Participants will enter their choice for each concept, and then the next word will appear with an intertrial interval of 500 ms (i.e., the time between the offset of the first concept and onset of the next concept, when the fixation cross is showing). Responses will time out after 5 seconds and move on to the next trial. After ten 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 six blocks of 100 trials (600 words), 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 (course credit, payment), and they will be given instructions on how to indicate they have completed the study to the appropriate lab. Participants will be allowed to take the study multiple times (see below). These values will be customized based on data collection type (i.e., Mechanical Turk, participant pool, etc.). An estimate for the amount of time required for the study is approximately twenty to thirty minutes including practice trials, instructions, and breaks.

This procedure can be considered single stream lexical decision task wherein every concept (prime and target) are judged for lexicality (i.e., word/nonword). Many priming studies often present prime words for a short period of time prior to the presentation of target words for lexicality judgement. Evidence from the Semantic Priming Project suggests that the stimulus onset asynchrony (time between non-judged prime word and target word) did not affect overall priming rates (25 versus 23 ms for 200 ms and 1200 ms). Further, adding the lexicality judgment

[2] We left the demonstration with the *s* and */* keys, as it appeared most consistent across keyboards for the proposal.

to each presented concept creates a less obvious link between prime and target. Even though they appear sequentially in the task, they are not explicitly paired by being a non-judged prime word followed by a judged target word. Additionally, this procedure varies from the data collected in the Semantic Priming Project, thus, extending their work to different conditions.

A primary goal of this project is to provide a complete dataset of priming and other important related linguistic variables. 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, such as age of acquisition (approximate age you learned a concept), imageability (how easy the concept is to imagine), concreteness (how concrete is the concept), valence (emotion), arousal, dominance (controlled versus dominated), and familiarity. For concepts that are missing these values in a target language, participants will be asked to provide ratings on a single metric (i.e., they would only see instructions for familiarity or arousal). Each participant will be asked to provide 25-50 ratings of concepts, given the need for a particular language, while also controlling for the length of the task to prevent fatigue in the experiment. These will only be presented at the end of the experiment to prevent interactions with priming effects. We will use the available large databases of these variables to estimate sample size necessary for these ratings using the same simulation procedure detailed above.

Results

An example of the data and processing for English can be found at https://github.com/doomlab/SPAML/tree/master/data_processing.

Trial/subject data

Files containing the entire data from the experiment will be available for download from the experiment website, GitHub, and through an *R* package as part of the publication. 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* → *katze* → *gatta*). All data will be archived on GitHub, 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. Participants are expected to incorrectly answer trials, and these trials will be marked for potential exclusion. Further, computer errors or trials due to missing data (i.e., participant inattentiveness and timeout trials, internet disconnection, computer crashes) will be marked as such in the final data with NA values. The response latencies from each participant's session will be *z*-scored in line with recommendations from Faust, Balota, Spieler, and Ferraro (1999). 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, and the recommended *z*-score procedure should control for subject variability at this level; therefore, repeated participation would not be detrimental to data collection.

Item data

An item level data file will also be prepared for publication and data releases. The item file will contain lexical information about all stimuli (length, frequency, orthographic neighborhood, bigram frequency). The descriptive statistics calculated from the trial level data will then be included: average response latency, average standardized response latency, sample size, standard errors of response latencies, and accuracy rate. For averages and standard errors, the incorrect and missing trials will be excluded. No data will be excluded for being a potential outlier, however, we will recommend cut off criterion for *z*-score outliers at 2.5 and 3.0 and 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 indicated because these values do not exist for nonwords.

Priming data

In a separate file, we will also prepare information about priming specific results. Priming is defined as the subtraction of average *z*-scored related response latency for an item from the corresponding item in the unrelated condition. The similarity scores calculated during stimuli selection will be included, 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 (Montefinese, Ambrosini, Fairfield, & Mammarella, 2013), German (Kremer & Baroni, 2011), Spanish (Vivas, Vivas, Comesaña, Coni, & Vorano, 2017) and Dutch (Ruts et al., 2004)**.**

Conclusion

In conclusion, this proposal would support the creation of a large dataset on semantic priming across multiple languages through the use of open source technologies and an online web portal. The Small World of Words project has seen success with this setup, and with PSA support, we can accelerate the data collection process to provide high quality data at a faster rate. This dataset will contain complete norms with coverage across variables that have been denoted as theoretically important for cognitive science research, thus supporting a secondary goal to participate in a data analysis challenge using the data upon release. Last, we will develop tools to help all researchers use the data, including the creation of an *R* package with corresponding Shiny web interface for users of all types. Other beneficial, novel outputs from this project include a framework for estimating sample size for database collections, and the creation of distributional models in non-English languages that can be provided for others to download (for example, see http://meshugga.ugent.be/snaut/).

References

Aguasvivas, J. A., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A. (2018). SPALEX: A Spanish lexical decision database from a massive online data collection. *Frontiers in Psychology*, *9*, 2156. https://doi.org/10.3389/fpsyg.2018.02156

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

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

Brysbaert, M., & New, Boris. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/brm.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Buchanan, E. M. (2019). *Using the Semantic Priming Project to understand variability in priming*. https://doi.org/10.17605/OSF.IO/P7BH5

Buchanan, E. M., Valentine, K. D., & Maxwell, N. (2019a). LAB: Linguistic Annotated Bibliography – a searchable portal for normed database information. *Behavior Research Methods*, *51*(4), 1878–1888. https://doi.org/10.3758/s13428-018-1130-8

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019b). English semantic feature

production norms: An extended database of 4436 concepts. *Behavior Research Methods*, *51*(4), 1849–1863. https://doi.org/10.3758/s13428-019-01243-z

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *13*(2), 149–154. https://doi.org/10.1177/1745691617706516

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS One*, *5*(6), e10729. https://doi.org/10.1371/journal.pone.0010729

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *9*(1), 40–48. https://doi.org/10.1177/1745691613513470

Chandler, J. J., & Paolacci, G. (2018). *Lie for a Dime: When most prescreening responses are honest but most "eligible" respondents are lies*. https://doi.org/10.31234/osf.io/mcvwk

Cohen-Shikora, E. R., Balota, D. A., Kapuria, A., & Yap, M. J. (2013). The past tense inflection project (PTIP): speeded past tense inflections, imageability ratings, and past tense consistency measures for 2,200 verbs. *Behavior Research Methods*, *45*(1), 151–159. https://doi.org/10.3758/s13428-012-0240-y

Copland, D. (2003). The basal ganglia and semantic engagement: potential insights from semantic priming in individuals with subcortical vascular lesions, Parkinson's disease, and cortical lesions. *Journal of the International Neuropsychological Society: JINS*, *9*(7), 1041–1052. https://doi.org/10.1017/S1355617703970081

Cree, G. S., & Armstrong, B. C. (2012). Computational Models of Semantic Memory. In M. Spivey, K. McRae, & M. Joanisse (Eds.), *The Cambridge Handbook of Psycholinguistics* (pp. 259–282). https://doi.org/10.1017/cbo9781139029377.014

Crepaldi, D., Amenta, S., Pawel, M., Keuleers, E., & Brysbaert, M. (2015). SUBTLEX-IT. Subtitle-based word frequency estimates for Italian. *Proceedings of the Annual Meeting of the Italian Association For Experimental Psychology*, 10–12.

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

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006. https://doi.org/10.3758/s13428-018-1115-7

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology. General*, *145*(9), 1228–1254. https://doi.org/10.1037/xge0000192

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*(2), 480–498. https://doi.org/10.3758/s13428-012-0260-7

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behavior: the case of greek. *Frontiers in Psychology*, *1*, 218. https://doi.org/10.3389/fpsyg.2010.00218

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, *6*(12), e26752. https://doi.org/10.1371/journal.pone.0026752

Faust, M. E., Balota, D. A., Spieler, D. H., & Richard Ferraro, F. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*(6), 777–799.

https://doi.org/10.1037//0033-2909.125.6.777

Frings, C., Schneider, K. K., & Fox, E. (2015). The negative priming paradigm: An update and implications for selective attention. *Psychonomic Bulletin & Review*, *22*(6), 1577–1597. https://doi.org/10.3758/s13423-015-0841-4

Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, *48*(3), 963–972. https://doi.org/10.3758/s13428-015-0621-0

Hauser, D., Paolacci, G., & Chandler, J. J. (2019). Common concerns with MTurk as a participant pool: Evidence and solutions. In F. R. Kardes, P. M. Herr, & N. Schwarz (Eds.), *Handbook of Research Methods in Consumer Psychology*. https://doi.org/10.4324/9781351137713-17

Haverkort, M. (2017). Linguistic representation and language use in aphasia. In A. Cutler (Ed.), *Twenty-First Century Psycholinguistics* (pp. 57–68). Routledge.

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P., & Hilbig, B. E. (2018, November). *Who said browser-based experiments can't have proper timing? Implementing accurate presentation and response timing in browser*. Presented at the Society for Computers in Psychology, New Orleans, LA. Retrieved from https://lab.js.org/resources/performance/

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2019). lab.js: A free, open, online study builder. Retrieved from https://psyarxiv.com/fqr49

Heyman, T., Bruninx, A., Hutchison, K. A., & Storms, G. (2018). The (un)reliability of item-level semantic priming effects. *Behavior Research Methods*, *50*(6), 2173–2183. https://doi.org/10.3758/s13428-018-1040-9

Heyman, T., Hutchison, K. A., & Storms, G. (2016). Uncovering underlying processes of semantic priming by correlating item-level effects. *Psychonomic Bulletin & Review*, *23*(2), 540–547. https://doi.org/10.3758/s13423-015-0932-2

Hilbig, B. E. (2016). Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724. https://doi.org/10.3758/s13428-015-0678-9

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review*, *10*(4), 785–813. https://doi.org/10.3758/BF03196544

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *Quarterly Journal of Experimental Psychology* , *61*(7), 1036–1066. https://doi.org/10.1080/17470210701438111

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., … Buchanan, E. M. (2013). The semantic priming project. *Behavior Research Methods*, *45*(4), 1099–1114. https://doi.org/10.3758/s13428-012-0304-z

Jones, L. L., & Golonka, S. (2012). Different influences on lexical priming for integrative, thematic, and taxonomic relations. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00205

Kelley, K. (2007). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods*, *39*(4), 755–766. https://doi.org/10.3758/BF03192966

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*(2), 226–243. https://doi.org/10.1037/met0000127

Keuleers, E., & Brysbaert, M. (2010). Wuggy: a multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: a new measure for Dutch word

frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Kiefer, M., Ansorge, U., Haynes, J.-D., Hamker, F., Mattler, U., Verleger, R., & Niedeggen, M. (2011). Neuro-cognitive mechanisms of conscious and unconscious visual perception: From a plethora of phenomena to general principles. *Advances in Cognitive Psychology*, *7*, 55–67. https://doi.org/10.2478/v10053-008-0090-4

Kloumann, I. M., Danforth, C. M., Harris, K. D., Bliss, C. A., & Dodds, P. S. (2012). Positivity of the English language. *PloS One*, *7*(1), e29484. https://doi.org/10.1371/journal.pone.0029484

Kremer, G., & Baroni, M. (2011). A set of semantic norms for German and Italian. *Behavior Research Methods*, *43*(1), 97–109. https://doi.org/10.3758/s13428-010-0028-x

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

Liu, B., Wu, G., Meng, X., & Dang, J. (2013). Correlation between prime duration and semantic priming effect: evidence from N400 effect. *Neuroscience*, *238*, 319–326. https://doi.org/10.1016/j.neuroscience.2013.02.010

Lucas, C., & Tingley, C. (2014). translateR: Bindings for the Google and Microsoft translation APIs (Version 1.0). Retrieved from https://CRAN.R-project.org/package=translateR

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, *7*(4), 618–630. https://doi.org/10.3758/BF03212999

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). 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.

https://doi.org/10.1016/j.jml.2016.04.001

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). Subtlex-pl: subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, *47*(2), 471–483. https://doi.org/10.3758/s13428-014-0489-4

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

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

McNamara, T. P. (2005). *Semantic Priming: Perspectives from Memory and Word Recognition*. Psychology Press.

McRae, K., & Jones, M. (2013). Semantic Memory. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology*. https://doi.org/10.1093/oxfordhb/9780195376746.013.0014

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. https://doi.org/10.1037/h0031564

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013). Semantic memory: a feature-based analysis and new norms for Italian. *Behavior Research Methods*, *45*(2), 440–461. https://doi.org/10.3758/s13428-012-0263-4

New, Boris, Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661–677.

https://doi.org/10.1017/s014271640707035x

Rogers, T. T. (2008). Computational Models of Semantic Memory. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 226–266). https://doi.org/10.1017/cbo9780511816772.012

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 506–515. https://doi.org/10.3758/BF03195597

Siew, C. S. Q., & Vitevitch, M. S. (2016). Spoken word recognition and serial recall of words from components in the phonological network. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(3), 394–410. https://doi.org/10.1037/xlm0000139

Singh, L. (2014). One world, two languages: cross-language semantic priming in bilingual toddlers. *Child Development*, *85*(2), 755–766. https://doi.org/10.1111/cdev.12133

Spruyt, A., De Houwer, J., Everaert, T., & Hermans, D. (2012). Unconscious semantic activation depends on feature-specific attention allocation. *Cognition*, *122*(1), 91–95. https://doi.org/10.1016/j.cognition.2011.08.017

Steinhauer, K., Royle, P., Drury, J. E., & Fromont, L. A. (2017). The priming of priming: Evidence that the N400 reflects context-dependent post-retrieval word integration in working memory. *Neuroscience Letters*, *651*, 192–197. https://doi.org/10.1016/j.neulet.2017.05.007

Tan, E. J., Neill, E., & Rossell, S. L. (2015). Assessing the Relationship between Semantic Processing and Thought Disorder Symptoms in Schizophrenia. *Journal of the International Neuropsychological Society: JINS*, *21*(8), 629–638. https://doi.org/10.1017/S1355617715000648

Taylor, J. (2019). LexOPS: A package and Shiny app for generating psycholinguistic stimuli

(Version 0.0.0.9005) [R]. Retrieved from https://github.com/JackEdTaylor/LexOPS/

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Vankrunkelsven, H., Verheyen, S., Storms, G., & De Deyne, S. (2018). Predicting lexical norms: A comparison between a word association model and text-based word co-occurrence models. *Journal of Cognition*, *1*(1). https://doi.org/10.5334/joc.50

Vitevitch, M. S., Goldstein, R., Siew, C. S. Q., & Castro, N. (2014). Using complex networks to understand the mental lexicon. *Yearbook of the Poznan Linguistic Meeting*, *1*(1), 119–138. https://doi.org/10.1515/yplm-2015-0007

Vivas, J., Vivas, L., Comesaña, A., Coni, A. G., & Vorano, A. (2017). Spanish semantic feature production norms for 400 concrete concepts. *Behavior Research Methods*, *49*(3), 1095–1106. https://doi.org/10.3758/s13428-016-0777-2

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Yap, M., Hutchison, K. A., & Tan, L. C. (2017). Individual differences in semantic priming performance: Insights from the Semantic Priming Project. In M. N. Jones (Ed.), *Frontiers of cognitive psychology. Big data in cognitive science* (pp. 203–226). New York: Psychology Press.