Taboo language across the globe: A multi-lab study

Simone Sulpizio, Fritz Günther, Linda Badan, Benjamin Basclain, Marc Brysbaert, Yuen Lai Chan, Laura Anna Ciaccio, Carolin Dudschig, Jon Andoni Duñabeitia, Fabio Fasoli, et al.

Note: This is the author's preprint version of the article (date: 21.01.2024) The final article is accepted for publication in *Behavior Research Methods.*

Taboo language across the globe: A multi-lab study

Simone **Sulpizio***[a,b], Fritz **Günther***[c], Linda **Badan**[d], Benjamin **Basclain**[e], Marc **Brysbaert**[f], Yuen Lai **Chan**[g], Laura Anna **Ciaccio**[h], Carolin **Dudschig**[i], Jon Andoni **Duñabeitia**[j], Fabio **Fasoli**[k,l], Ludovic **Ferrand**[m], Dušica **Filipović Đurđević**[n], Ernesto **Guerra**[o], Geoff **Hollis**[p], Remo **Job**[q], Khanitin **Jornkokgoud**[r], Hasibe **Kahraman**[e], Naledi **Kgolo-Lotshwao**[s], Sachiko **Kinoshita**[e], Julija **Kos**[t], Leslie **Lee**[u], Nala H. **Lee**[u], Ian Grant **Mackenzie**[i], Milica **Manojlović**[n], Christina **Manouilidou**[t], Mirko **Martinic**[o], Maria del Carmen **Méndez**[v], Ksenija **Mišić**[n], Natinee **Na Chiangmai**[r], Alexandre **Nikolaev**[w], Marina **Oganyan**[x], Patrice **Rusconi**[y], Giuseppe **Samo**[z], Chi-shing **Tse**[g], Chris **Westbury**[1], Peera **Wongupparaj**[2], Melvin J. **Yap**[3], & Marco **Marelli***[a,b]

[a]Department of Psychology, University of Milano-Bicocca, Milan, Italy

[b]Milan Center for Neuroscience (NeuroMI), University of Milano-Bicocca, Milan, Italy

[c]Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

[d]Department of Humanities, University of Trento, Trento, Italy

[e]School of Psychological Sciences, Macquarie University, Sydney, Australia

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

[g]Department of Educational Psychology, The Chinese University of Hong Kong, Hong Kong, China

[h]Brain Language Laboratory, Department of Philosophy and Humanities, Freie Universität Berlin, Berlin, Germany

[i]Department of Psychology, University of Tübingen, Tübingen, Germany

[j]Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Madrid, Spain

[k]School of Psychology, University of Surrey, Guildford, United Kingdom

[l]Centro de Investigação e Intervenção Social, Instituto Universitário de Lisboa, Lisbon, Portugal

[m]Laboratoire de Psychologie Sociale et Cognitive, CNRS, Université Clermont Auvergne, Clermont-Ferrand, France

[n]Department of Psychology, University of Belgrade, Belgrade, Serbia

[o]Center for Advanced Research in Education, Institute of Education, Universidad de Chile, Santiago, Chile

[p]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada

[q]Department of Psychology and Cognitive Science, University of Trento, Trento, Italy

[r]Cognitive Science and Innovation Research Unit (CSIRU), College of Research Methodology and Cognitive Science, Burapha University, Chonburi, Thailand

[s]Faculty of Humanities, University of Botswana, Gaborone, Botswana

[t]Department of Comparative and General Linguistics, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia

[u]Department of English, Linguistics, & Theatre Studies, National University of Singapore, Singapore, Singapore

[v]Facultad de Filosofia y Letras I, Universidad de Alicante, Alicante, Spain

[w]School of Humanities, Foreign Languages and Translation Studies, University of Eastern Finland, Joensuu, Finland

[x]Department of Linguistics, University of Washington, Seattle, United States of America

[y]Department of Cognitive Sciences, Psychology, Education and Cultural Studies University of Messina, Messina, Italy

[z]Department of Linguistics, Beijing Language and Culture University, Beijing, China

[1]Department of Psychology, University of Alberta, Edmonton, Canada

[2]Department of Psychology, Faculty of Humanities and Social Sciences, Burapha University, Chonburi, Thailand

[3]Department of Psychology, National University of Singapore, Singapore, Singapore

Author note

Simone Sulpizio, Fritz Günther, and Marco Marelli equally contributed to the article.

Correspondence concerning the article should be addressed to Simone Sulpizio, Department of Psychology, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milan, Italy, email: simone.sulpizio@unimib.it; Fritz Günther**,** Department of Psychology, Humboldt-Universität zu Berlin, Unter den Linden 6, 10117 Berlin, Germany, email: fritz.guenther@hu-berlin.de; Marco Marelli, Department of Psychology, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milan, Italy, email: marco.marelli@unimib.it;

Abstract

The use of taboo words represents one of the most common and arguably universal linguistic behaviors, fulfilling a wide range of psychological and social functions. However, in the scientific literature, taboo language is poorly characterized, and how it is realized in different languages and populations remains largely unexplored. Here we provide a database of taboo words, collected from different linguistic communities (Study 1, N = 1,046), along with their speaker-centered semantic characterization (Study 2, N = 455 for each of six rating dimensions), covering 13 languages and 17 countries from all five permanently inhabited continents. Our results show that, in all languages, taboo words are mainly characterized by extremely low valence and high arousal, and very low written frequency. However, a significant amount of cross-country variability in words' tabooness and offensiveness proves the importance of community-specific sociocultural knowledge in the study of taboo language.

Keywords: taboo words; swearing; semantics; best-worst scaling; emotion

Everyday communication is full of socially inappropriate words that are considered linguistic taboo. We are taught not to use them in conversations, even though we produce taboo words from the very moment we start speaking (Jay & Jay, 2013), and keep doing it throughout our lives. We also produce them while sleeping (Arnulf et al., 2017) or when acquired language disorders severely impair any other word production (Van Lancker & Cummings, 1999). As adults, 0.5% of the words we produce (i.e., ~80 words per day; Mehl et al., 2006) and 1% of the words we write on Twitter are taboo words (Wang et al., 2014). We use taboo words despite it being socially inappropriate, forbidden, and (in some countries) even legally punished. We do so because taboo language is an extremely powerful linguistic tool that fulfills an unparalleled wide range of psychological and social functions, as no other word category can do. Swearing allows us to: Induce emotional reactions (Sheidlower, 2009), insult others (Croom, 2011), increase the vividness of what is said (Azzaro, 2018), intensify emotional communication (Jay & Janschewitz, 2007), reinforce message effectiveness (Cavazza & Guidetti, 2004), increase the perceived credibility of the speaker (Rassin et al., 2005), regulate emotions and reduce pain (Stephens & Umland, 2011), promote group bonding and reinforce group identity (Daly et al., 2004; Montagu, 2001), and elicit humor (Blake, 2018). Moreover, differently from all other words, taboo words (and in particular swear words) are used almost only with a connotative function (i.e., they do not refer to their literal meaning; Finkelstein, 2018; Jay & Janschewitz, 2008).

We do not all swear the same. Frequency of swearing is associated with personality traits (e.g., high scores of agreeableness and conscientiousness, as measured by the Big Five personality test, are associated with low frequency of swearing; Mehl et al., 2006), social factors (e.g., group identity; Daly et al., 2004), gender (men swear more frequently in public and use more offensive words than women; Jay, 2009) and idiosyncratic pragmatic factors, such as the conversational topic, the setting of the conversation (i.e., public/private,

formal/informal), or the speaker-listener relationship (Jay & Janschewitz, 2008; Johnson & Lewis, 2010).

Studies investigating how taboo words are processed indicate peculiar properties of this category. Taboo words are remembered better than other words (MacKay et al., 2004), capture people's attention (Carretié et al., 2008; MacKay et al., 2004), exert a detrimental effect on word recognition (e.g., Sulpizio et al., 2019) and speech production tasks (e.g., White et al., 2017), require higher level of cognitive control (Dhooge & Hartsuiker, 2011; Scaltritti et al., 2021), increase the arousal level of the sympathetic nervous system (Harris et al., 2003; McGinnies, 1949), and persist in severe acquired language disorders hindering any other linguistic production (Van Lancker & Cummings, 1999).

Despite its wide use and relevance in fulfilling multiple social and psychological functions, we know very little about what taboo language is and what constitutes it across different populations, languages, and cultures. All our empirical knowledge on taboo language comes from a relatively small set of studies, almost entirely conducted in English and with limited cultural diversity. This poses two main theoretical problems. First, taboo language is highly conditioned by sociocultural factors, so what constitutes taboo can only be determined within a specific sociocultural environment. Hence, the currently available evidence offers an extremely restricted picture of the phenomenon. Second, because of this, even the composition, and thus the definition, of the taboo taxonomy is blurred. There is no agreement on the types and the number of categories characterizing taboo words (Jay, 2009; Stapleton, 2009). Finally, related to this last issue, it is still unclear what makes a word taboo. In terms of semantic properties, emotional aspects have been suggested to play a central role (Hansen et al., 2017; Jay & Jay, 2015). However, emotionality might not be enough to precisely characterize taboo words, which would otherwise be indistinguishable from other emotional words. Other properties that are typically considered are offensiveness (i.e., how a

person perceives a word as inappropriate) and tabooness (i.e., how a person believes the society considers that word inappropriate; Jay, 1992). Nonetheless, while the use of the latter property makes the definition tautological, the former seems not to be a necessary property of swearing. For example, the English words *sex* or *vagina* are generally not offensive but are taboo in some social circumstances. In the data presented below, these are the words with the largest discrepancy between tabooness and offensiveness. The specific lexico-semantic characterization of taboo words is still to be determined, as it is still unknown whether and to what extent taboo words can be differentiated from non-taboo words on the basis of their lexical and semantic properties.

The present study aims at providing a first step towards filling these gaps, by collecting and characterizing taboo words in 17 different countries and 13 different languages (including some typically overlooked ones), covering all five permanently inhabited continents. In addition to offering a window into taboo language around the world, our study offers the unique chance to tease apart cross-linguistic from cross-cultural differences by analyzing the behavior of participants that speak country-based varieties of the same language (e.g., English in Canada and in Singapore). Importantly, taboo words in our study are defined in a strictly bottom-up manner based on speakers' productions (Study 1). This allows us to establish what each community actually considers taboo without introducing any bias due to the researchers' idiosyncrasies and normative definitions, and to identify commonalities and differences across languages and countries. In Study 2, we systematically collect intuitions about several semantic measures for each of the produced words to determine the combination of semantic features that best characterize the taboo dimension, and to evaluate their consistency across languages and cultures. Taken together, our results achieve two important goals: Theoretically, they contribute to a better general definition and understanding of taboo words and swearing across languages and cultures. Methodologically,

they form a very rich database to study taboo language both *per se* and in relation with its several social and psychological functions.

Study 1 – Identifying taboo words

Methods

Participants.

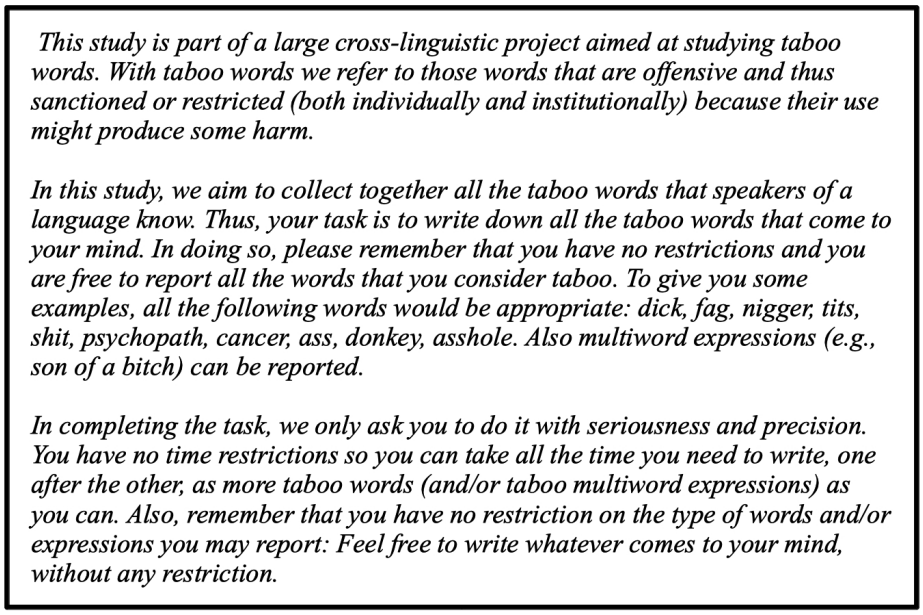
We collected data in 18 labs from 17 countries (Australia, Belgium, Botswana, Canada, China [two labs, in Beijing and Hong Kong], Chile, France, Germany, Finland, Italy, Serbia, Singapore, Slovenia, Spain, Thailand, United Kingdom, United States of America), covering all the five permanently inhabited continents and 13 different languages (Cantonese, Dutch, English, Finnish, French, German, Italian, Mandarin, Serbian, Setswana, Slovenian, Spanish, Thai), with some of these (i.e., English, and Spanish) spoken in multiple countries. These languages are spoken, as native speakers, by more than 2 billion people in the world (i.e., ~25% of the global population, data from Wikipedia).

The total number of participants was 1,046 (see Supplementary Table 1 for details), with each lab collecting data from at least 40 participants (40 to 167). Only native speakers of the language in question who lived in the country in question and who were not suffering from language-related and/or learning disabilities were included. Supplementary Table 1 reports participants' details *per* lab as well as information concerning the ethics approvals obtained by each lab involved in the project.

Procedure.

In each of the labs, a local coordinator managed all the study's aspects. The coordinator was a native speaker of the language in question living in the culture in which data collection occurred or was flanked by another researcher who was a native speaker of the language in question, and was living in the culture in which data collection occurred.

Participants were asked to freely write down all the taboo words they could think of. Both single word and multi-word expressions were accepted, and examples were provided for both cases. There was neither time pressure nor any time restriction to complete the task. In the instructions, we specified that participants were free to write whatever came to their minds and encouraged them to avoid self-censorship. Instructions (in English) were reported in Figure 1.


This study is part of a large cross-linguistic project aimed at studying taboo words. With taboo words we refer to those words that are offensive and thus sanctioned or restricted (both individually and institutionally) because their use might produce some harm.

In this study, we aim to collect together all the taboo words that speakers of a language know. Thus, your task is to write down all the taboo words that come to your mind. In doing so, please remember that you have no restrictions and you are free to report all the words that you consider taboo. To give you some examples, all the following words would be appropriate: dick, fag, nigger, tits, shit, psychopath, cancer, ass, donkey, asshole. Also multiword expressions (e.g., son of a bitch) can be reported.

In completing the task, we only ask you to do it with seriousness and precision. You have no time restrictions so you can take all the time you need to write, one after the other, as more taboo words (and/or taboo multiword expressions) as you can. Also, remember that you have no restriction on the type of words and/or expressions you may report: Feel free to write whatever comes to your mind, without any restriction.

Figure 1: Instructions of Study 1.

The instructions were provided to all the labs, which were asked to translate them in the local language and then back-translate them into English (translation and back-translation were not required for labs collecting data in English). Translation and back translation were provided by different persons, and the back translation was compared to the original version as a sanity check. Details about the data collection modality for each lab are reported in Supplementary Table 1.

For each sample, all participants' productions were combined. In each lab a researcher went through the list and: a) checked all the productions and corrected for possible minor

errors (e.g., typos); b) for non-English languages, provided an English translation of each word; c) on the basis of their intuition as native-speaker knowledgeable of the respective culture, classified each word (using a simplified taxonomy based on Jay (2009)) as belonging to one of the following categories for which definitions were provided to researchers: insult; slur; sexual; scatological referents and disgusting objects; profanities/blasphemies. When appropriate, researchers were invited to classify the same production in more than one category. When words could not be classified within the existing categories, researchers were allowed to create new categories. All annotated data are available at https://osf.io/ecr32/?view_only=60b964248cc64a8793a9013075132a1c.

Note that, since there was no one speaker knowing *all* involved languages, we cannot guarantee that the very same classification and translation criteria were applied for all languages. Therefore, the information collected in b) and c) only provides a pointer to the word meaning, so that readers who do not speak the language have an opportunity to understand all the items in the dataset. However, we emphasize that this information should only be considered as a general reference and treated very carefully for any form of quantitative analysis.

Statistical Considerations.

In the linear mixed effect model (LMM) analyses reported here, the 18 different samples served as our basic unit of observation; therefore, all LMMs reported here contain random intercepts for the samples in addition to the fixed effects specified in the individual analyses. We estimated the LMMs in R (R Core Team, 2022) using the packages *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017).

Results

In Study 1, participants from 17 countries (see Figure 2) were asked to freely generate any taboo words they could think of. The total number of words produced varies greatly between samples (see Figure 3).

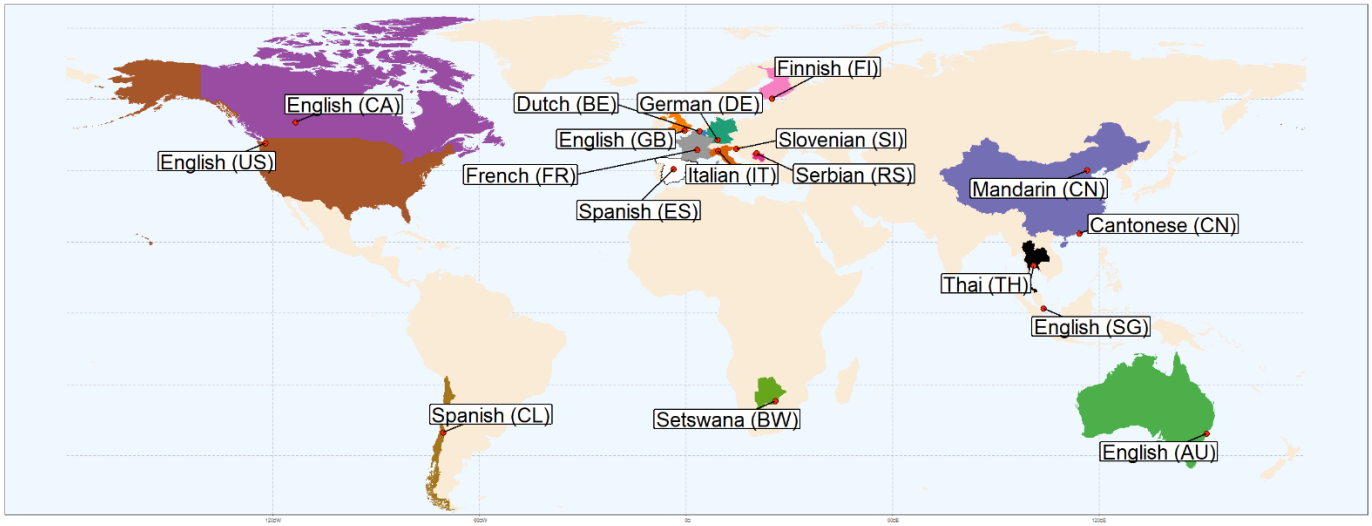


Figure 2: Map of the countries involved in Study 1 and 2. The 18 labs, and the respective 17 countries and 13 languages in which data were collected.

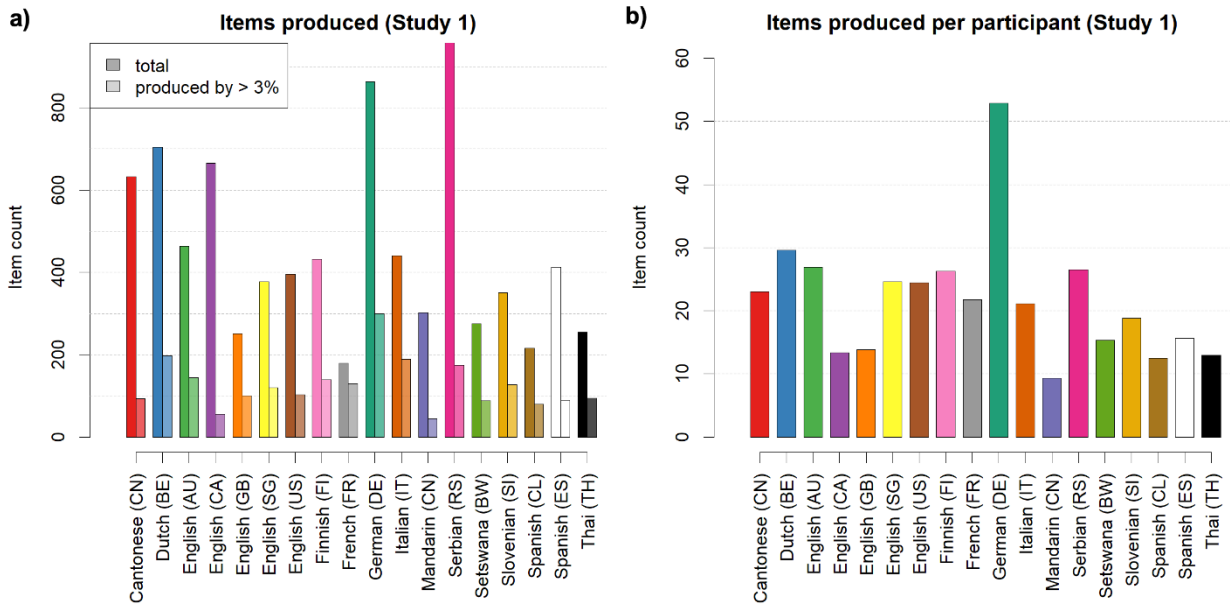


Figure 3: Number of items produced in Study 1 for each lab. a) Total number of items (darker colors) and the subset of items produced by at least 3% of all participants (lighter colors). **b)** Average number of items produced per participant.

In a qualitative exploratory analysis, to assess the cross-language variability in our data, we manually inspected the 10 most frequently produced words in each sample and categorized them by means of their English translations (treating words with near-synonym translations as the same word). As can be seen in Figure 4, there is a certain degree of consensus across samples: Some words are found among the most frequent words in many if not most languages. Variations of *cunt* (especially when also considering *mother's cunt*) were seen in all samples, and those of *bitch* in almost all samples. Six additional items (*dick, faggot, nigger, fuck, shit,* and *ass*) were produced by about half of the samples. Also, with only a few exceptions, most samples produce around one third to a half of the 17 items (6-10 items), again suggesting some overlap in participants' intuitions across languages and cultures. Note that almost *all* of these words are produced by some participants in every sample, but not frequently enough to appear among the ten most frequent words.

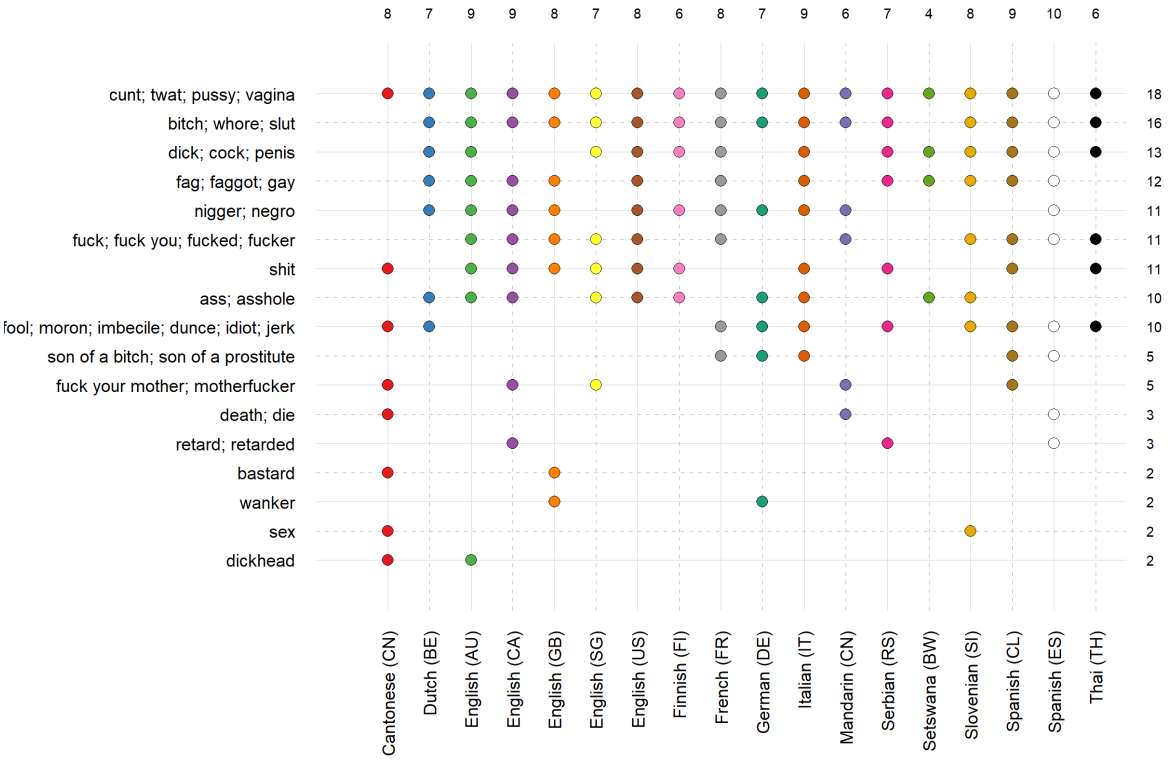


Figure 4: Samples in which a word is among the 10 most frequently produced ones. The figure shows the samples in which a word (or a semantically closely related word) is among the 10 most frequently produced words in Study 1, alongside the number of samples for which the word has been produced (right) and number of these items produced per sample (top). This figure only includes words appearing in the top-10 in at least two samples. Note that the taboo words reported in the figure refer to sets of meaning-related words (not exact translations), that were created on the basis of our intuition, and should only be considered as qualitative.

Study 2 – Evaluating taboo words

Methods

Participants.

Data were collected by the same labs for all the same languages and countries as in Study 1. The total number of participants was 455 for each of the six rating dimensions (see below), with the number of participants per lab depending on the length of their item list (see below). Only participants who self-declared to be in the 18-40 age range (the age range was fixed to this window as taboo language use may vary with age; Barbieri, 2008; Thelwall, 2008; note that this age range is the most typically used in the majority of psychological and cognitive science research), a native speaker of the language, and not suffering from language-related and/or learning disabilities were included. For each lab, approximately half of the participants self-identified as male and the other half self-identified as female. Supplementary Table 2 reports all participants' demographic information for each sample and rating dimension.

Materials.

The taboo stimuli were based on the results of Study 1. We selected only those stimuli that were produced by at least 3% (rounded down) of participants (note that for samples up to 66 participants, this criterion is equivalent to only excluding *hapax legomena*, i.e., words that were produced by only one participant). Using a relative threshold allowed us both to account

for between-samples variability in number of participants, and to keep item sets manageable also for labs that had a large participant sample in Study 1.

To further keep the data collection manageable for all labs, the maximum number of taboo words to be included was set to 252. In cases where the procedure based on the 3% production criterion threshold produced less than 252 items, all taboo words above that threshold were included. On the other hand, in cases where the selection procedure based on the 3% production criterion threshold produced more than 252 words, the lab was required to apply a slightly more conservative threshold. For example, if following the 3% threshold rule and excluding words produced up to 2 times resulted in a list of 300 words, then the researchers were asked to exclude words produced up to 3 times. In cases where applying this slightly more conservative procedure then again produced a list shorter than 252 words, the researchers were required to fill the list up to 252 words by randomly sampling between the exceeding stimuli (i.e., from those that were produced between 3% and this more conservative threshold; see Supplementary Table 2 for the final number of selected taboo stimuli for each language; the stimuli for each language are available at https://osf.io/ecr32/?view_only=60b964248cc64a8793a9013075132a1c).

Taboo stimuli were presented along with filler non-taboo stimuli. The purpose of including filler words was three-fold: i) introducing variety in the item set to mitigate any list effect during the rating task (for example, avoid the lists to only include very negative words; this is especially relevant when using the best-worst paradigm to collect ratings, as described below); (ii) providing in the resource norms for non-taboo elements, that can serve as matched control items for future studies; and (iii) allowing us to directly assess the external reliability of our ratings by comparing them to other existing word norms, which often do not include a large number of taboo words.

The total number of filler words was half of the number of taboo words eventually selected, rounded up. Filler words were selected to be representative examples of the language under investigation with respect to the two most important emotional dimensions, namely valence and arousal. Note that, as a representative sample, they are not matched to the taboo words, which we expect to score relatively higher on these dimensions – this expectation is confirmed in the analysis predicting the taboo word status of words (see the Results section of Study 2). Moreover, any a-priori matching between taboo and filler words was impossible because of the lack of norms for taboo words in most of the languages under investigation.

To prepare the filler set, each coordinator was asked to rely on existing word norms. Where available, the adapted version of the ANEW database (Bradley & Lang, 1999) was used and filler words were selected in order to have a distributional profile of valence and arousal (qualitatively) similar to that of the whole ANEW database. If the ANEW database was not available, the researchers could use any other database reporting information about valence and arousal and follow the same selection procedure described above. If no database with emotional dimensions was available for a given language, the English version of the ANEW database was used as inspiration, and then the selection was based on the intuition of a native-speaker researcher. In the case where the taboo stimuli included multi-word expressions, some filler multi-word expressions were also included (with the same proportion as the multi-word taboo expressions in the list). The list of filler stimuli for each language is available at https://osf.io/ecr32/?view_only=60b964248cc64a8793a9013075132a1c.

Instructions and debriefing information (in English, see Supplementary methods) were provided to all the labs, which were asked to translate them into the local language and then back-translate them into English (translation and back-translation were not required for

labs collecting data in English). Translation and back-translation were provided by different persons, and the back-translation compared to the original version served as a sanity check.

Procedure.

We collected ratings on six dimensions:

- Age of acquisition (when a word was learned; Brysbaert & Biemiller, 2017)
- Concreteness (to what extent a word referent can be perceived with the senses; Brysbaert et al., 2014)
- Valence (how pleasant a word referent is; Warriner et al., 2016)
- Arousal (the amount of excitement evoked by a word referent; Warriner et al., 2016)
- Offensiveness (how offensive a rater finds the word personally; Janschewitz, 2008)
- Tabooness (how taboo a word is, defined as to what extent it is *not* acceptable to use it in most social situations; Janschewitz, 2008)

Rating data were collected using the best-worst scaling technique (Hollis, 2018; Hollis & Westbury, 2018). In best-worst scaling studies, participants are presented with *N* items in each trial, and have to select the item which, in their opinion, scores the highest and the item which scores the lowest on a given dimension (for example, the least and most offensive item in the presented list). This produces information for 2(*N*-1) pairwise comparisons per trial (the "best" item versus all other items, and the "worst" item versus all other items). Presenting each item in many different constellations of such sets of *N* items provides implicit rank information, making it possible to induce a rating scale from these best-worst judgments (Hollis & Westbury, 2018).

The optimal choice in terms of data quality and data-collection efficiency has been identified as $N = 6$ items per trial (Hollis, 2020). For this reason, the total number of items for each lab had to be divisible by 6. In case this criterion was not naturally met by the selection of taboo and filler stimuli (see Materials section above), a few additional fillers were added until the criterion was met.

We collected 30 observations for each individual item (that is, each item was presented in 30 different sets of 6 items; see Hollis, 2018, Experiment 4), which is the sufficient amount for near-asymptotic elimination of measurement errors (Hollis & Westbury, 2018). We used the software provided by Hollis (2018) to optimally arrange the item lists into sets of 6, in a no-repetition setup that avoids presenting a combination of two words in more than one set, if possible (see Hollis, 2018, Experiment 4). Note that each participant was thus presented with a unique list of item sets as their trials.

Since the task could be tiring for participants, we decided that each single data collection session should not exceed 15 minutes (which corresponds to about 45 trials). This meant that, in a lab with 252 taboo words and 126 fillers, at least 42 data collection sessions (and thus, the same number of participants) per semantic dimension were needed (252 data collection sessions in total). For smaller item sets, this number decreases in a linear manner.

In each session, the participant only responded to one of the semantic dimensions (e.g., only arousal, or only offensiveness). Participants could not participate in more than one session for the same dimension, but they could take part in several subsequent sessions, one for each semantic dimension (e.g., a participant was prevented to evaluate "valence" more than once, but could evaluate both "valence", "arousal", and "tabooness" in different sessions if they desired). This allowed us to collect data also in labs with limited access to eligible and willing participants (which can result from organizational, financial, or even political constraints). Participants were informed beforehand that the study would contain words that

many people consider offensive or inappropriate, that they would be directly confronted with very harsh language and offensive material, and that they could stop the experiment at any time if feeling uncomfortable.

All data were collected in a web-based experiment using *jsPsych* (version 6.3.0; de Leeuw, 2015). All labs sent all the experimental materials (including translations of instructions, scales, and debriefing) to one of the authors (FG) who implemented the experiment and sent back to the labs a ready-to-use link for data collection. This guaranteed a standardization of the experimental procedure, which was the same for all the labs. All data collections were hosted on servers of the University of Tübingen (FG's affiliation at the time). Details on data collection modality are reported in Supplementary Table 2.

Constructing the rating scales.

For each dimension in each sample, we estimated rating scores for all items from the explicit best-worst judgments using software provided by Hollis (2018). As a result, the items received a rating score between 0 (always selected as "worst") and 1 (always selected as "best"). Note that this has important implications for the interpretation of these rating scores: Rating scores are inherently relative to the item set for which they were obtained. For example, if there is a clear least offensive item in a list, this item will always receive a low score, even if the list only consists of highly offensive items. This is the reason why we included fillers words that were selected to be overall representative of a language. Thus, the rating scores should be treated like values that are standardized *within each language (*without being directly comparable *between languages)*. In addition, note that also the age of acquisition data is bound between 0 and 1, rather than directly indicating the age at which the word was learned (however, as shown later, correlations between these AoA scores and

traditional AoA ratings are very high, so these scores can be transformed into an actual age via linear regression).

The rating scales were constructed from the judgments using the Value learning algorithm (Hollis, 2018). For each dimension in each sample, we applied this scaling algorithm (a) on the entire dataset, (b) on the data collected from female and male participants separately, and (c) on two equally-sized subsamples of the data, sampled by randomly assigning half of the participants to each subsample, in order to compute a split-half reliability for the rating scales (Günther et al., 2022).

Internal and external reliabilities.

As a first step, reliabilities for each rating dimension were estimated by splitting the participant sample into two random halves, scoring these halves independently, and computing the correlation between these two sets of scores (see Günther et al., 2022). For the rating dimensions of AoA, valence, tabooness, and offensiveness, these reliabilities were high ($r > .80$) for most samples. For the rating dimensions of arousal and concreteness, these reliabilities tended to be overall lower; however, there was considerable variation between samples (see Supplementary analyses and Supplementary Figure 1 for details). Note that this is however not unique to our study; especially arousal reliabilities tend to be lower in other large-scale rating studies using a general vocabulary (Warriner et al., 2013).

We further assessed these reliabilities quantitatively by means of the LMM reported in the "Gender Differences" section of the Results to Study 2. As noted there, this model contained a fixed-effect interaction between type of correlation (by gender vs random split-half) and dimension rated, as well as a random intercept and correlation type random slopes for the samples. The factors *AoA* for dimension rated and *split-half reliability* for correlation type served as reference level for the model; thus, all main effects are interpreted as

differences to these reference levels. The absolute differences between rating dimensions are significant ($F(5,187) = 42.68$, $p < .001$), with considerably lower values on the dimensions "arousal" ($b = -.13$, $t = -4.05$, $p < .001$) and "concreteness" ($b = -.21$, $t = -6.58$, $p < .001$) as compared to the model's intercept (i.e., the combination of the two reference conditions) of $b = .84$. There were no significant effects for the other rating dimensions (offensiveness: $b = 0.03$, $t = 1.09$, $p = .279$; tabooness: $b = 0.04$, $t = 1.22$, $p = .225$; valence: $b = -0.04$, $t = -1.15$, $p = .250$).

In addition to these internal reliabilities, we estimated the external reliability of our rating scores as the correlation with word norms collected in 35 other studies (see Supplementary analyses for more details, including a full list of these norms, their shared dimensions with our dataset, and the number of shared items). Note that, for some languages generally or some semantic dimensions in some languages specifically, no external sources were available. In these cases, our data (notably including the data for the filler items) provide a first resource that can serve as a reference point for future studies.

We observed that external reliabilities were not significantly lower than internal split-half reliabilities. We thus have no evidence to indicate that the participants performing our rating tasks diverged substantially from the participant samples from other studies, despite the differences in item set structure and rating task (see Supplementary analyses for more information).

Word frequencies.

For each item in Study 2, we considered two distinct measures of word frequency. On the one hand, we considered the production frequency in Phase 1 as the percentage of participants producing a given word in the respective sample (to account for differences in Phase 1

participant sample sizes across labs). On the other hand, we considered a written text frequency (the standard variant of word frequency measure, derived from written text corpora). In order to obtain largely comparable measures for written text frequencies, we based our estimates on the WaCKy web corpus family (Kilgarriff et al., 2010) or, where available, the structurally very similar but larger TenTen corpus family (Jakubíček et al., 2013) on SketchEngine (Kilgariff et al., 2014). SketchEngine also provides the possibility to extract frequencies for multi-word expressions. Where applicable and available, we selected subcorpora for a specific language variant (namely for American English, Australian English, British English, Canadian English, European Spanish, and Chilean Spanish). To accommodate for differences in corpus size, all word frequencies were measured as frequency per million tokens. In line with previous research on frequency effects, all written corpus frequencies were log-Laplace-transformed via log(freq + 1) for all analyses reported (Brysbaert & Diependaele, 2013).

Statistical Considerations.

Again, as reported for Study 1, the LMMs (as well as generalized linear mixed effect Models, GLMMs) reported here contain random intercepts for the samples in addition to the fixed effects specified in the individual analyses, unless specified otherwise.

Results

In Study 2, using a best-worst scale (Hollis, 2018; Hollis & Westbury, 2018), participants were asked to evaluate taboo words as well as representative filler items (see Methods for details) on six dimensions: age of acquisition (AoA), concreteness, valence, arousal, offensiveness, and tabooness. We started by qualitatively examining the most taboo and offensive words in each sample to look for possible cross-language and cross-countries

similarities. Then, by means of multiple quantitative analyses, we identified the semantic features that best characterize the taboo dimension and evaluated their consistency across languages and cultures.

For this quantitative analysis, it needs to be considered that although the reliabilities of our rating scales are generally high, some individual reliability scores are also low (see Supplementary Figure 1). Even though this introduces noise in our variables, the analyses presented here mainly serve to describe and illustrate our dataset. Therefore, we included low-reliability rating scales in these analyses. For follow-up studies using our resource to investigate specific theoretical questions, we however strongly recommend researchers to select subsets of our dataset that meet their criteria and requirements, to be careful if using languages with low reliability and, in this case, consider the possibility to collect further data.

Qualitative Examination

To understand what makes a word taboo or offensive, we started with a qualitative examination and manually inspected the semantic content of the 20 most taboo and 20 most offensive words in each sample. Keeping in mind that any qualitative summary will inevitably paint a somewhat oversimplified picture, some patterns do clearly emerge in the data. There are two classes of words that raters in most samples consider particularly taboo or offensive: sex-related words and slurs.

In virtually all samples, sex-related words take a central position, including those referring to specific sexual acts (*ass fuck*, *blowjob*), or genitalia (*cunt, dick*). Especially prominent are words that include forms of sexual violence or abuse (*rape, pedophile*), or words of all aforementioned categories that involve family members, including sexual behavior towards the addressee's family members as in *"I fuck your mother*", sexual behavior of the other towards their family members as in *motherfucker*, the sexual behavior of their

family members as in "*son of a bitch"*, or references towards their family members' genitalia as in *"Your mother's smelly cunt*". While present in the highest-rated words across all samples, sex-related words dominate the lists in the East Asian (China, Singapore, Thailand), Slavic (Slovenia, Serbia) and Spanish-speaking samples (Chile, Spain). A remarkable pattern within this class of words is a severe gender bias. The list of the most taboo/offensive words is almost exclusively populated by words referring to female genitalia (*cunt*) or female family members (*motherfucker, "I fuck your mother", son of a bitch*). While the male counterparts to some of these words are also produced (*dick, cock, "I fuck your father"*, words for male sex workers), their offensiveness/tabooness ratings are not as high as for the female versions, and they consequently do not appear in the top-20 lists discussed here.

The second major class of highly offensive and/or taboo words are slurs that mostly refer to race (*nigger, wog*), gender (*bitch*, *whore*), and gender/sexual orientation (*fag, tranny*). Across many samples, racial slurs tend to address Black and Jewish people, as well as relevant ethnic groups in the respective countries (for example, slurs for Mexican people in the US, Pakistani people in the GB, or Arab people in Belgium). Notably, slurs take a very prominent position in the Anglosphere (Australia, Canada, Singapore, GB, US) and Central European countries (Belgium, France, Germany, Italy, Slovenia).

A third class of frequent slurs refer to a violation of societal group expectations. Sexist slurs (*bitch, whore*) refer to women's sexual looseness that goes against traditional gender roles while homophobic (*faggot, dyke*) and transphobic (*tranny*) slurs refer to the violation of heteronormative and cisgender expectations related to sexual orientation and gender identity, respectively.

Beyond these major classes of taboo words, words referring to mental and/or physical disabilities (*retard*, *spastic*) are found in the most offensive/taboo words across most languages. Some other common types of words that appear across multiple different samples

are death wishes, either directed at the addressees themselves (*"go die"*) or their family members (*"may your whole family die"* from Australia, Cantonese-speaking China, Belgium, and Spain), words referring to specific political ideologies (*Nazi* from Germany, France, Italy, and the US), or words suggesting low personal qualities of the addressee (*stupid, cheap type, ugly* from Cantonese-speaking China, Chile, Finland, Serbia, and Thailand). An interesting observation is that blasphemies (*fucking god, shit Christ*) only appear in the 20 most taboo and offensive words in Italy, while they are absent from all other lists inspected here.

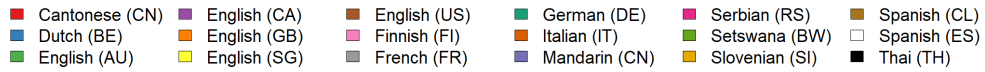
Quantitative Analyses

Gender differences. We first examined if there were structural differences when participant samples were split by gender (male vs. female) instead of randomly (which would point to systematic gender differences). To this end, we fitted a linear mixed-effects model (LMM) with a fixed-effect interaction between type of correlation (by gender vs. random split-half) and dimension rated, as well as a random intercept and by correlation-type random slopes for the samples. This is the same LMM as reported in the *Internal and external reliabilities* section in the Methods of Study 2. The factors *AoA* for dimension rated and *split-half reliability* for correlation type served as reference level for the model.

Across languages and rating dimensions, the correlations between male and female rating scores were comparable to the random split-half reliabilities. The fixed effect for correlation-type was not significant (b = -0.13, $F(1,187) = 2.97$, p = .086) and neither was its interaction with rating dimension ($F(5,187) = 0.12$, p = .988), indicating no structural gender differences beyond the expected random variation between any two groups of raters. Note, however, that the rating scores obtained with the best-worst scale are inherently relative to

the item set for which they were obtained and thus do not allow making direct absolute group comparisons concerning the *mean values* of these ratings.

Relations between the semantic dimensions and frequencies. Figure 5 shows the pairwise Pearson correlations between semantic dimensions for the taboo words (i.e., without fillers), and (a) the production frequency from Study 1 and (b) the written corpus frequency. Some clear general trends can be observed across all labs (for example, a strong negative correlation between valence and offensiveness), but also a few cases with considerable variability (for example, when considering the correlation between concreteness and offensiveness).



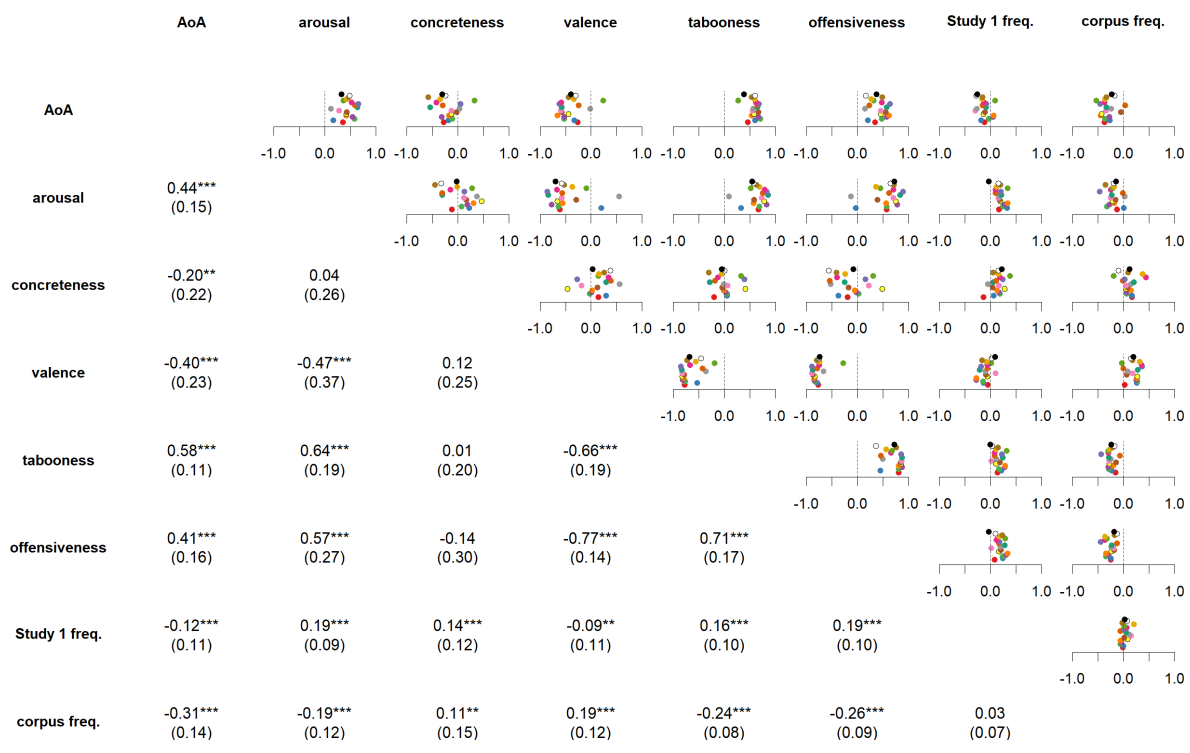


Figure 5: Correlations between rating dimensions and frequency measures. Pairwise Pearson correlations between all rating dimensions and the production frequency in Study 1 (*Study 1 freq.*) and the written corpus frequency (*corpus freq.*). The upper triangle shows correlation values for single samples (each sample represented by a colored circle); lower triangle represents the means of these correlations with their standard deviation in parentheses. Mean correlations significantly different from zero ($p < .001$ in a t-test) are marked with ***.

First, it is worth noting that, although there is some degree of association between offensiveness and tabooness, these variables are far from overlapping (mean $r = .71$). In addition, the two frequency variables show interesting patterns. First, the correlations between Study 1 production frequency and written corpus frequency are not significantly different from zero across all samples (mean $r = .032$, $t(17) = 1.84$, $p = .083$). While written word frequency is typically a good proxy for familiarity (Baayen et al., 2006; Balota & Chumbley, 1984), this relation clearly breaks down for taboo words. Taboo words produced more frequently in Study 1 are arguably those that speakers are more familiar with. However, these results indicate that a frequently produced taboo word in Study 1 is not frequently used

in linguistic corpora. While one might suspect that participants in Study 1 refrained from producing very familiar taboo words because they are too offensive or taboo (i.e., self-censorship), this explanation appears unlikely given the consistently *positive* correlations between Study 1-production frequency and offensiveness and tabooness (see Figure 5).

The relation between these variables and written frequency is the opposite. Tabooness and offensiveness show consistent negative correlations with written corpus frequency (see Figure 5), indicating that "worse" words are less likely to appear in written language. This fits the very definition of tabooness. One is not supposed to produce taboo words in public language and the written corpus frequencies were collected from publicly accessible sources. Thus, the dissociation between how written frequency and production frequency pattern with tabooness and offensiveness shows that speakers are aware of which words *should no*t be used in public language and tend to avoid using them, but *can easily* produce them when explicitly asked to. This is substantiated by data on word prevalence (i.e., the number of people knowing the word) showing that, among the 491 unique English taboo expressions of Study 1 present in Brysbaert et al.'s (2019) prevalence list, 86% are known by more than 90% of the speakers.

In an additional analysis, we re-examined the prominent finding that the relation between the dimensions of valence and arousal is U-shaped rather than linear (Yik et al., 2023), to investigate whether this pattern holds for taboo words as well as non-taboo filler words across the wide variety of samples in our dataset. Although there is some heterogeneity across languages (see Supplementary analyses for details), overall, a U-shaped pattern indeed emerges for both classes of words across all samples (see Figure 6).

What makes a word taboo/offensive? Beyond exploring bivariate correlations, a main objective of our study was to test which lexical and semantic variables are systematically associated with tabooness and offensiveness. We investigated this question in three steps.

In the first step, we set up an LMM to predict the tabooness/offensiveness rating values of all words (taboo words and fillers) from the other rating dimensions (valence, arousal, concreteness, AoA) and written corpus frequency. Study 1 production frequency was not included as a predictor, because obtaining this variable requires instructing speakers to specifically produce taboo words, rendering it far less general than the other variables considered here. To further investigate dissociations between the tabooness and offensiveness dimensions, our LMM additionally included interaction terms between each of the described fixed effect predictors and a dummy variable coding for tabooness vs. offensiveness ratings (and a by-sample random slope for this dummy variable). The results of this analysis are displayed in Table 1.

		LMM predicting tabooness/offensiveness			GLMM predicting taboo word status		
Predictor type	Predictor	*b*	*t*	*p*	*b*	*z*	*p*
Intercept	Intercept	0.398	32.43	< .001	1.834	3.43	< .001
Main effects	*Dummy*	0.349	20.15	< .001			
	Valence	-0.515	-50.77	< .001	-9.586	-19.08	< .001
	Arousal	0.496	40.68	< .001	13.111	19.65	< .001
	Concreteness	0.068	6.58	< .001	-1.322	-2.77	< .001
	AoA	0.181	20.04	< .001	-2.364	-5.81	< .001
	Corpus freq.	-0.009	-10.53	< .001	-0.309	-7.90	< .001
Interactions	*Dummy*: Valence	-0.209	-14.38	< .001			
	Dummy: Arousal	-0.142	-8.26	< .001			
	Dummy: Concreteness	-0.139	-9.55	< .001			
	Dummy: AoA	-0.200	-15.65	< .001			

	Dummy: Corpus freq.	-0.003	-2.39	.017			

Table 1. Results of the statistical models. *LMM predicting tabooness/offensiveness*: Predictors of tabooness and offensiveness ratings across samples, for the dataset of all words (words produced in Study 1 and fillers). "Dummy" is a dummy variable coding for tabooness ratings (coded as 0, the reference condition) or offensiveness ratings (coded as 1); therefore, the intercept and main effects except "dummy" describe tabooness ratings, while the "dummy" effect and all interactions describe how offensiveness ratings differ from tabooness ratings. *GLMM predicting taboo word status*: Predictors of taboo word status across labs (1: taboo, 0: filler).

As can be seen in Table 1, the main predictors for tabooness are valence (higher tabooness ratings for lower valence) and arousal (higher tabooness ratings for higher arousal). In addition, tabooness ratings are higher for words that are more concrete, are learned later (higher AoA), and appear less often in written language. Moreover, the influence of all these predictors is different for offensiveness ratings. For these, we found stronger negative effects of valence and written corpus frequency, as well as weaker positive effects of arousal, concreteness (to the point where the effect is negative; b = -0.07, t = -6.93 for the main effect in the same model with offensiveness as the reference condition for "dimension"), and AoA (to the point where it is slightly negative; b = -0.02, t = -2.08 for the main effect when offensiveness is the reference condition). Since the variables identified here could just be the ones telling apart taboo from non-taboo words, we repeated the analysis only on taboo words. The same results emerged (see Supplementary analyses).

In a second step, to specifically identify the factors discriminating taboo words from non-taboo (filler) words, we set up a generalized linear mixed-effects model (GLMM) as a categorical regression model to predict the taboo vs. non-taboo status, using the same fixed effects of the previous models (valence, arousal, concreteness, and AoA ratings, as well as written corpus frequency). All fixed effects significantly predicted taboo word status (see Table 1). A word is more likely to be a taboo word for higher values of arousal, but less

likely for higher values of all other dimensions. Note, however, that in a bivariate comparison, AoA ratings tend to be higher for taboo words than for fillers (M = .541 vs. M = .414); therefore, the negative effect of AoA is likely the result of collinearity with other predictors. In terms of effect size, the z parameters in Table 1 show that low valence and high arousal are by far the most relevant predictors of taboo word status, followed by low written corpus frequency (see Figure 6). The results of this analysis are thus very similar to the analysis predicting the tabooness of words.

The accuracy of this GLMM in predicting taboo word status is very high (ACC = .863, significantly higher than the No Information Rate NIR = .664, p < .001; $F1$ = .900[38])[1]. This is mainly because the very low valence and very high arousal of taboo words lie considerably outside the regular range of non-taboo (filler) words (see Supplementary analyses and Supplementary Figure 2). Thus, by neglecting or even explicitly excluding taboo words, standard word norms systematically exclude the low end of the valence dimension and the high end of the arousal dimension.

To reduce overfitting and to investigate how well the effects of these variables predicting taboo status generalize across samples, we replicated this analysis while employing a leave-one-out cross-validation (LOOCV) procedure. We estimated the GLMM on all samples except one and used the parameters of the resulting model to classify taboo words in the left-out sample. The accuracy rates in the left-out samples ranged from .750 (Setswana as left-out, $F1$ = .832) to .927 (English (GB) as left-out, $F1$ = .945) (see Figure 6) and were all significantly higher than the NIRs (all ps < .001; Kuhn, 2022). Thus, the results generalize very well across samples.

[1] The F1-score is the harmonic mean of precision (the proportion of true positives in all positive classification results – how many identified items are relevant?) and recall (the proportion of true positives in all true cases – how many relevant items are identified?) for a binary classification

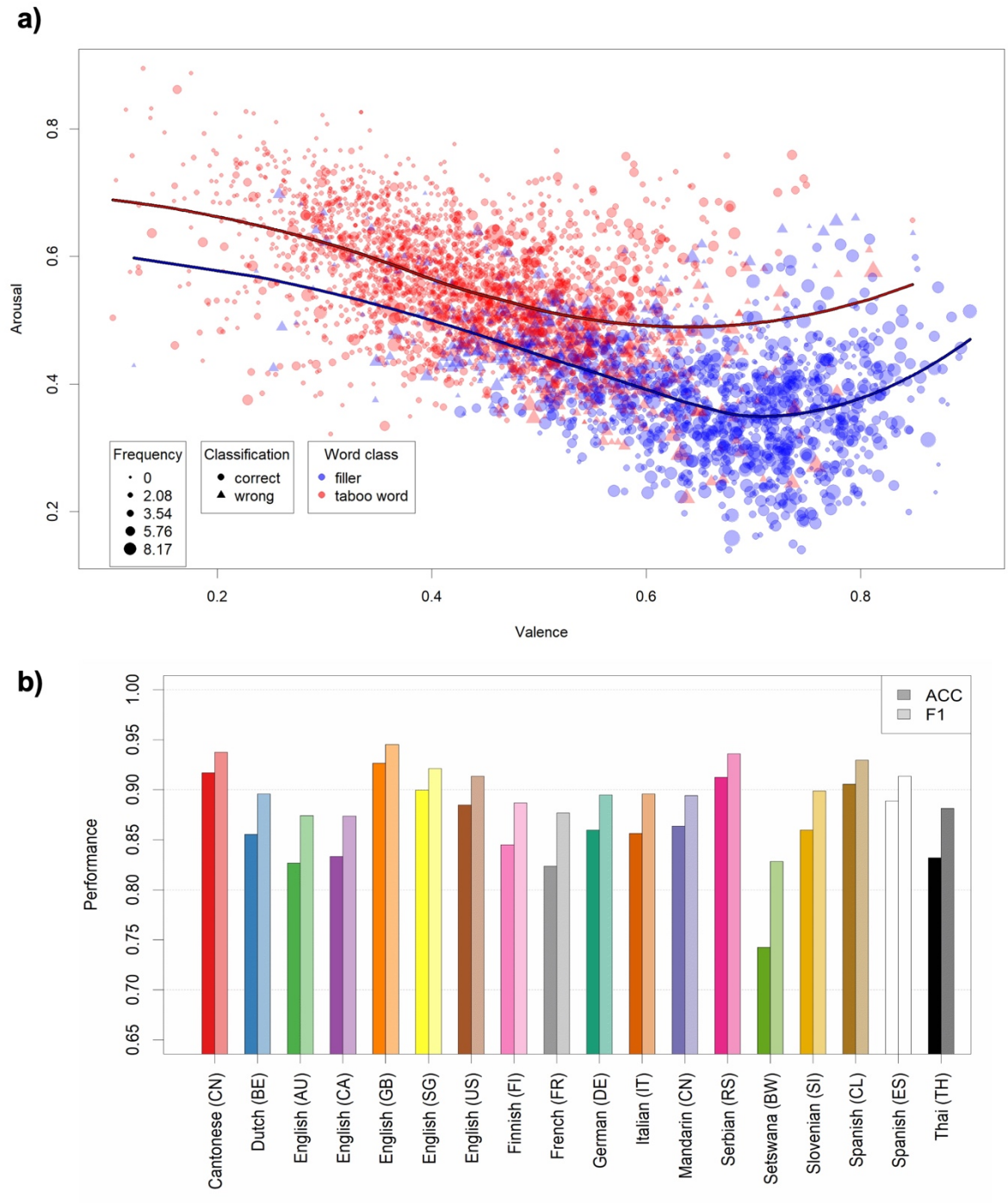


Figure 6: Illustrations of the categorical regression analyses predicting taboo word status from semantic variables and written corpus frequency. a) Distribution of valence (x-axis), arousal (y-axis), and written corpus frequency (point size) for taboo words and fillers (color-coded) in the combined dataset of all samples, alongside their classification accuracy (point type). Regression lines predicting arousal from valence are fitted with local polynomial regression (loess) fitting. **b)** Accuracy rates (darker colors) and F1-scores (lighter colors) for the LOOCV analysis, predicting taboo word status in the left-out sample with a GLMM trained on all other samples (including as predictors valence, arousal, concreteness, AoA ratings, and written corpus frequency).

Differences between varieties of the same language. There are some instances where we collected data for the same languages from different samples around the world (five variants of English and two variants of Spanish). Our data can therefore provide us with an opportunity to disentangle linguistic and sociocultural factors.

Figure 7 displays correlations between the different variants of English on the different rating dimensions, computed on their shared item sets (for the results on Spanish, see Supplementary Figure 3). For context, it should be kept in mind that upper limits to these correlations are posed by the reliabilities of these rating scores (see Supplementary analyses). Although the overall agreement is high, some variability emerges for all dimensions. Note, however, that these correlations are still relatively high in absolute terms (all > .70 for offensiveness, and > .65 for tabooness; all *p*s < .001).

A more fine-grained investigation allows us to explore which individual shared words are perceived similarly or differently across all five variants. Figure 7 and 8 exemplify how the content of the different types of taboo words is shaped by sociocultural sensitivity. Considering Figure 7, for which data from more countries are available: Despite the first two words listed in the figure are those with the highest production frequency, some of them show a very high cross-country variability. For example, while *nigger* is perceived in all countries as extremely offensive in a similar way, it is considered less taboo in the GB than in all other countries. On the contrary, *dick* is perceived as similarly taboo in all countries, but its offensiveness varies a lot. Finally, in some cases, there is an asymmetry between countries in the way a country ranks a word in terms of offensiveness and tabooness. This is the case, e.g., for *retard*, which is more offensive but less taboo in the US than in Canada or Australia. Similar considerations also hold for Spanish (Figure 8): *maricón* is similarly taboo in Spain and in Chile, but is considered more offensive in the former than the latter; on the other hand, *puto* is similarly taboo in the two countries, but more offensive in Chile than in Spain.

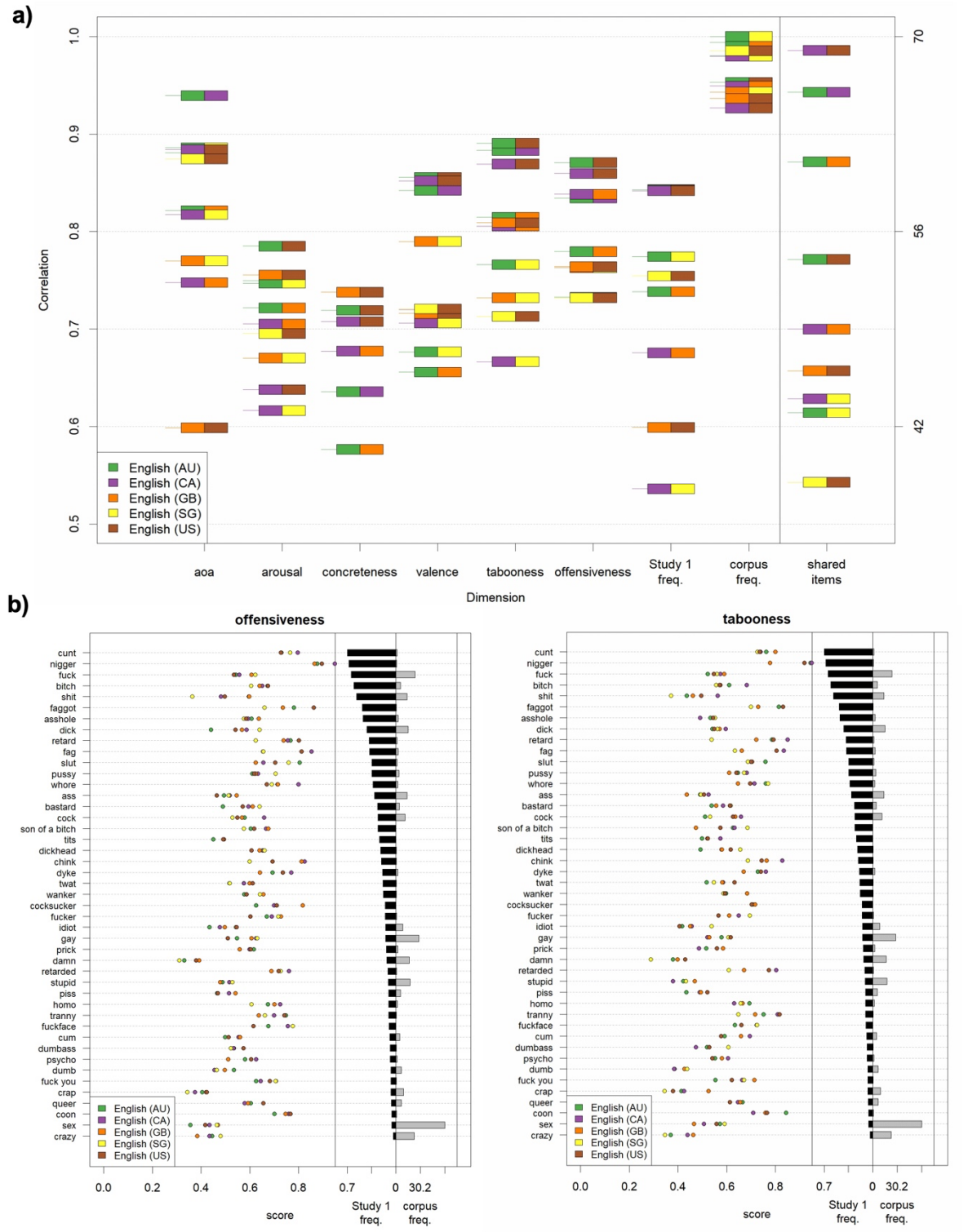


Figure 7: Differences and agreements between different varieties of English. a) Left-hand side: Correlations between the values on each rating dimension, Study 1 production frequency, and written frequency, for the different variants of the same language (English), computed on the shared items between these variants (AU: Australia; CA: Canada; GB: Great Britain; SG: Singapore; US: United States of America); the short horizontal lines indicate the correlation value between a pair of varieties, the box next to the line indicates the pair of

variants for which the correlation was computed. Right-hand side: The number of these shared items between pairs of variants; note that the number of shared items is exactly the same for SG-US and SG-GB, thus the latter is not visible in the plot. **b)** Left-hand side of each plot: Offensiveness (left plot)/tabooness (right plot) ratings by language variant for all the items appearing in at least four out of the five variants of English. Right-hand side of each plot: Production frequency in Study 1 and written corpus frequency for these items (mean values).

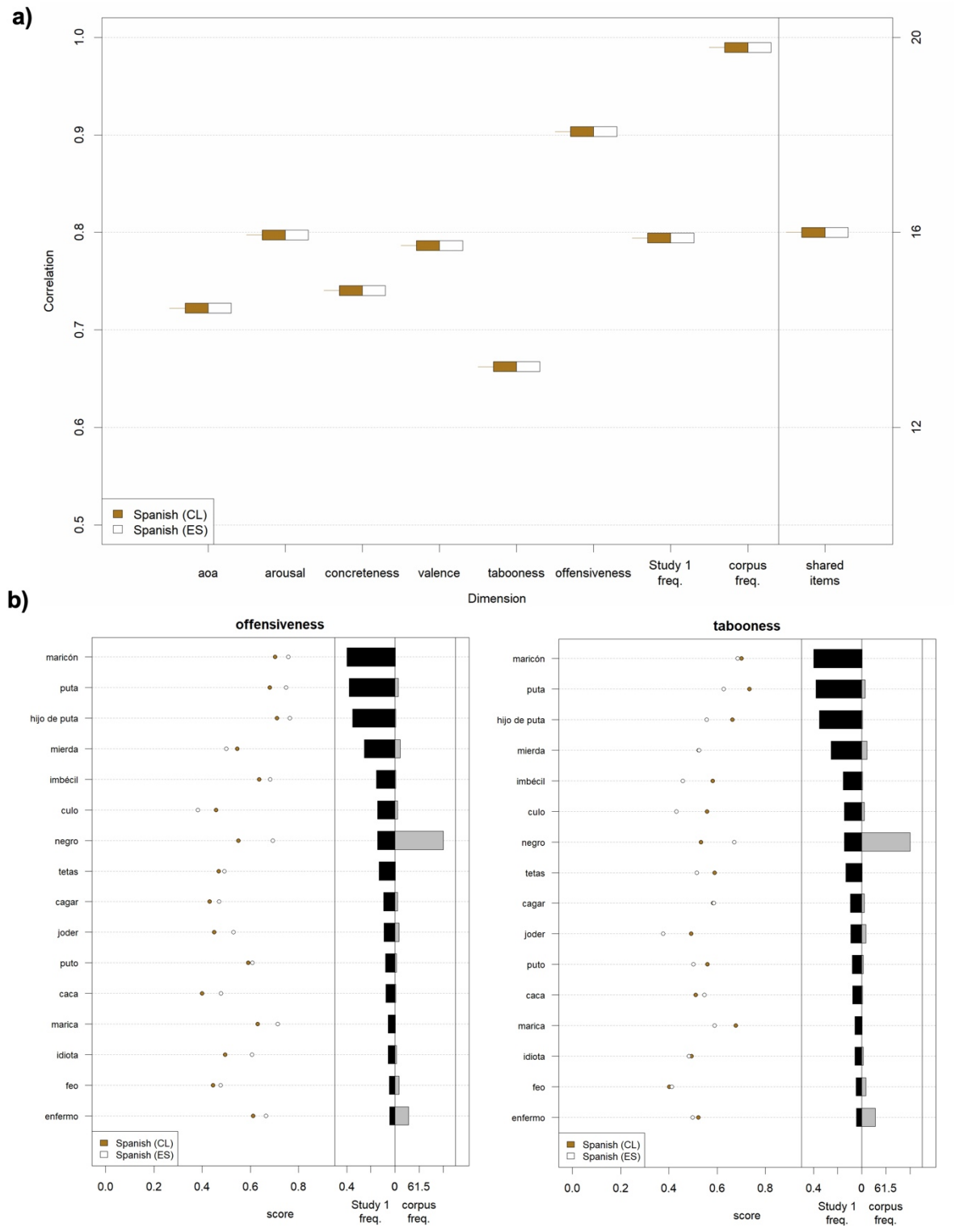


Figure 8: Differences and agreements between different groups of Spanish speakers. a) Left-hand side: Correlations between the values on each rating dimension, Study 1 production frequency, and written frequency for the two different variants of Spanish, computed on the shared items between these variants. Right-hand side: The number of these shared items. **b)** Left-hand side of each plot: Offensiveness (left plot)/tabooness (right plot) ratings by language variant for all items that appear in both variants of Spanish. Right-hand

side of each plot: Production frequency in Study 1 (black bars) and written corpus frequency (grey bars) for these items (mean values).

Discussion

The present study explored, for the first time, the taboo lexicon across several languages and countries, including some typically under-studied languages and populations. We provide a large resource to study taboo words. Moreover, our data offer a rich picture of what taboo language is and improve our knowledge about such a common linguistic behavior in at least three directions: a) identifying what words are considered taboo and to what extent; b) characterizing what makes a word taboo; c) dissociating the impact of culture to vis-a-vis language on the characterization of word tabooness and offensiveness (at least for English). In so doing, our results also contribute to better specify models of swearing (Senberg et al., 2021; Vingerhoets et al., 2013) by better characterizing contextual factors of swearing.

Differently from previous literature (Bertels et al., 2009; Janschewitz, 2008; Roest et al., 2018; Sulpizio et al., 2020), we have identified taboo words using a speaker-based (descriptive) rather than an expert-based (normative) approach. In this way, taboo words were defined on the basis of the intuition and the sensitivity of the linguistic community, in a way directly following from the definition of a taboo as a proscription of behavior for a specific community (Allan & Burridge, 2006; for a similar approach, see Jay, 1992). The results show that, across samples, the category of taboo words ranges between ~50 and ~300 agreed-upon words, with most samples ranging between ~100 and ~200 words. Interestingly, amidst all the variation and differences among samples, there is also some considerable cross-language consistency on which words are seen as most taboo and offensive. Among the words considered the worst (and/or the most forbidden) by people all over the world, we found sex-related terms and slurs. Sex and sexuality have been considered taboo for a long time (Foucault, 1978) in many societies, arising from both social and self-proscription. Sexuality

has to do not only with human body, but also with psychological, social, and moral dimensions (Weeks, 2011), which all may contribute to defining its tabooness. Slurs, on the other hand, explicitly target and degrade social groups (typically minorities). The fact that slurs were common among taboo words highlights the fact that such type of words are used to address individuals and groups 'deviating' from traditional roles or norms. Importantly, slurs represent sort of verbalized thought-crime (Nunberg, 2018) that can have consequences on interpersonal and intergroup relations (Fasoli et al., 2015; 2016). The use of slurs has thus relevant implications for psychological, social, ethical and legislative dimensions (Council of Europe, 2021; Dzenis & Nobre Faria, 2020; Rosenblum et al., 2020).

Sex-related words and slurs show two interesting asymmetries. First, while the former tend to be more represented in the "worst words" in East Asian, Slavic, and Spanish-speaking countries, the latter tend to be more represented in Central European and English-speaking countries. Second, while a bias for sex words emerged such that they were mostly related to women words, slurs pointed to minorities and their 'deviation' from social norms (i.e., women should not be sexually loose/promiscuous, people should be straight, people should be cisgender).

Of note, among the languages and countries we tested, although blasphemy has always been considered one of the main categories of taboo language (Jay, 1992), it is almost completely absent in our sample of "worst words", and relatively infrequent in general. This is in line with diachronic linguistic analyses indicating that religion-related swearwords are becoming more and more socially acceptable, especially in the English-speaking world (Mohr, 2013).

Within each language, taboo words were similarly ranked by males and females, suggesting that they perceive taboo words in a similar way. This is different from what was reported for emotional words (always using the best-worst scale, Mohammed, 2018). The

absence of a gender difference, differently from language production, may be explained by the fact that everyone, regardless of their gender, has learned which words are taboo as this is shared knowledge in a given society (Špago, 2020).

Our results allow us to characterize, from a psycholinguistic perspective, what makes a word taboo. Other than of course being characterized by high scores of tabooness and offensiveness (two dimensions that our data empirically show to be different constructs, and which we excluded from our analysis to predict taboo word status), taboo words are particularly related to valence and arousal – the more taboo and offensive words are also the more negative and high arousing ones. In fact, the very low end of the valence distribution and the very high end of the arousal distribution are essentially only occupied by taboo words. Thus, by not considering taboo words in word norms or item lists, one will systematically ignore the far ends of the spectrum for these variables. At the same time, however, there is still considerable overlap between taboo and non-taboo words at the region of medium-to-high arousal and medium-to-low valence (cf. Figure 6), demonstrating that the categorical distinction between taboo and non-taboo words cannot be based solely on these emotional dimensions. Such a distinction would require the assumption of a clear threshold, which is however absent in our data. Interestingly, tabooness and offensiveness are positively correlated with production frequency, but negatively correlated with written frequency, the latter being the third most important predictor in the characterization of taboo words. This critical dissociation between the two frequency measures clearly shows that the most taboo words are those that everybody would consider as such and be familiar with, but that most speakers would avoid using in a regular public context. In this way, taboo words behave markedly in contrast to non-taboo words, for which written frequency is typically strongly associated with word familiarity (Baayen et al., 2006; Balota & Chumbley, 1984). This has important implications for corpus-based analyses and computational language models, which

typically require a high number of observations to obtain meaningful observations and results. Taken together, valence, arousal, and written frequency are the dimensions mostly contributing to predicting tabooness and offensiveness (with smaller additional contributions of age of acquisition and concreteness), thus being largely sufficient to discriminate between a taboo and a non-taboo word. This emerges across all the samples we investigated (to the extent that taboo status in one sample can be predicted from a classifier trained on the other samples), suggesting that there is a common lexico-semantic characterization of taboo words across languages and cultures.

The availability of data from different countries speaking the same language allows us to dissociate the impact of culture to vis-a-vis language on the characterization of word tabooness and offensiveness. The presence of a significant amount of cross-country variability in these two dimensions (despite the fact that we are considering exactly the same words) is clear proof of cross-cultural variability and indicates that the study of taboo language cannot overlook sociocultural and pragmatics specific knowledge to the community that is to be investigated. Interestingly, these data also offer a hint at the interaction between cross-countries shared tendencies – detected at a superordinate lexico-semantic level – and sociocultural specificity – detected at a subordinate one. At the superordinate level, sex-related terms and slurs are considered taboo by all our samples. At the subordinate level, these two categories are filled according to sociocultural specific norms and sensitivities, showing a certain degree of variability in the category composition.

When used in actual communication, it is likely that all taboo words we report may be offensive, hurting, or violate social norms (after all, participants were explicitly instructed to produce exactly such words). Importantly, the general population might perceive these words as more taboo and offensive than reported here, as all our participants voluntarily produced or exposed themselves, and hence might not have been too bothered by taboo and offensive

words to begin with. Therefore, our data cannot be used in any way to justify any violent and/or offensive behavior – low ratings on offensiveness or tabooness do *not* warrant or excuse the inappropriate use of these words and cannot serve as an argument that persons taking offense "are in the wrong" and that the words "aren't so bad". For the same reason, our results can be relevant for communication policies and strategies to monitor, detect, flag, and potentially prevent offensive linguistic behaviors in social network sites. We identified possible critical categories and words that require specific attention. Importantly, although some categories with cross-linguistic stability (as, e.g., slurs and sex-related terms) might be considered as a general guide for inspection and control, the critical contents should be specifically identified for each language with careful consideration of the words that do or do not harm that specific language community.

Before concluding, we believe is important to highlight some limitations of the present work. A first possible limitation refers to the way in which instructions of Study 1 were phrased, which might have biased participants in their productions. In particular, making explicit reference to offensiveness and giving some examples of taboo words might have either favored some categories (e.g., slurs and insults, which were represented among the examples given in the instructions) or penalized others (e.g., blasphemies, which were not present among the examples). Although we cannot exclude that instruction phrasing might have had an impact on which words were generated by participants, we believe this effect (if any) was limited. First,, participants often generated words that were not mentioned in the instructions (e.g., blasphemies and political terms), suggesting that instructions did not substantially constrain participants in their generation; second, participants did not produce all taboo words provided in the instructions equally often (e.g., *donkey* was sporadically produced despite being one of the provided examples), suggesting that the words included in the instructions did not force participants' productions. Future word-generation studies might

try to test this issue by comparing the productions obtained in the present study with materials obtained through differently-phrased instructions and/or offering different examples.

A second possible limitation of our study is that, although our data show a clear relation between tabooness and emotional dimensions, they are silent about causality. In other words, we do not know whether some words are taboo because they are highly negative and highly arousing, or the other way round. Although some data on this are available in the literature and seem to suggest that emotional connotation is a consequence of tabooness (i.e., taboo words are originally neutral and acquire emotional connotation because they are paired with punishment, e.g., Jay et al., 2006; Jay & Jay, 2013), the evidence is still scanty and mostly based on retrospective studies. Future research should tackle this issue by adopting longitudinal and/or experimental approaches, so that a causal relation (if any) may be detected.

It is also worth noting that, although this study offers an unprecedented set of data on taboo linguistic behaviors all around the world, it still misses out several languages typically unrepresented in the psychological literature. Although the issue might be hard to overcome because dealing with taboos is particularly problematic for some communities, we believe the present work represents an attempt to reduce the gap between the most researched languages and those that have been traditionally neglected, and may directly support future research on this topic, by demonstrating its feasibility and relevance. Because of its rich psychological and sociocultural characterization, the use of taboo language has relevant implications for several fields such as psychology, linguistics, neuropsychology, and neuroscience, but also gender studies, sociology, and anthropology. Although more psycholinguistically-oriented, our results may be of relevance for all these disciplines and boost the study of taboo language, a possibly universal linguistic behavior.

Open Practices Statement

All data and materials for all studies are fully available.

References

Allan, K., & Burridge, K. (2006). *Forbidden words: Taboo and the censoring of language*. Cambridge University Press.

Arnulf, I., Uguccioni, G., Gay, F., Baldayrou, E., Golmard, J. L., Gayraud, F., & Devevey, A. (2017). What does the sleeping brain say? syntax and semantics of sleep talking in healthy subjects and in parasomnia patients. *Sleep, 40*, zsx159.

Azzaro, G. (2018). Taboo language in books, films, and the media. In K. Allan (ed.) *The Oxford handbook of taboo words and language.* Oxford University Press (pp. 285-310).

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*, 290-313.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human perception and performance, 10*, 340-357.

Barbieri, F. (2008). Patterns of age-based linguistic variation in American English 1. *Journal of sociolinguistics, 12*, 58-88.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statiscal Software, 67*, 1-4.

Bertels, J., Kolinsky, R., & Morais, J. (2009). Norms of emotional valence, arousal, threat value and shock value for 80 spoken French words: Comparison between neutral and emotional tones of voice. *Psychologica Belgica, 49*, 19-40.

Blake, B. J. (2018). Taboo language as source of comedy. In K. Allan (ed.) *The Oxford handbook of taboo words and language.* Oxford University Press (pp. 353-371).

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Vol. 30, No. 1, pp. 25-36). Technical report C-1, the center for research in psychophysiology, University of Florida.

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior research methods, 49*, 1520-1523.

Brysbaert, M., & Diependaele, K. (2013). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior research methods, 45*, 422-430.

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior research methods, 51,* 467-479.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods, 46*, 904-911.

Carretié, L., Hinojosa, J. A., Albert, J., López-Martín, S., De La Gándara, B. S., Igoa, J. M., & Sotillo, M. (2008). Modulation of ongoing cognitive processes by emotionally intense words. *Psychophysiology*, *45*, 188-196.

Cavazza, N., & Guidetti, M. (2014). Swearing in political discourse: Why vulgarity works. *Journal of Language and Social Psychology*, *33*, 537-547.

Council of Europe (2021). Combating rising hate against LGBTI people in Europe. *Parliamentary Assembly Council of Europe,* 1-18.

Croom, A. M. (2011). Slurs. *Language Sciences*, *33*, 343-358.

Daly, N., Holmes, J., Newton, J., & Stubbe, M. (2004). Expletives as solidarity signals in FTAs on the factory floor. *Journal of Pragmatics*, *36*, 945-964.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods*, *47*, 1-12.

Dhooge, E., & Hartsuiker, R. J. (2011). How do speakers resist distraction? Evidence from a taboo picture-word interference task. *Psychological Science*, *22*, 855-859.

Dzenis, S., & Nobre Faria, F. (2020). Political correctness: The twofold protection of liberalism. *Philosophia*, *48*, 95-114.

Fasoli, F., Maass, A., & Carnaghi, A. (2015). Labelling and discrimination: Do homophobic epithets undermine fair distribution of resources?. *British Journal of Social Psychology, 54*, 383-393.

Fasoli, F., Paladino, M. P., Carnaghi, A., Jetten, J., Bastian, B., & Bain, P. G. (2016). Not "just words": Exposure to homophobic epithets leads to dehumanizing and physical distancing from gay men. *European Journal of Social Psychology*, *46*, 237-248.

Finkelstein, S. R (2018). Swearing as emotion acts. In A. Pizarro Pedraza (ed.), *Linguistic taboo revisited: Novel insights from cognitive perspectives*. De Gruyter Mouton (pp. 108-139).

Foucault, M. (1978). *The history of sexuality: Vol. 1. An introduction*. Penguin Press.

Günther, F., Marelli, M., Tureski, S., & Petilli, M. A. (2023). ViSpa (Vision Spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. *Psychological Review*, *130*, 896-934.

Hansen, S. J., McMahon, K. L., Burt, J. S., & de Zubicaray, G. I. (2017). The locus of taboo context effects in picture naming. *Quarterly Journal of Experimental Psychology, 70*,

75-91.

Harris, C. L., Ayçíçeğí, A., & Gleason, J. B. (2003). Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language. *Applied Psycholinguistics*, *24*, 561-579.

Hollis, G. (2018). Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior research methods*, *50*, 711-729.

Hollis, G. (2020). The role of number of items per trial in best–worst scaling experiments. *Behavior research methods*, *52*, 694-722.

Hollis, G., & Westbury, C. (2018). When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior research methods*, *50*, 115-133.

Janschewitz, K. (2008). Taboo, emotionally valenced, and emotionally neutral word norms. *Behavior research methods*, *40*, 1065-1074.

Jay, T. (1992). *Cursing in America*. John Benjamins.

Jay, T. (2009). The utility and ubiquity of taboo words. *Perspectives of Psychological Science, 4,* 153-161.

Jay T., Janschewitz K. (2007). Filling the emotion gap in linguistic theory: Commentary on Potts' expressive dimension. *Theoretical Linguistics*, 33, 215-221

Jay T., Janschewitz K. (2008). The pragmatics of swearing. *Journal of Politeness Research*, 4, 267-288.

Jay, K. L., & Jay, T. B. (2013). A child's garden of curses: A gender, historical, and age-related evaluation of the taboo lexicon. *The American Journal of Psychology*, *126*, 459-475.

Jay, K. L., & Jay, T. B. (2015). Taboo word fluency and knowledge of slurs and general pejoratives: Deconstructing the poverty-of-vocabulary myth. *Language Sciences*, *52*, 251-259.

Jay, T., King, K., & Duncan, T. (2006). Memories of punishment for cursing. *Sex Roles*, *55*, 123-133.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013, July). The TenTen corpus family. In *7th international corpus linguistics conference CL* (pp. 125-127).

Johnson, D. I., & Lewis, N. (2010). Perceptions of swearing in the work setting: An expectancy violations theory perspective. *Communication Reports*, *23*, 106-118.

Kilgarriff, A., Reddy, S., Pomikálek, J., & Avinesh, P.V.S (2010). A Corpus Factory for Many Languages. In *LREC workshop on Web Services and Processing Pipelines*, Malta.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubícek, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*, 7-36.

Kuhn, M. (2015). Caret: classification and regression training. *Astrophysics Source Code Library*, ascl-1505.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*, 1-26.

McGinnies, E. (1949). Emotionality and perceptual defense. *Psychological Review, 56*, 244-251.

MacKay, D. G., Shafto, M., Taylor, J. K., Marian, D. E., Abrams, L., & Dyer, J. R. (2004). Relations between emotion, memory, and attention: Evidence from taboo Stroop, lexical decision, and immediate memory tasks. *Memory & cognition*, *32*, 474-488.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, *90*, 862-877.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 174-184).

Mohr, M. (2013). *Holy sh* t: A brief history of swearing.* Oxford University Press.

Montagu, A. (2001). *The anatomy of swearing*. University of Pennsylvania Press.

Nunberg, G. The social life of slurs. In D., Fogal, D. W. Harris, & M. Moss (eds.), *New work on speech acts.* Oxford University Press.

R CoreTeam (2022). R: A language and environment for statistical computing.

Rassin, E., & Heijden, S. V. D. (2005). Appearing credible? Swearing helps!. *Psychology, Crime & Law*, *11*, 177-182.

Roest, S. A., Visser, T. A., & Zeelenberg, R. (2018). Dutch taboo norms. *Behavior Research Methods*, *50*, 630-641.

Rosenblum, M., Schroeder, J., & Gino, F. (2020). Tell it like it is: When politically incorrect language promotes authenticity. *Journal of Personality and Social Psychology*, *119*, 75-103.

Scaltritti, M., Job, R., & Sulpizio, S. (2021). Selective suppression of taboo information in visual word recognition: Evidence for cognitive control on semantics. *Journal of*

Experimental Psychology: Human Perception and Performance, *47*, 934-945.

Senberg, A., Muenchau, A., Muente, T., Beste, C., & Roessner, V. (2021). Swearing and coprophenomena–A multidimensional approach. *Neuroscience & Biobehavioral Reviews*, *126*, 12-22.

Sheidlower, J. (2009). *The F-word*.

Špago, D. (2020). Gender-related differences in the use and perception of verbal insults: the Bosnian perspective. *Lingua Posnaniensis*, *62*, 81-94.

Stapleton, K. (2010). *Swearing*. In M.A. Locher, S.L. Graham (Eds.), Interpersonal Pragmatics, De Gruyter Mouton, Berlin, Germany (pp. 289-306).

Stephens, R., & Umland, C. (2011). Swearing as a response to pain—Effect of daily swearing frequency. *The Journal of Pain*, *12*, 1274-1281.

Sulpizio, S., Toti, M., Del Maschio, N., Costa, A., Fedeli, D., Job, R., & Abutalebi, J. (2019). Are you really cursing? Neural processing of taboo words in native and foreign language. *Brain and language*, *194*, 84-92.

Sulpizio, S., Vassallo, E., Job, R. & Abutalebi, J. (2020). ITABÙ: Preliminary data for an Italian database for taboo words. *Giornale Italiano di Psicologia, 47*, 559-614.

Thelwall, M. (2008). Fk yea I swear: Cursing and gender in MySpace. *Corpora, 3*, 83-107.

Van Lancker, D., & Cummings, J. L. (1999). Expletives: Neurolinguistic and neurobehavioral perspectives on swearing. *Brain research reviews*, *31*, 83-104.

Vingerhoets, A. J., Bylsma, L. M., & De Vlam, C. (2013). Swearing: A biopsychosocial perspective. *Psihologijske teme*, *22*, 287-304.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2014, February). Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 415-425).

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, *45*, 1191-1207.

White, K. K., Abrams, L., Koehler, S. M., & Collins, R. J. (2017). Lions, tigers, and bears, oh sh! t: Semantics versus tabooness in speech production. *Psychonomic bulletin & review*, *24*, 489-495.

Weeks, J. (2011). *The language of sexuality*. Routledge.

Yik, M., Mues, C., Sze, I. N., Kuppens, P., Tuerlinckx, F., De Roover, K., ... & Russell, J. A. (2023). On the relationship between valence and arousal in samples across the globe. *Emotion*, *23*, 332-344.

Acknowledgments

Jon Andoni Duñabeitia was partially supported by grants PID2021126884NB-I00 from the Spanish Government, ISERIE from Ayudas Fundación BBVA a Proyectos de Investigación Científica 2021, and H2019/HUM-5705 from the Comunidad de Madrid. Dušica Filipović Đurđević was supported by Ministry of Education, Science and Technological Development, Republic of Serbia, Grant no. 451-03-68/2020-14/200163 (University of Belgrade, Faculty of Philosophy). Ernesto Guerra was supported by ANID/PIA/Basal Funds for Centers of Excellence Project FB0003. Fritz Günther was supported by DFG Emmy-Noether grant "What's in a name?" (project nr. 459717703). Simone Sulpizio was partially supported by the grant PRIN nr. 2022N87CR9 from Italian Ministry of University and Research (MUR).

Author contribution

S.S.: Conceptualizaion, Methodology, Software, Investigation, Data curation, Supervision, Project administration, Writing – Original Draft, Writing – Review & Editing

Ma.Ma.: Conceptualizaion, Methodology, Supervision, Project administration, Writing – Original Draft, Writing – Review & Editing

F.G.: Conceptualizaion, Methodology, Software, Investigation, Data curation, Formal analyses, Supervision, Project administration, Writing – Original Draft, Writing – Review & Editing

B.B.: Methodology, Software, Investigation, Data curation, Writing – Review & Editing

L.B.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing

M.B.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing

L.A.C.: Methodology, Software, Data curation, Supervision, Writing – Review & Editing

N.N.C.: Software, Investigation, Data curation, Writing – Review & Editing

C.D.: Methodology, Data curation

D.F.Đ.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing

J.A.D.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing

F.F.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing

L.F.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing

E.G.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing

G.H.: Methodology, Software, Writing – Review & Editing

K.J.: Software, Investigation, Data curation, Writing – Review & Editing

R.J.: Conceptualizaion, Writing – Review & Editing

H.K.: Methodology, Software, Investigation, Data curation, Writing – Review & Editing

J.K.: Methodology, Software, Investigation, Data curation, Writing – Review & Editing

N.K.: Methodology, Software, Investigation, Data curation, Writing – Review & Editing

S.K.: Supervision, Writing – Review & Editing, Funding acquisition

C.Y.L.: Software, Investigation, Data curation, Writing – Review & Editing

L.L.: Methodology, Writing – Review & Editing

N.L.: Methodology, Writing – Review & Editing

C.M.: Supervision, Writing – Review & Editing

I.G.M.: Methodology, Data curation

K.M.: Investigation, Writing – Review & Editing

Ma.Mi.: Investigation, Data curation, Writing – Review & Editing

Mi.Ma.: Methodology, Writing – Review & Editing

A.N.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing

M.C.M.: Methodology, Software, Investigation, Data curation, Writing – Review & Editing

M.O.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing

P.R.: Methodology, Software, Writing – Review & Editing

G.S.: Software, Investigation, Data curation, Writing – Review & Editing

C.S. T.: Methodology, Supervision, Writing – Review & Editing, Funding acquisition

C.W.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing

P.W.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing

M.Y.: Methodology, Software, Investigation, Data curation, Supervision, Writing – Review & Editing, Funding acquisition

Supplementary materials

Supplementary method

Study 1

Definition of the categories used to classify Study 1 productions

Insult: an insolent/rude word/expression, that can offend and/or sound as an affront. Note that insults may refer to different types of (perceived or actual) deviations (physical, psychological, social; e.g., wimp, retarded). This category also includes some animal names that can be metaphorically used to refer to characterize people physical or psychological abilities (e.g., *donkey, monkey, pig*).

Slur: a pejorative term that targets people on the basis of their group (nationality, ethnicity, religion, gender, sexual orientation, etc; e.g., faggot, nigger).

Sexual references: any word/expression having as a referent a sex-related body part (e.g., cunt) or a sexual practice (e.g., *blow job*).

Scatological referents and disgusting objects (e.g., *crap, shit*).

Profanities/blasphemies: any irreverent word/expression toward/around God and/or sacred things (e.g., *goddamn*).

Supplementary Table 1. Summary of the Study 1 data collection – language, participants demographics, number of produced items, and ethical approval

Language	Recruitment, testing modality, reimbursement	N (male, female, other)	Age (mean, SD)	Items produced - unique types total (average per participant); produced by >3% of participants	Ethical approval
Cantonese (CN)	University students, online, money	41 (21, 20, 0)	20.1 (1.2)	632 (23.0); 93	Survey and Behavioural Research Ethics, The Chinese University of Hong Kong; Reference number: EDU2020-098
Spanish (CL)	Social media, online, none	73 (5, 25, 0); 43 N.A.	37.3 (5.9); 43 N.A.	216 (12.5); 80	Comité de Ética de la Investigación en Ciencias Sociales y Humanidades, Facultad de Filosofía y Humanidades, Universidad de Chile; Reference number 16/2020

Dutch[2] (BE)	Prolific, online, money	48 (-)	-	704 (29.6); 198	Ethical Committee Faculty of Psychology and Educational Sciences, University of Gent; Reference number 2020/115
English (AU)	University students, online, course credit	45 (14, 31, 0)	22.2 (6.5)	464 (26.9); 144	Macquarie University Human Research Ethics Committee; Reference number #52020795419110
English (CA)	Social media, online, none	167 (69, 92, 6)	39.5 (17.7)	665 (13.3); 56	University of Alberta REB; #Pro00115115
English (GB)	University students, online, course credit	59 (13, 46, 0)	35.1 (12.3)	251 (13.9); 100	University of Surrey Ethics Committee; reference number FHMS 20-21 093 EGA
English (SG)	University students, online, money	40 (20, 20, 0)	25 (2.2)	377 (24.6); 120	Institutional Review Board, National University of Singapore; Reference Code: NUS-IRB-2021-2
English (US)	Social media + Prolific, online, none or money	71 (24, 42, 5)	23.1 (5.5)	396 (24.5); 103	University of Washington Institutional Review Board; status Exempt, study id: STUDY00011426
Finnish (FI)	University students, online, none	49 (22, 27, 0)	22.8 (1.7)	432 (26.2); 139	University of Helsinki Ethical Review Board in Humanities and Social and Behavioral Sciences; Statement 50/2021
French (FR)	University students, pen and paper, course credit	40 (13, 27, 0)	22.6 (4.5)	179 (21.7); 130	Research Ethics Committee of Université Clermont Auvergne; Reference number IRB00011540-2020-51
German (DE)	University students, online, money or course credit	41 (21, 20, 0)	22.02 (2.57)	863 (52.9); 299	Ethics committee for Psychological Research, University of Tübingen; Az Guenther_2020_0726_198
Italian (IT)	University students, online, course credit	62 (13, 48, 1)	21.9 (1.1)	441 (21.1); 189	Committee for minimal risk Research Evaluation of the Department of Psychology, University of Trento; Protocol Number RM-2020-325
Mandarin (CN)	University students, online, none	44 (10, 33, 1)	24.8 (3.4)	301 (9.3); 44	Ruled as “not required”
Serbian (RS)	Social media, online, personal contacts, none	88 (19, 69, 0)	25.8 (5.87)	975 (26.5); 174	Institutional Review Board of the Department of Psychology, University of Belgrade; Protocol #2020-43
Setswana (BW)	University students + personal contacts, pen and paper, money	45 (15, 30, 0)	23.6 (6.2)	275 (15.3); 88	Office of Research and Development of the University of Botswana; reference UBR/RES/IRB/SOC/096

[2] Information about participants gender and age were not collected for Dutch as requested by the local ethical committee to increase anonymity.

Slovenian (SI)	University students + personal contacts, online, none	43 (19, 24, 0)	21.7 (2.5)	352 (18.8); 127	Ethical Committee Faculty of Arts, University of Ljubljana; Reference number: 209-2020
Spanish (ES)	University students + personal contacts, online, none	45 (17, 28, 0)	29.6 (7.4)	413 (15.6); 89	Research Ethics Committee of Universidad Nebrija; Reference UNNE-2020-009
Thai (TH)	University students, pen and paper, none	45 (19, 24, 2)	26.7 (6.8)	254 (13.0); 94	Burapha University Ethics committee (The SIDCER-FERCAP); Protocol number IRB2-016/2564

Note: For each lab, the ethical approval encompassed both Study 1 and Study 2.

Study 2

Instruction for Study 2 data collection

Thank you for participating in this study!

In the following study, you will be presented with sets of six words. There are no right or wrong answers, we are interested in your honest opinion. If you feel that several answers are possible, choose the one that best represents your opinion.

Note that some words can be offensive or insulting.

CONCRETENESS

Your task is to judge which of these words is the most concrete and which is the least concrete. A concrete word refers to something that exists in reality; you can have immediate experience of it through your senses (smelling, tasting, touching, hearing, seeing) and the actions you do. Therefore, “most concrete” refers to the word that you can experience the most using actions and senses, whereas “least concrete” refers to the word that you can experience the least using actions and senses.

AROUSAL

Your task is to judge which of these words is the most arousing and which is the least arousing. The most arousing word is the one evoking the strongest excitation or stimulation in you, no matter if good or bad. The least arousing is the word evoking the weakest excitation or stimulation in you, no matter if good or bad.

VALENCE

Your task is to judge which of these words is the most pleasant and the least pleasant. The most pleasant word is the one making you the happiest or more satisfied or more contented, whereas the least pleasant word is the one making you the most unhappy, unsatisfied, or sad.

AGE OF ACQUISITION

Your task is to judge which of these words you have learned or heard first in your life, and which you learned or heard at the latest point in your life.

TABOONESS

Your task is to judge which of these words is the most taboo and the least taboo one. Taboo means that the word use is not acceptable in most social situations. When completing the task, try to think about the use of the word in different contexts.

OFFENSIVENESS

Your task is to judge which of these words when used, is the most personally offensive and which is the least personally offensive.

Debriefing of Study 2

Thank you very much for your participation. The aim of this study is to assess words that are commonly considered as taboo words, namely socially inappropriate words, across different languages. This cross-linguistic project will examine taboo words used in more than 20 countries across the 5 continents. Despite their frequent use in everyday (oral and written) communication and their social relevance, taboo words have been almost neglected by empirical research on language. This project aims to start characterizing taboo words in many different languages. This will allow us to identify taboo words and empirically investigate how these taboo words are perceived and judged in multiple languages and cultures.

Supplementary Table 2. Summary of the Study 2 ratings – language, participants demographics, split-half reliability coefficients, and gender correlation

Language	Recruitment, testing modality, reimbursement	Number of items (taboo, filler (superscript indicates the source for fillers))	Dimension	N (male, female, other)	Age (mean, SD, range)	Split-half reliability	Gender correlation
Cantonese (CN)	University students, online, money	144 (95, 49) (Hutton & Bolton, 2005; Ng, 2006; 2010)	AoA	16 (8, 8, 0)	20.19 (1.83), 18 - 24	0.746	0.677
			arousal	16 (8, 8, 0)	20.31 (1.40), 18 - 23	0.857	0.744
			concreteness	16 (8, 8, 0)	20.31 (1.40), 18 - 22	0.736	0.676
			offensiveness	16 (8, 8, 0)	20.44 (1.82), 18 - 24	0.883	0.829
			tabooness	16 (8, 8, 0)	20.19 (1.28), 18 - 22	0.905	0.854
			valence	16 (8, 8, 0)	20.44 (1.82), 18 - 24	0.891	0.732
Dutch (BE)	Prolific, online, money	380 (249, 131) (Moors et al., 2013)	AoA	42 (21, 21, 0)	25.43 (5.69), 18 - 40	0.886	0.883
			arousal	42 (21, 21, 0)	25.36 (5.93), 18 - 40	0.598	0.585
			concreteness	42 (21, 21, 0)	26.29 (5.87), 18 - 40	0.815	0.797
			offensiveness	42 (21, 21, 0)	26.14 (5.58), 19 - 39	0.857	0.859
			tabooness	42 (21, 21, 0)	25.52 (6.33), 18 - 40	0.889	0.885

			valence	42 (21, 20, 1)	25.67 (6.17), 18 - 40	0.884	0.877
English (AU)	Prolific, online, money	219 (149, 70) (Bradley & Lang, 1999)	AoA	24 (13, 11, 0)	21.54 (5.56), 18 - 36	0.847	0.873
			arousal	25 (13, 12, 0)	22.44 (5.73), 18 - 39	0.758	0.780
			concreteness	25 (12, 13, 0)	20.60 (4.11), 18 - 34	0.342	0.388
			offensiveness	24 (11, 13, 0)	19.54 (3.35), 18 - 30	0.877	0.847
			tabooness	24 (11, 13, 0)	19.67 (4.04), 18 - 38	0.904	0.905
			valence	24 (12, 12, 0)	21.17 (5.44), 18 - 38	0.861	0.853
English (CA)	University students, in-lab, course credit	378 (252, 126) (Hollis et al., 2017)	AoA	42 (19, 20, 3)	20.17 (3.62), 18 - 32	0.905	0.892
			arousal	42 (16, 22, 4)	20.45 (4.04), 18 - 37	0.862	0.819
			concreteness	42 (22, 19, 1)	19.55 (1.84), 18 - 25	0.640	0.608
			offensiveness	42 (21, 21, 0)	18.86 (3.07), 17 - 36	0.891	0.899
			tabooness	41 (17, 22, 2)	19.46 (2.46), 17 - 29	0.918	0.897
			valence	42 (20, 20, 2)	20.14 (2.65), 18 - 30	0.868	0.855
English (GB)	Clickworker, online, money	150 (100, 50) Bradley & Lang, 1999)	AoA	18 (10, 8, 0)	28.67 (6.97), 18 - 40	0.649	0.668
			arousal	17 (8, 9, 0)	31.71 (5.47), 20 - 38	0.371	0.305
			concreteness	16 (7, 9, 0)	33.44 (5.03), 25 - 40	0.548	0.412
			offensiveness	17 (9, 8, 0)	32.65 (5.65), 22 - 40	0.833	0.842
			tabooness	16 (8, 8, 0)	30.94 (6.27), 17 - 39	0.833	0.780
			valence	18 (8, 10, 0)	31.06 (5.77), 19 - 39	0.810	0.826
English (SG)	University students, online, money	179 (116, 63) (Warriner et al., 2017)	AoA	20 (10, 10, 0)	22.40 (1.93), 20 - 28	0.787	0.806
			arousal	20 (10, 10, 0)	22.30 (1.59), 19 - 26	0.786	0.771
			concreteness	20 (10, 10, 0)	21.95 (1.05), 20 - 24	0.441	0.507
			offensiveness	20 (10, 10, 0)	22.40 (1.90), 20 - 26	0.880	0.879
			tabooness	20 (10, 10, 0)	22.40 (1.67), 19 - 26	0.887	0.848
			valence	20 (10, 10, 0)	22.55 (2.21), 20 - 28	0.813	0.798
English (US)	Prolific, online, money	156 (102, 54) Bradley & Lang, 1999)	AoA	18 (8, 10, 0)	25.06 (6.76), 18 - 40	0.893	0.893
			arousal	18 (7, 10, 1)	27.39 (7.08), 18 - 40	0.760	0.788
			concreteness	18 (8, 9, 1)	26.33 (7.58), 18 - 40	0.577	0.621
			offensiveness	18 (8, 9, 1)	26.06 (7.67), 18 - 40	0.889	0.878
			tabooness	18 (8, 9, 1)	25.22 (6.42), 18 - 40	0.925	0.924
			valence	18 (8, 9, 1)	26.61 (6.43), 18 - 40	0.884	0.838
Finnish (FI)	University students, online, none	222 (148, 74) (Eilola & Havelka, 2010)	AoA	13 (0, 0, 13)	28.69 (7.08), 20 - 40	0.876	0.860
			arousal	13 (0, 0, 13)	27.92 (6.32), 20 - 38	0.717	0.636
			concreteness	13 (1, 0, 12)	26.62 (4.50), 23 - 38	0.791	0.775
			offensiveness	13 (0, 0, 13)	26.38 (7.03), 20 - 40	0.911	0.897
			tabooness	13 (1, 0, 12)	28.23 (6.27), 19 - 40	0.939	0.895
			valence	14 (1, 0, 13)	26.50 (6.45), 19 - 40	0.816	0.708
French (FR)	University students, online, course credit	204 (128, 76) (Monnier & Syssau, 2014)	AoA	23 (12, 11, 0)	19.48 (1.65), 18 - 23	0.909	0.883
			arousal	23 (12, 11, 0)	18.83 (1.27), 18 - 23	0.522	0.434
			concreteness	23 (12, 11, 0)	19.48 (1.97), 17 - 26	0.786	0.772
			offensiveness	23 (11, 12, 0)	19.52 (3.26), 17 - 30	0.913	0.900
			tabooness	23 (11, 12, 0)	21.96 (3.89), 18 - 30	0.913	0.854
			valence	23 (11, 12, 0)	22.48 (4.63), 18 - 33	0.868	0.815
German (DE)	University students + Prolific, online, course credit or money	378 (254, 124) (Kanske & Kotz, 2010)	AoA	42 (20, 22, 0)	24.31 (5.82), 18 - 37	0.885	0.865
			arousal	42 (21, 21, 0)	22.14 (3.49), 18 - 36	0.840	0.806
			concreteness	42 (21, 21, 0)	22.05 (3.60), 18 - 32	0.803	0.743
			offensiveness	42 (23, 19, 0)	23.31 (4.36), 18 - 36	0.909	0.895
			tabooness	41 (20, 21, 0)	23.34 (5.48), 18 - 40	0.905	0.915
			valence	42 (22, 20, 0)	22.95 (4.14), 18 - 40	0.898	0.882
Italian (IT)	Prolific, online, money	306 (212, 94) (Montefinese et al., 2014)	AoA	32 (16, 16, 0)	24.03 (4.12), 18 - 34	0.892	0.897
			arousal	32 (16, 16, 0)	25.56 (4.60), 19 - 38	0.852	0.834
			concreteness	32 (16, 16, 0)	24.62 (4.49), 18 - 37	0.808	0.812
			offensiveness	32 (16, 16, 0)	23.69 (4.72), 19 - 37	0.922	0.905

			tabooness	32 (16, 16, 0)	23.41 (3.28), 18 - 31	0.901	0.892
			valence	32 (16, 16, 0)	25.34 (4.63), 18 - 39	0.888	0.797
Mandarin (CN)	University students, online, none	66 (43, 23) (Lin & Yao, 2016)	AoA	8 (4, 4, 0)	29.38 (2.88), 25 - 33	0.772	0.806
			arousal	8 (3, 5, 0)	29.38 (2.72), 25 - 33	0.919	0.879
			concreteness	8 (4, 4, 0)	30.38 (3.78), 25 - 35	0.677	0.602
			offensiveness	8 (3, 5, 0)	29.50 (3.59), 25 - 34	0.897	0.894
			tabooness	8 (3, 5, 0)	30.00 (3.38), 25 - 35	0.886	0.847
			valence	8 (4, 4, 0)	31.12 (3.60), 25 - 35	0.838	0.897
Serbian (RS)	University students + social media, online, none	263 (174, 89) (Filipović Đurđević & Kostić, 2017; Popović Stijačić & Filipović Đurđević)	AoA	30 (15, 15, 0)	29.60 (7.35), 19 - 40	0.849	0.865
			arousal	30 (16, 14, 0)	28.50 (5.81), 18 - 40	0.853	0.841
			concreteness	30 (15, 15, 0)	28.87 (5.02), 20 - 40	0.760	0.695
			offensiveness	30 (15, 15, 0)	29.90 (6.47), 18 - 40	0.914	0.889
			tabooness	30 (15, 15, 0)	29.60 (6.75), 18 - 40	0.895	0.914
			valence	30 (15, 15, 0)	27.63 (6.28), 19 - 40	0.850	0.807
Setswana (BW)	University students + personal contact, online + in-lab, money	132 (87, 45)	AoA	13 (6, 7, 0)	21.23 (2.42), 19 - 28	0.753	0.625
		Fillers translated from Bradley and Lang (1999)	arousal	15 (7, 8, 0)	20.67 (1.45), 19 - 25	0.312	0.378
			concreteness	15 (7, 8, 0)	20.73 (1.62), 19 - 24	0.065	-0.054
			offensiveness	15 (7, 8, 0)	22.93 (4.82), 19 - 37	0.671	0.658
			tabooness	15 (7, 8, 0)	21.67 (2.47), 19 - 28	0.664	0.588
			valence	15 (8, 7, 0)	22.33 (3.29), 19 - 30	0.014	0.190
Slovenian (SI)	University students + personal contact + social media + Prolific, online, course credit or money	378 (248, 130)	AoA	42 (21, 21, 0)	22.07 (2.99), 18 - 30	0.912	0.893
		Fillers translated from Bradley and Lang (1999)	arousal	42 (20, 22, 0)	22.52 (3.56), 18 - 30	0.696	0.716
			concreteness	42 (20, 22, 0)	22.07 (3.37), 19 - 30	0.676	0.522
			offensiveness	42 (20, 21, 1)	21.50 (4.63), 0 - 30	0.902	0.875
			tabooness	42 (22, 20, 0)	22.05 (3.28), 19 - 30	0.889	0.868
			valence	42 (21, 21, 0)	22.60 (3.58), 19 - 38	0.844	0.803
Spanish (CL)	Social media, online, none	180 (120, 60[15])	AoA	20 (12, 8, 0)	23.35 (4.18), 18 - 32	0.867	0.851
			arousal	20 (9, 11, 0)	28.95 (6.13), 21 - 40	0.737	0.777
			concreteness	20 (8, 12, 0)	31.45 (6.23), 23 - 40	0.725	0.629
			offensiveness	20 (9, 10, 1)	27.50 (5.03), 20 - 39	0.864	0.849
			tabooness	20 (10, 10, 0)	30.40 (6.04), 21 - 40	0.886	0.883
			valence	20 (11, 9, 0)	30.45 (6.27), 21 - 39	0.837	0.820
Spanish (ES)	University students + social media, online, none	126 (82, 44) (Redondo et al., 2007)	AoA	14 (7, 7, 0)	22.86 (4.22), 19 - 32	0.909	0.915
			arousal	14 (7, 7, 0)	23.07 (4.10), 19 - 32	0.756	0.769
			concreteness	14 (7, 7, 0)	23.36 (4.29), 19 - 32	0.743	0.719
			offensiveness	14 (7, 7, 0)	24.29 (3.91), 19 - 32	0.890	0.874
			tabooness	14 (7, 7, 0)	22.79 (4.68), 19 - 32	0.849	0.883
			valence	14 (7, 7, 0)	23.57 (1.22), 22 - 26	0.830	0.784
Thai (TH)	University students, online, none	375 (254, 121) (Ngamprom et al., 2017)	AoA	38 (16, 20, 2)	24.39 (6.00), 18 - 40	0.818	0.771
			arousal	37 (16, 19, 2)	26.46 (6.73), 18 - 40	0.673	0.639
			concreteness	36 (16, 17, 3)	26.33 (6.20), 18 - 39	0.504	0.456
			offensiveness	38 (16, 20, 2)	26.63 (6.42), 18 - 39	0.866	0.855
			tabooness	38 (19, 17, 2)	26.16 (6.00), 17 - 39	0.855	0.874
			valence	36 (17, 17, 2)	25.47 (5.96), 16 - 39	0.809	0.773

Supplementary analyses – Study 2

Calculating external reliabilities – correlations with other data sets

To calculate external reliabilities, we collected all published word norms sharing the rating dimensions with our study. An overview of these word norms, alongside the number of shared items with our datasets and the shared rating dimensions, is provided in Supplementary Table 3. Note that we also considered some rating dimensions that only partially overlap with our dimensions, but share a very similar construct (*emotional charge* instead of *arousal* in Eilola and Havelka (2011); *imageability* instead of *concreteness* in the taboo word norms by Janschewitz (2008); *insult* instead of *offensiveness*, and *personal taboo* and *public taboo* instead of *tabooness* in the taboo norms by Roest and colleagues (2018)).

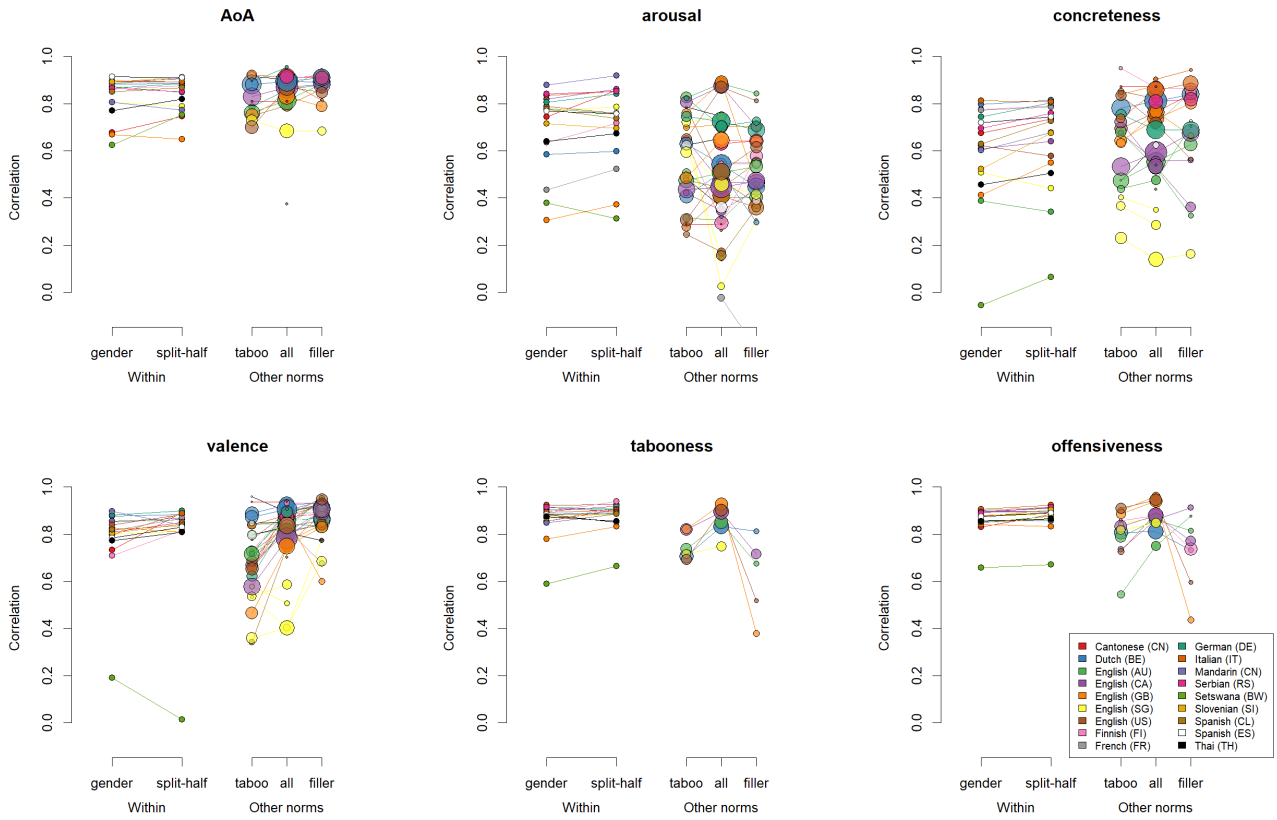
Supplementary Table 3. Overview of the word norms used in the current study for each sample to calculate the external reliability, together with the dimensions and the number of items shared with each resource used. Dashes (-) indicate cases where the word norms use slightly different terms than the ones established in our study, or additional specifications of these terms.

Language	**Dataset**	**Shared dimensions**	**Shared items (taboo, filler)**
Cantonese (CN)	Cai et al. (2022)	AoA	6 (6, 0)
	Su et al. (2022)	AoA; concreteness	11 (11, 0)
	Xu and Li (2020)	concreteness	9 (9, 0)
	Xu et al. (2022)	arousal; valence	10 (10, 0)
	Yao et al. (2017)	arousal; concreteness; valence	2 (2, 0)
	Yee (2017)	arousal; concreteness; valence	0 (0, 0)
Dutch (BE)	Brysbaert et al. (2014)	concreteness	299 (173, 126)
	Brysbaert et al. (2014)	AoA	302 (176, 126)
	Roest et al. (2018)	arousal; offensiveness - insulting; tabooness - general; tabooness - personal; valence	89 (70, 19)
	Moors et al. (2013)	AoA; arousal; valence	192 (67, 125)
English (AU)	Bradley and Lang (1999)	arousal; valence	94 (30, 64)
	Brysbaert et al. (2014)	concreteness	164 (97, 67)
	Eilola and Havelka (2010)	arousal - emotional charge; concreteness; offensiveness; valence	37 (26, 11)
	Janschewitz (2008)	arousal; concreteness - imageability; offensiveness; tabooness; valence	67 (48, 19)
	Kuperman et al. (2012)	AoA	162 (98, 64)
	Warriner et al. (2013)	arousal; valence	154 (89, 65)
English (CA)	Bradley and Lang (1999)	arousal; valence	151 (28, 123)
	Brysbaert et al.(2014)	concreteness	265 (141, 124)
	Eilola and Havelka (2010)	arousal - emotional charge; concreteness; offensiveness; valence	40 (20, 20)

	Janschewitz (2008)	arousal; concreteness - imageability; offensiveness; tabooness; valence	95 (56, 39)
	Kuperman et al. (2012)	AoA	266 (143, 123)
	Warriner et al.(2013)	arousal; valence	247 (123, 124)
English (GB)	Bradley and Lang (1999)	arousal; valence	59 (9, 50)
	Brysbaert et al. (2014)	concreteness	108 (58, 50)
	Eilola and Havelka (2010)	arousal - emotional charge; concreteness; offensiveness; valence	31 (23, 8)
	Janschewitz (2008)	arousal; concreteness - imageability; offensiveness; tabooness; valence	59 (37, 22)
	Kuperman et al. (2012)	AoA	109 (60, 49)
	Warriner et al. (2013)	arousal; valence	105 (55, 50)
English (SG)	Bradley and Lang (1999)	arousal; valence	25 (13, 12)
	Brysbaert et al. (2014)	concreteness	84 (50, 34)
	Eilola and Havelka (2010)	arousal - emotional charge; concreteness; offensiveness; valence	19 (18, 1)
	Janschewitz (2008)	arousal; concreteness - imageability; offensiveness; tabooness; valence	37 (35, 2)
	Kuperman et al. (2012)	AoA	80 (49, 31)
	Warriner et al.(2013)	arousal; valence	88 (48, 40)
English (US)	Bradley and Lang (1999)	arousal; valence	63 (11, 52)
	Brysbaert et al. (2014)	concreteness	115 (62, 53)
	Eilola and Havelka (2010)	arousal - emotional charge; concreteness; offensiveness; valence	31 (22, 9)
	Janschewitz (2008)	arousal; concreteness - imageability; offensiveness; tabooness; valence	58 (42, 16)
	Kuperman et al. (2012)	AoA	115 (65, 50)
	Warriner et al. (2013)	arousal; valence	114 (62, 52)
Finnish (FI)	Eilola and Havelka (2010)	arousal - emotional charge; concreteness; offensiveness; valence	73 (14, 59)
	Söderholm et al. (2013)	arousal; valence	20 (0, 20)
French (FR)	Bonin et al. (2018)	arousal; concreteness; valence	25 (7, 18)
	Ferrand et al. (2008)	AoA	21 (7, 14)
	Monnier and Syssau (2014)	arousal; valence	45 (1, 44)
German (DE)	Birchenough et al. (2017)	AoA	84 (22, 62)
	Kanske & Kotz (2010)	arousal; concreteness; valence	159 (43, 116)
	Schmidtke et al. (2014)	arousal; valence	51 (19, 32)
	Schröder et al. (2012)	AoA	15 (10, 5)
Italian (IT)	Della Rosa et al. (2010)	AoA; concreteness	17 (4, 13)
	Montefinese et al. (2014)	arousal; concreteness; valence	127 (35, 92)
	Montefinese et al. (2019)	AoA	131 (39, 92)
Mandarin (CN)	Cai et al. (2022)	AoA	12 (8, 4)
	Su et al. (2022)	AoA; concreteness	9 (7, 2)
	Xu and Li (2020)	concreteness	11 (4, 7)
	Xu et al. (2022)	arousal; valence	13 (5, 8)
	Yao et al. (2017)	arousal; concreteness; valence	5 (3, 2)
Spanish (CL)	Alonso et al. (2015)	AoA	52 (19, 33)
	Guasch et al. (2016)	arousal; concreteness; valence	12 (4, 8)
	Hinojosa et al. (2016)	arousal; concreteness; valence	1 (1, 0)
	Redondo et al. (2007)	arousal; valence	40 (7, 33)

	Stadthagen-Gonzalez et al. (2017)	arousal; valence	53 (20, 33)
Spanish (ES)	Alonso et al. (2015)	AoA	66 (35, 31)
	Guasch et al. (2016)	arousal; concreteness; valence	23 (10, 13)
	Hinojosa et al. (2016)	arousal; concreteness; valence	6 (5, 1)
	Redondo et al. (2007)	arousal; valence	51 (20, 31)
	Stadthagen-Gonzalez et al. (2017)	arousal; valence	69 (38, 31)
Thai (TH)	Ngamprom et al. (2017)	valence	19 (2, 17)
	Ngamprom et al. (2017)	arousal	5 (1, 4)

To compare these external reliabilities with the internal split-half reliabilities, we estimated a linear mixed effects model (LMM; Bates et al., 2015; Kuznetsova et al., 2017) predicting the correlation from a fixed effect of the type of reliability (internal vs. external), plus random intercepts and random slopes by type for the rating dimension and sample. We only considered data points where at least 10 items were shared with the word norms from other studies. Here, we observed that external reliabilities were not significantly lower (b = -0.070, t = -1.846, p = .107) than internal split-half reliabilities (for which we observe an intercept of 0.792). We thus have no evidence to indicate that the participants performing our rating tasks diverged substantially from the participant samples from other studies. This is relevant considering that participants in our study encountered a different context with far more taboo words when providing their judgments and performed a different type of rating task (best-worst scaling instead of standard Likert scales).

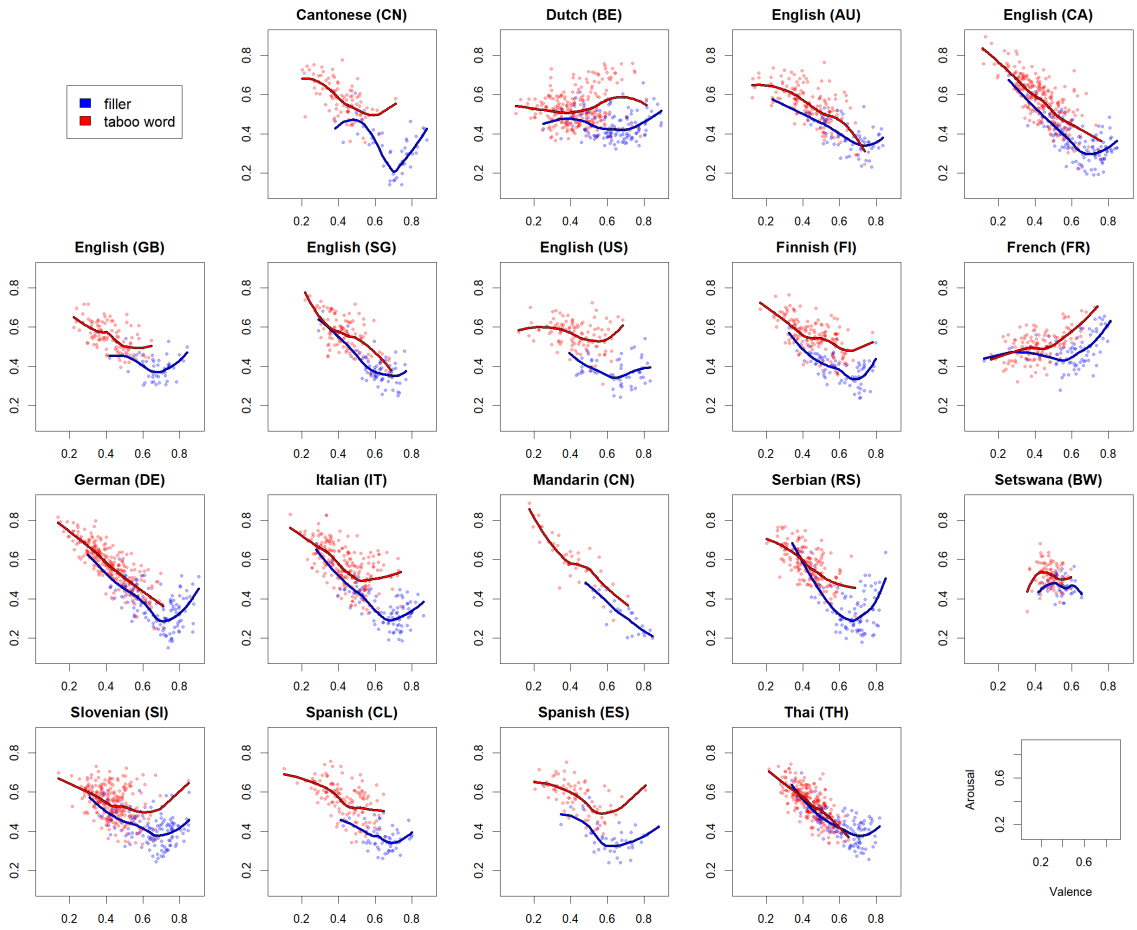


Supplementary Figure 1: **Internal and external reliabilities, and gender agreement.** For each sample and rating dimension, the correlation between male and female rating scores (left-most point), split-half reliability (second point from the left), and the correlation to rating scores from other word norms (right part), split by taboo words, all words, or filler words. The size of the points on the right part indicates the number of shared items (larger points = more items). Only points estimated from more than n = 10 shared items are displayed.

The relation between valence and arousal.

As can be seen in the top panel of Figure 5, there is indeed a U-shaped relation between valence and arousal when pooling together data from all samples. This is confirmed in a statistical analysis: Adding a quadratic fixed effect for valence to a LMM predicting arousal from a linear fixed effect for valence and a random effect for languages improves these baseline models for taboo words ($X^2(1) = 40.77$, $p < .001$), non-taboo words ($X^2(1) = 130.74$, $p < .001$), and both types of words combined ($X^2(1) = 30.25$, $p < .001$). As indicated in Supplementary Figure 2, the same pattern emerges in most individual datasets too (with some noteworthy exceptions such as Mandarin, where we observe clear negative relations between valence and arousal). Note that, overall, only few taboo words tended to have very high valence ratings (in some datasets such as Cantonese, Spanish (CL), Mandarin, Serbian,

English (GB), or German, such items are essentially missing entirely), so the right end of the distribution that could display a positive relation between valence and arousal is missing here.



Supplementary Figure 2: Relation between valence and arousal for taboo and filler words. The relation between valence (x-axis) and arousal (y-axis) by sample; points indicate individual items. Note that regression lines predicting arousal from valence are fitted with local polynomial regression (loess) fitting (the loess() function in R), which display more general non-linear effects than the quadratic terms analysed here.

Predicting tabooness and offensiveness – analysis of the taboo word subset

Since the variables predicting tabooness/offensiveness for all items could just be the ones telling apart taboo words from non-taboo words, we repeated the mixed-effect models analysis described in the main text but restricted it to taboo words only. The results of this analysis are displayed in Supplementary Table 3. As can be seen, the general pattern of results is very similar to the analysis of all words. Therefore, the very same variables predict

differences in tabooness and offensiveness ratings (and the difference between them) within the set of taboo words generated in Study 1 of this study.

Supplementary Table 3. Predictors of tabooness and offensiveness ratings across samples, for the dataset of taboo words (produced in Study 1) only. "Dummy" is a dummy variable encoding for tabooness ratings (coded as 0, the reference condition) or offensiveness ratings (coded as 1); therefore, the intercept and main effects except "dummy" describe tabooness ratings, while the "dummy" effect and all interactions describe how offensiveness ratings differ from tabooness ratings.

predictor type	predictor	*b*	*t*	*p*
intercept	intercept	0.348	25.34	< .001
main effects	dummy	0.414	21.65	< .001
	valence	-0.388	-29.83	< .001
	arousal	0.415	27.00	< .001
	concreteness	0.138	10.83	< .001
	AoA	0.219	21.27	< .001
	corpus freq.	-0.006	-5.18	< .001
interactions	dummy : valence	-0.271	-14.74	< .001
	dummy : arousal	-0.100	-4.62	< .001
	dummy : concreteness	-0.211	-11.80	< .001
	dummy : AoA	-0.252	-17.44	< .001
	dummy : corpus freq.	-0.004	-2.46	.014

References

Alonso, M. A., FernAndez, A., & Díez, E. (2015). Subjective age-of-acquisition norms for 7,039 Spanish words. *Behavior Research Methods*, *47*, 268-274.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statiscal Software, 67*, 1-4.

Birchenough, J. M., Davies, R., & Connelly, V. (2017). Rated age-of-acquisition norms for over 3,200 German words. *Behavior Research Methods*, *49*, 484-501.

Bonin, P., Méot, A., & Bugaiska, A. (2018). Concreteness norms for 1,659 French words: Relationships with other psycholinguistic variables and word recognition times. *Behavior Research Methods*, *50*, 2366-2387.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Vol. 30, No. 1, pp. 25-36). Technical report C-1, the center for research in psychophysiology, University of Florida.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*, 80-84.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904-911.

Cai, Z. G., Huang, S., Xu, Z., & Zhao, N. (2022). Objective ages of acquisition for 3300+ simplified Chinese characters. *Behavior Research Methods*, *54*, 311-323.

Della Rosa, P. A., Catricalà, E., Vigliocco, G., & Cappa, S. F. (2010). Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods*, *42*, 1042-1048.

Eilola, T. M., & Havelka, J. (2010). Affective norms for 210 British English and Finnish nouns. *Behavior Research Methods*, *42*, 134-140.

Eilola, T. M., & Havelka, J. (2011). Behavioural and physiological responses to the emotional and taboo Stroop tasks in native and non-native speakers of English. *International Journal of Bilingualism*, *15*, 353-369.

Ferrand, L., Bonin, P., Méot, A., Augustinova, M., New, B., Pallier, C., & Brysbaert, M. (2008). Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic French words and their relation with other psycholinguistic variables. *Behavior Research Methods*, *40*, 1049-1054.

Filipović Đurđević, D., & Kostić, A. (2017). Number, relative frequency, entropy, redundancy, familiarity, and concreteness of word senses: Ratings for 150 Serbian polysemous nouns. In S. Halupka-Rešetar and S. Martínez-Ferreiro (Eds.) *Studies in Language and Mind 2* (pp. 13-77).

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

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

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology*, *70*(8), 1603-1619.

Hutton, C. & Bolton, K. *Dictionary of Cantonese slang: The language of Hong Kong movies, street gangs and city life.* (University of Hawaii Press, 2005).

Janschewitz, K. (2008). Taboo, emotionally valenced, and emotionally neutral word norms. *Behavior Research Methods*, *40*, 1065-1074.

Kanske, P., & Kotz, S. A. (2010). Leipzig affective norms for German: A reliability study. *Behavior Research Methods*, *42*, 987-991.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978-990.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1-26.

Lin, J., & Yao, Y. (2016). Encoding emotion in Chinese: a database of Chinese emotion words with information of emotion type, intensity, and valence. *Lingua Sinica*, *2*, 1-22.

Monnier, C., & Syssau, A. (2014). Affective norms for French words (FAN). *Behavior Research Methods*, *46*(4), 1128-1137.

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). The adaptation of the affective norms for English words (ANEW) for Italian. *Behavior Research Methods*, *46*, 887-903.

Montefinese, M., Vinson, D., Vigliocco, G., & Ambrosini, E. (2019). Italian age of acquisition norms for a large set of words (ItAoA). *Frontiers in Psychology*, *10*, 278.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A. L., ... & Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of

acquisition for 4,300 Dutch words. *Behavior research methods*, *45*, 169-177.

Ng, H. (2006). *A study of Hong Kong Cantonese I.* (Subculture).

Ng, H. (2010). *A Study of Hong Kong Cantonese II*. (Subculture).

Ngamprom, C., Chadcham, S. & Wongupparaj, P. (2017). Development of the Affective Norms for Thai Words (THAI-ANW) Bank System. *Research Methodology in Cognitive Science, 15*, 162-178.

Popović Stijačić, M. & Filipović Đurđević, D. Perceptual strength, concreteness, imageability, context availability, age of acquisition, familiarity, emotional valence, and arousal ratings for 2100 Serbian nouns and their effect on visual lexical decision latencies.

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

Roest, S. A., Visser, T. A., & Zeelenberg, R. (2018). Dutch taboo norms. *Behavior Research Methods*, *50*, 630-641.

Schmidtke, D. S., Schröder, T., Jacobs, A. M., & Conrad, M. (2014). ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior Research Methods*, *46*, 1108-1118.

Schröder, A., Gemballa, T., Ruppin, S., & Wartenburger, I. (2012). German norms for semantic typicality, age of acquisition, and concept familiarity. *Behavior Research Methods*, *44*, 380-394.

Söderholm, C., Häyry, E., Laine, M., & Karrasch, M. (2013). Valence and arousal ratings for 420 Finnish nouns by age and gender. *PloS One*, *8*, e72859.

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

Su, I. F., Yum, Y. N., & Lau, D. K. Y. (2022). Hong Kong Chinese character psycholinguistic norms: Ratings of 4376 single Chinese characters on semantic radical transparency, age-of-acquisition, familiarity, imageability, and concreteness. *Behavior Research Methods*, 1-20.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191-1207.

Xu, X., & Li, J. (2020). Concreteness/abstractness ratings for two-character Chinese words in MELD-SCH. *PloS One*, *15*, e0232133.

Xu, X., Li, J., & Chen, H. (2022). Valence and arousal ratings for 11,310 simplified Chinese

words. *Behavior research methods*, *54*, 26-41.

Yao, Z., Wu, J., Zhang, Y., & Wang, Z. (2017). Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*, *49*, 1374-1385.

Yee, L. T. (2017). Valence, arousal, familiarity, concreteness, and imageability ratings for 292 two-character Chinese nouns in Cantonese speakers in Hong Kong. *PloS One*, *12*, e0174569.