**Data exclusion participant-level**

*One participant was removed from the participant-based (Fl) ANOVA due to insufficient data; one due to a high rate of incorrect responses (42%)*

Insufficient data is not operationalized, so not included. High rate of incorrect responses is again no clear cutoff. It mentions 42% but that is presumably not the cutoff. Hence, this segment did not yield any concrete processing step.

*One participant whose logtransformed response times did not correlate with the average log-transformed response times of all other participants (r = -.05) was removed from the analysis.*

Similar as the previous one, it mentions -.05, but that is presumably not the cutoff. Hence, this segment did not yield any concrete processing step.

*2 did not complete all the stages (1 health deterioration, 1 died);*

Not applicable to the current study (hopefully).

*Some data were not usable due to technical problems with the Ibex server and with the interface between Ibex and Amazon Mechanical Turk; exclusion of the non-natives (English); exclusion of those who completed the study twice*

The first one is rather ad hoc and doesn’t apply to our study (presumably). The second option (excluding non-natives) is included. As to the last option, it isn’t possible to check in the current study whether someone participated more than once (presumably this did not occur often).

*power and equipment failures (Exp.1,4) or experimentor error (Exp.2); Exp.2: participant who didnt show up for the 2nd session*

Participants not showing up for the 2nd session is not applicable. The other options are also not applicable, I believe. If there would be an error somewhere, I would argue it shouldn’t be part of the multiverse. For example, if a certain stimulus is not displayed correctly, it doesn’t make sense to include the corresponding data in the analysis. As such, it shouldn’t be an option in the multiverse (also not as a forced choice with arbitrary timing; see below); there is a clearly a superior choice and no plausible alternative.

**Data exclusion item-level**

*For the word item analysis, four items were removed due to insufficient data in the first test block*

Insufficient data is not operationalized, so not included.

*removed because item had more than 50% incorrect response rate*

In contrast to some of the earlier cases (e.g., 42%), the 50% is plausibly the cutoff they used, so this option is included.

*After the experiment was done a prime awareness interview -> no awareness (all good)*

Not applicable to the current study.

*Any words missing from either our corpora or Elexicon were excluded from the analysis in question (157 items, 2.4% of the total list of prime–target pairs).*

Not applicable to the current study.

*Items whose lemma or surface frequencies were missing from CELEX were excluded from the analysis that included transition probability of the suffix. In an exploratory alternative analysis, they used MorphoLex frequencies*

Not applicable to the current study.

*exclusion of trials with a non-real-word target*

This is by default the case in the current study (eventually).

*Fillers*

Removing fillers is not applicable to the current study. Note though that it could impact certain outlier exclusion criteria if fillers/nonwords are included when e.g., calculating 3 SD above/below the mean.

There is a difference between absolute and relative processing steps. Absolute steps (e.g., removing participants with a different native language) are not influenced by the order in which they are carried out. Relative steps (e.g., removing outliers 3SD above the mean) involve some kind of reference or baseline, and are as such influenced by the order in which they are carried out. Certain decisions such as removing non-words in this study don’t have plausible alternatives (i.e., it doesn’t make sense to include non-words in the eventual analysis). However, the timing of them affects the relative processing steps. I’m currently thinking of making a separate category for these “forced choices” of which only the timing is an arbitrary decision.

**Data exclusion trial-level**

*Analysis only on correct "word" responses*

Coded this as exclude incorrect trials.

*correct trials, (RTs) longer than 3000 ms and shorter than 200 ms (0.7% of the data) were excluded from the analyses*

As trials automatically timed out after 3000 ms there is no way to include those trials as an alternative processing step. As such only 200 ms is included as a category.

*equipment malfunction; trials on which subjects failed to follow instructions and pressed inappropriate buttons or* *pressed the button before the stimulus was displayed*

Equipment malfunction is treated as experimenter error (see above). Failing to follow instructions and pressing inappropriate buttons will result in an incorrect response, so no need for a separate category. Regarding the last option “pressing the button before the stimulus was displayed”: this was not recorded in the current study, so not part of the multiverse.

*Erroneously completed targets; targets preceded by an incorrectly completed prime; responses faster than 250 ms and slower than 4000 ms*

Same as before for the 4000 ms criteria. Only 250 criterion is included.

**Outlier treatment participant-level**

*unclear if this should be included before or here, but because their mean was 2SD above the condition mean*

Condition here refers to a between-subjects variable (version of the experiment). We don’t really have that, so it was coded as 2SD above participants’ mean in general.

*The authors described this decision in terms of them being outliers: Data from two participants, with overall error rates of 14% and 19% respectively. These rates of incorrect responses were substantially higher than the average error rate for all participants, which was 5%.*

Again, not a specific cutoff. It is reasonable to assume the cutoff was 10%, which is already included as a criterion anyway (under data exclusion participant-level).

*(Exp1) 1 participant performing the task at chance level; 2 participants with mean reaction times slower than 2 standard deviations from the sample mean*

Performing at chance level isn’t operationalized. The other criterion is already included.

*1 control participant made a large number of errors*

Not operationalized, so not included

*Removed any participant whose mean RT in any one condition was greater than 2.5 SD from the mean for that condition across participants.*

Again, we don’t have any conditions except related vs unrelated (condition here referred to different kind of primes). Hence, it was coded as 2.5 SD above participants’ mean in general.

*Subjects who failed to reach a preselected accuracy criterion (data-driven. Exp.1: 64%; Exp.2: 67%; Exp.3-4: 63%)*

One-sided proportion test (for exp 1: prop.test(x= 27, n=42, p=.5, alternative = "greater")). Authors did this separately for words and non-words though.

*experiment 2: Two subjects were removed from analysis since they had error rates of higher than 25% in non-word trials.*

Coded as “for nonwords: Participants with an error rate above 25% removed”, though it is possible that they implicitly used the same criterion for word trials, but since none exceeded this threshold it wasn’t mentioned.

**Outlier treatment item-level**

*Free association material test: For one of our pairs (navy-gun), the target was given by 6 of 45 subjects. --> excluded*

Not applicable to the current study

*Emu-ostrich and starling-crow pairs: only few people knew the meaning of the words*

Not operationalized. In addition, there is a criterion “Items with an error rate above 50% removed”, which also captures this notion.

*targets with >25% errors; Additionally, newspapers LSA space analysis: if at least one word in the pair could not be found in the semantic space; Exp. 2: additionally, Cook’s Distance of more than double the value of the second largest*

First criterion is included, the second is not applicable to the current study. The third one could be considered part of the analysis choices as it would involve fitting a model and considering the residuals. Hence, it is not included here.

*8 paires were eliminated from all analyses because, due to chance repairing, the “unrelated” pairs (e.g., flesh–knife) had higher LSA scores than the corresponding related pairs (e.g., dagger–knife, with LSA values of .49 and .27 for the unrelated and related pairs, respectively).*

Not applicable to the current study (similarity was checked beforehand, not based on LSA though).

*"plow fort" because it lacked a "modal response" to finding the mediator in a different exp*

Not applicable to the current study.

**Outlier treatment trial-level**

*For correct RTs, a mean and standard deviation were calculated for each subject within each SOA and session, and any RT greater than 3 SDs above or below the mean for that subject during that SOA and session was identified as an outlier. I also know that 2.5 is presented in the data*

This is coded as “Per participant: RTs +- 3 SD from the participant-specific mean are removed ” and “Per participant: RTs +- 2.5 SD from the participant-specific mean are removed ”.. We don’t have separate sessions or SOA conditions. Exclusion of incorrect responses is considered a separate step.

*We used Van Selst and Jolicoeur’s (1994) non-recursive procedure yielding a separate cutoff criterion for every participant.*

Not included because it is from an article that doesn’t fit the scope of the literature review (it got included because it was part of a dissertation that featured some articles that did fit the scope).

*Exclusion: For control participants, any responses under 200ms or no responses within 2000ms; aphasia group: 200ms to 10000ms; Replaced: lexical decision latencies which deviated more than two standard deviations from an individual participant’s mean were replaced with the mean +/- 2SD, also did the replacement by item (replacing the trials that were too long)*

I went back to the paper, because I wasn’t sure I understood the replacement by item aspect. The author says “In the by-participant analyses, lexical decision latencies which deviated more than two standard deviations from an individual participant’s mean were replaced with the mean plus or minus two standard deviations as appropriate.” and “Each item’s mean and standard deviation were computed and where an item’s lexical decision latencies deviated more than two standard deviations from their mean, the value was replaced with the mean plus or minus two standard deviations as appropriate.”

*RT > 3 SDs above the condition mean or grand mean (Exp.2-3) were replaced by the cutoff value.*

The condition in question refers to factors that are applicable to the current study (i.e., type of task and type of presentation format). So, it doesn’t refer to related/unrelated. As such only the latter criterion referring to the grand mean is included.

*Exp1. No; Exp.2 "obvious outliers"; Exp.3-4 RT outside the upper, outer fence (Tukey, 1977) for each condition*

The latter is operationalized as Q3+1.5\*IQR. Here condition does refer to (different type of) related vs unrelated pairs.

*"Exp1: RTs below 250ms or above 2500ms, or those above or below 2 standard deviations of the participants’ mean for each condition;*

*Exp2-5: all incorrect responses, RTs < 250ms or > 2500ms, and those more than +/-2 SDs of the individual subject’s mean for each condition."*

Although not entirely clear, condition again seems to refer to related vs unrelated (among other factors).

*RTs >1,500 ms and any remaining RTs +- 2.5 SDs each participant’s condition mean.*

Condition again seems to refer to related vs unrelated.

*RT >3SD from the mean for each condition*

This also seems to refer to the mean per condition (related vs. unrelated) for each participant.

*After having fitted the final linear mixed effect model, we performed a residual analysis to identify possible outliers (i.e., observations with absolute standardized residuals greater than 2.5) --> re-fitted after exclusion*

Similar to the one about Cook’s distance. It is considered part of the analysis choices and hence, not included here.

**Missing data**

*For the word item analysis, four items were removed due to insufficient data in the first test block; One participant was removed from the participant-based (Fl) ANOVA due to insufficient data.*

Insufficient data is not operationalized, so not included (see earlier).

*sort of, they excluded conditions when participants didn't have a matching pair because they performed poorly on a different task*

Not applicable to the current study.

*Some data was lost due to power and equipment failures; the affected participants were excluded. Exp2: 1 participant didnt return for the 2 session*

Not applicable to the current study.

It seems that if there were missing data, it’s either due to equipment failure, or unavailability of supporting information (e.g., from norming data). If that happened, the corresponding data were excluded, and not imputed in some shape or form. For the current study, these issue shouldn’t arise, and if they do, the affected data will be removed.

**Data transformations**

*logarithmically transformed lexical frequency values of the target words (predictors)*

Not applicable to the current study.

*RTs were then log-transformed using Box–Cox power transformations*

As log-transformation is a special case of the Box-Cox transformation when lambda=0, this was simply coded as log-transformation.

*For correct RTs, a mean and standard deviation were calculated for each subject within each SOA and session*

As we don’t have session or SOA, this was coded as by-participant z-score transformation RTs

*The mean RT for the related condition was subtracted from the mean of the corresponding unrelated condition, then divided by the mean RT for the unrelated targets, and converted into a percentage.*

This was adapted for the current purposes as follows: The mean RT for the target in the related condition was subtracted from the mean of the target in the unrelated condition, then divided by the mean RT for the target in the unrelated condition, and converted into a percentage.