Language and sensorimotor simulation in conceptual processing: Multilevel analysis and statistical power

The present readme file is the most general one in the project. More specific notes are available in the readme files and R scripts within some folders.

R project

All the materials are available at https://doi.org/10.17605/OSF.IO/UERYQ. To reproduce or explore the analyses in R/RStudio, download the complete project from OSF by selecting the space 'OSF Storage (Germany - Frankfurt)', and clicking on 'Download as zip'.

After downloading and unzipping the files, the R project can be opened by clicking on the file 'thesis.Rproj'. This action activates the working directory and the package library, which are crucial for the reproducibility of the analyses. The packages are managed by the 'renv' package (see 'Packages' section below).

All the file paths specified in the `.R` scripts are relative to the root directory (namely, the most general folder, in which the file 'thesis.Rproj' is located).

Project-specific functions

The 'R_functions' folder contains several custom functions, which can be loaded at once by running `load_functions()`.[1]

[1] The function `load_functions()` is defined in .Rprofile (root directory).

Primary and final data sets

Each of the three studies (Semantic priming, Semantic decision and Lexical decision) involve a core data set and several variables from other data sets. In each study, all the primary data sets were downloaded and saved using the code stored in the 'data' folder within each study. Below is an example of the code used to download the data:

```
GET('https://www.montana.edu/attmemlab/documents/all%20ldt%20subs_all%20trials3.xlsx',
    write_disk(tf <- tempfile(fileext = ".xlsx")))
Hutchison_lexicaldecision = read_excel(tf)
write.csv(Hutchison_lexicaldecision,
          'semanticpriming/data/primary_datasets/Hutchison_lexicaldecision.csv')
```

To prevent the influence from any future changes to the original data sets online, the code used to download the primary data sets such as the above one is currently *protected* (i.e., commented out) in the scripts by the snippet `# Protected code #`. If desired, the code can be run by removing the snippet.

Exceptionally, one data set could not be downloaded using code. This was a data set used in the lexical decision study, and named 'ELP measures for lexical decision study.csv'. These data could only be downloaded through a web application from the English Lexicon Project. The procedure for this download is detailed in the following script:

```
lexicaldecision/data/lexicaldecision_data_preparation.R
```

The final data sets are stored in `data/final_dataset` within each study.

Analysis and results

The main analysis scripts are in the 'frequentist_analysis' folder within each study (e.g., 'semanticpriming/frequentist_analysis/'). The names of these files end in '_lmerTest.R' (e.g., 'semanticpriming_lmerTest.R'). Further folders include 'bayesian_analysis', 'frequentist_bayesian_plots' and 'power_analysis'.

The results produced by code—i.e., statistical models and plots—are stored reasonably close to the corresponding code script, and the names of the files are reasonably transparent. For instance, the plot that is stored in

```
semanticpriming/frequentist_analysis/model_diagnostics/results/plots/semanticpriming_residuals.png
```

can be produced in the script

```
semanticpriming/frequentist_analysis/model_diagnostics/semanticpriming_residuals.R
```

Directory tree

Below is a directory tree containing the main folders of the project (no files shown). The tree was produced in the RStudio 'Terminal' console using the following one-line command, and the main folders were then copied from the output.

```
find . -type d | sed -e "s/[^-][^\/]*\//  |/g" -e "s/|\([^ ]\)/| - \1/"
```

```
.
  | - bayesian_priors
  |  | - plots
  | - semanticpriming
  |  | - analysis_with_visualsimilarity
  |  |  | - model_diagnostics
  |  |  |  | - results
  |  |  |  | - plots
  |  |  | - results
  |  |  | - plots
  |  |  | - correlations
  |  |  |  | - plots
  |  | - frequentist_bayesian_plots
  |  |  | - plots
  |  | - frequentist_analysis
  |  |  | - model_diagnostics
  |  |  |  | - results
  |  |  |  | - plots
  |  |  | - lexical_covariates_selection
  |  |  |  | - results
  |  |  |  | - plots
  |  |  | - results
  |  |  | - plots
  |  | - power_analysis
  |  |  | - reduced_randomeffects_model
  |  |  |  | - model_diagnostics
  |  |  |  |  | - results
  |  |  |  |  | - plots
  |  |  |  | - results
  |  |  |  | - plots
```

```
|  |  | - individual_scripts
|  |  |  | - scripts_using_smaller_randomeffects_model
|  |  | - results
|  |  | - plots
|  | - bayesian_analysis
|  |  | - posterior_predictive_checks
|  |  |  | - results
|  |  |  | - plots
|  |  | - divergent_models
|  |  |  | - results
|  |  | - prior_predictive_checks
|  |  |  | - plots
|  |  | - results
|  | - data
|  |  | - primary_datasets
|  |  | - subset_with_visualsimilarity
|  |  | - final_dataset
|  | - correlations
|  |  | - plots
| - lexicaldecision
|  | - frequentist_bayesian_plots
|  |  | - plots
|  | - frequentist_analysis
|  |  | - model_diagnostics
|  |  |  | - results
|  |  |  | - plots
|  |  | - lexical_covariates_selection
|  |  |  | - results
|  |  |  | - plots
|  |  | - results
|  |  | - plots
|  | - power_analysis
|  |  | - individual_scripts
|  |  | - results
|  |  | - plots
|  | - bayesian_analysis
|  |  | - posterior_predictive_checks
|  |  |  | - results
|  |  |  | - plots
|  |  | - divergent_models
|  |  |  | - posterior_predictive_checks
|  |  |  |  | - plot
|  |  | - prior_predictive_checks
|  |  |  | - plots
|  |  | - results
|  |  | - plots
|  | - data
|  |  | - primary_datasets
|  |  | - final_dataset
|  | - correlations
|  |  | - plots
| - semanticdecision
|  | - frequentist_bayesian_plots
|  |  | - plots
```

```
|  | - frequentist_analysis
|  |  | - model_diagnostics
|  |  |  | - results
|  |  |  | - plots
|  |  | - lexical_covariates_selection
|  |  |  | - results
|  |  |  | - plots
|  |  | - results
|  |  | - plots
|  | - power_analysis
|  |  | - individual_scripts
|  |  | - results
|  |  | - plots
|  | - bayesian_analysis
|  |  | - posterior_predictive_checks
|  |  |  | - results
|  |  |  | - plots
|  |  | - divergent_models
|  |  |  | - results
|  |  | - prior_predictive_checks
|  |  |  | - plots
|  |  | - results
|  | - data
|  |  | - primary_datasets
|  |  | - final_dataset
|  | - correlations
|  |  | - plots
| - R_functions
| - thesis
|  | - thesis-core_files
|  |  | - figure-latex
| - general_datasets
| - renv
```

Original settings and reproducibility of results

The reproducibility of the analyses is more certain if the same versions of R and the relevant packages are used. Code scripts that required more than one hour (i.e., all statistical analyses and some plots) were run in R 4.0.2 or 4.1.0 (the latter was only used for the Bayesian analyses), whereas faster scripts (i.e., most plots) were run in R 4.1.2. For convenience, R 4.1.2 can likely be used to reproduce all scripts. Later versions could work too.

- Windows: https://cran.r-project.org/bin/windows/base/old/4.1.2/R-4.1.2-win.exe
- macOS: https://cran.r-project.org/bin/macosx/base/R-4.1.2.pkg
- Linux: https://cran.r-project.org/bin/linux/

  - Select Version 4.1.2 or the closest available
- Source files for other cases: https://cloud.r-project.org/src/base/R-4/R-4.1.2.tar.gz

Time-consuming analyses

Analyses requiring more than one hour were run on the High-End Computing facility (HEC) at Lancaster University (information available at https://answers.lancaster.ac.uk/display/ISS/High+End+Computing+%28HEC%29+help). Such a facility (generally known as High-Performance Computing) allows the running of multiple scripts in parallel. While using this is not required for the reproducibility of the analyses, it would be helpful, as some scripts take several hours to run (i.e., the lmerTest models for Studies 1 and 2), and other scripts take more than a month (i.e., the power curves for Study 1).

Importantly, it must be noted that the number of cores to be engaged is hard-coded in two sets of scripts. Firstly, in the 'allFit_convergence' scripts located in 'frequentist_analysis/model_diagnostics', the number of cores is specified with the parameter `ncpus`. Secondly, in the scripts located in 'bayesian_analysis', the number of cores is specified with the `cores` parameter.

If the code must be expedited (at the expense of precision), the following steps can be taken:

1. In the 'lmerTest' models (e.g., 'semanticpriming_lmerTest.R'),

- within the `summary` function, replace `ddf = 'Kenward-Roger'` with `ddf = 'lme4'` (see background);
- within the `confint.merMod` function, replace `method = 'profile'` with `method = 'Wald'` (see background).

2. In the 'brms' models ('semanticpriming_brms_informativepriors_exgaussian.R'),

- reduce the numbers in `warmup`, `iter`, `chains` and `cores`

3. In the individual power curves, replace `nsim = 200` with `nsim = 30` (see background).[2]

[2] An example of an individual power curve: 'powercurve1_400participants_200sim_semanticpriming_lmerTest.R'.

Packages

The 'renv' package is used to record all the packages used, and their versions, in the file 'renv.lock', within the root directory. To use 'renv' as the package library (recommended), the current project must be opened by double-clicking on the file 'thesis.Rproj'. When the project is opened for the first time, the R console will show information on how to install the necessary packages, following the records in 'renv.lock'. Basically, this will prompt you to run `renv::restore()`, which will install the necessary packages in their appropriate versions. Some of the output will look like the following:

```
- xml2              [* -> 1.3.3]
- xopen             [* -> 1.0.0]
- xts               [* -> 0.12.1]
- zip               [* -> 2.2.0]
- zoo               [* -> 1.8-10]

# GitHub ============================
- osfr              [* -> ropensci/osfr@HEAD]
- papaja            [* -> crsh/papaja@devel]
```

Packages whose versions are most important

The main packages used for the statistical analyses are listed below, along with their versions. Although the 'renv' library is the recommended option, the version of each package could otherwise be installed using the commands below.e

```
install.packages('devtools')
library(devtools)
```

- lme4 1.1-26: `devtools::install_version('lme4', '1.1-26')`
- lmerTest 3.1-3: `devtools::install_version('lmerTest', '3.1-3')`
- robustlmm 2.4.4: `devtools::install_version('robustlmm', '2.4.4')`
- afex 1.0-1: `devtools::install_version('afex', '1.0-1')`
- brms 2.17.0: `devtools::install_version('brms', '2.17.0')`
- simr 1.0-5: `devtools::install_version('simr', '1.0-5')`

Thesis

A reproducible document containing Pablo Bernabeu's PhD thesis is available in the 'thesis' folder.

References

The following references are mentioned in some scripts.

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

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., & Krivitsky, P. N. (2021). *Package 'lme4'*. CRAN. https://cran.r-project.org/web/packages/lme4/lme4.pdf

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods, 23*(3), 389–411. https://doi.org/10.1037/met0000159

Bürkner, P.-C., Gabry, J., Weber, S., Johnson, A., Modrak, M., Badr, H. S., Weber, F., Ben-Shachar, M. S., & Rabel, H. (2022). *Package 'brms'*. CRAN. https://cran.r-project.org/web/packages/brms/brms.pdf

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

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, 1–15. https://doi.org/10.3758/s13428-021-01587-5

Lo, S, & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior*

Research Methods, 49(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods, 52*, 1271–1291. https://doi.org/10.3758/s13428-019-01316-z.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Nicenboim, B., Schad, D. J., & Vasishth, S. (2022). *An introduction to Bayesian data analysis for cognitive science*. https://vasishth.github.io/bayescogsci/book/

Petilli, M. A., Günther, F., Vergallito, A., Ciapparelli, M., & Marelli, M. (2021). Data-driven computational models reveal perceptual simulation in word processing. *Journal of Memory and Language, 117*, 104194. https://doi.org/10.1016/j.jml.2020.104194

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: Concrete/abstract decision data for 10,000 English words. *Behavior Research Methods, 49*(2), 407–417. https://doi.org/10.3758/s13428-016-0720-6

Pexman, P. M., & Yap, M. J. (2018). Individual differences in semantic processing: Insights from the Calgary semantic decision project. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(7), 1091–1112. https://doi.org/10.1037/xlm0000499

Rodríguez-Ferreiro, J., Aguilera, M., & Davies, R. (2020). Semantic priming and schizotypal personality: Reassessing the link between thought disorder and enhanced spreading of semantic activation. *PeerJ, 8*, e9511. https://doi.org/10.7717/peerj.9511

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution, 11*, 9, 1141–1152. https://doi.org/10.1111/2041-210X.13434

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Sachar, M. S. (2021). *Package 'afex'*. CRAN. https://cran.r-project.org/web/packages/afex/afex.pdf

Singmann, H., & Kellen, D. (2019). An Introduction to Mixed Models for Experimental Psychology. In D. H. Spieler & E. Schumacher (Eds.), *New Methods in Cognitive Psychology* (pp. 4–31). Hove, UK: Psychology Press.

Stone, K., Malsburg, T. von der, & Vasishth, S. (2020). The effect of decay and lexical uncertainty on processing long-distance dependencies in reading. *PeerJ, 8*, e10438. https://doi.org/10.7717/peerj.10438

Stone, K., Veríssimo, J., Schad, D. J., Oltrogge, E., Vasishth, S., & Lago, S. (2021). The interaction of grammatically distinct agreement dependencies in predictive processing. *Language, Cognition and Neuroscience, 36*(9), 1159–1179. https://doi.org/10.1080/23273798.2021.1921816

Wingfield, C., & Connell, L. (2022). Understanding the role of linguistic distributional knowledge in cognition. *Language, Cognition and Neuroscience*, 1–51. https://doi.org/10.1080/23273798.2022.2069278