**Reproducibility in Second Language Research**

**Authors:** Anders Søndergaard Røn and Mette Hejberg Pedersen

**Keywords:** Second Language Research; Reproducibility; Priming

**Abstract**

English is an integral part of Danish society, and it is therefore important that researchers investigate how Danish speakers learn and speak English. This project investigated native Danish speakers’ perception of the English sound contrast between /s/ and /z/ and how to improve their perception through a priming study. Although the results show that the participants’ perception did not improve due to a priming effect, the project also sought to add to the literature within linguistics that attempts to close the reproducibility gap. Linguistics is moving towards sharing code and data to close this gap, which this project also does.

**1 Introduction**

Most people living in Denmark speak a second language. Many even speak more than two languages. Within the repertoire of languages that the citizens of Denmark speak, you will most likely find English. Therefore, it is important to understand how native Danish speakers learn English and what their struggles are. This paper investigates how we can improve native Danish speakers’ perception of English speech sounds that have been shown to be difficult for native Danish speakers to perceive and discriminate between. The most important aim through this investigation was to contribute to the reproducibility within the field of second language research.

**2 Problems and Background**

How we learn our second language is of interest to the public because language and culture are linked. English is an integral and important part of Danish society, and therefore it is important to understand how native Danish speakers learn, understand, and speak English. A deeper understanding of the mechanisms behind second language acquisition and second language speech informs teaching methods all the way from the first time a child is introduced to English in school to immigrants learning English or another second language later in life. Furthermore, the findings from second language research informs government policy.

Within many scientific fields there exists a reproducibility crisis and linguistics is no different (Winter 2019). This means that researchers have been unable to reproduce findings from studies, which naturally questions their validity. One way to achieve better reproducibility is to share and document your code according to the FAIR principles (Wilkinson et al. 2016). However, as it turns out, using programming as a tool for data analysis is only now taking on within linguistics. Therefore, it is a far way to go before we can solve the reproducibility crisis just within linguistics. This project sought to reproduce previous findings within second language research, test new methods, and contribute to the documentation of code to close part of the gap in reproducibility within linguistics.

The project sought to reproduce the findings that native Danish speakers perceive the English speech sounds /f/ and /v/ very well (Horslund, Ellegaard, and Bohn 2015), and that they also perceive the English speech sounds /s/ and /z/ if they are at the end of the syllable (Eger and Bohn 2015). The first case is interesting because /f/ and /v/ only differ slightly from English to Danish (Horslund, Ellegaard, and Bohn 2015, 2). Therefore, one should expect native Danish listeners to perform very well on this task. The second case is interesting because Danish only has /s/ and not /z/ and it is unclear why native Danish speakers have been found to perform very well on this task (Eger and Bohn 2015, 1). This project also sought to reproduce findings suggesting that if you example use duration to distinguish vowels (long or short vowels) in your first language, then you can transfer that knowledge to use duration to distinguish consonants in your second language (McAllister, Piske, Flege 2002). Specifically, we tested whether hearing /f/ and /v/ could improve perception of /s/ and /z/ in the beginning of a syllable, because both contrasts differ in voicing, which is whether your vocal cords vibrate or not while producing the sound. We also tested whether perception of /s/ and /z/ in the beginning of a syllable could improve by hearing /s/ and /z/ in the end of a syllable. To test and reproduce these findings, we used a priming study design. To briefly outline the study design, there were three groups, two experimental and one control group. One experimental group heard /f/ and /v/ and the other heard /s/ and /z/ in the end of a syllable. The control group heard an unrelated contrast. The study was divided into three sessions where session 1 and 3 were identical and provided the data for main part of the statistical analysis which would reveal whether there was a priming effect. In this session all participants heard /s/ and /z/ in the beginning of a syllable. Session 2 was the intervention session with either the priming contrasts or the unrelated contrast.

A research project takes a long time to complete, and it is therefore not possible to discuss all phases of the project. Originally the project also sought to optimize certain parts of the data manipulation that many researchers do which are very tedious tasks. This included automatic segmentation of the sound files from a single file into individual files in Bash and manipulating the quality of the sound files with the Parselmouth package for Python (Jadoul 2021). However, both tasks, although they are tedious and prone to error, do not contribute to the reproducibly crisis in linguistics. Therefore, together with the fact that the documentation of the Parselmouth packages remains sparse, we chose to exclude the optimization of these tasks in the present project, because we found the issues of reproducibility much more pressing. However, optimization of these tasks would be ideal and this would therefore constitute interesting and important future projects.

**3 Software Framework**

We wrote the code for this project on a 1,5-year-old MacBook Air 2019 MVFK2D 128GB, where R (version 4.1.1) and RStudio (version 1.4 1717) were installed. We were able to write all code on a single computer and therefore not use online collaboration tools. This was suitable for us due to our skill level and helped us improve our programming skills greatly. This also means that we had less need for version control.

**4 Data Acquisition and Processing**

We collected our own behavioral data for this project. However, the stimuli of speech sounds were retrieved from a data bases provided by Shannon et al (1999). We segmented and manipulated the files in *Praat* (Boersma and Weenink 2020)*.* We provide a prewritten script that conducted the quality manipulation for us (see appendix). Furthermore, the project was conducted online and ran through the PERCY software provided by Christoph Draxler at Munich University (Draxler 2014; 2021). The PERCY software automatically assigned groups to the participants depending on which group had the fewest participants. We retrieved the data through the links in the appendix.

**5 Empirical Results**

The results of our investigation showed that there was no effect of priming in either of the groups. This means that there was no significant difference between session 1 and 3. Furthermore, as expected there was neither a difference between session 1 and 3 for the control group. To reach this conclusion we wrangled our data using the tidyverse package (Wickham et al. 2019). The original data was compiled into two csv files, one for session 1 and one for sessions 2 and 3 because they were completed directly after each other. The first part of the wrangling separated the trial rounds that we ourselves completed of the experiment from the participants’ completions. We loaded the session 1 csv file into R and filtered the data according to dates. First, we converted the date information in the “startdate” column from characters to dates. Then we removed all completions of the experiment prior to 26/10 – 2021, because we knew this was the date of the first entry we wanted to include. We then selected only the columns we were interested in which contained information about participant identification, stimulus, expected response, participant response, language background, and hearing. We renamed those columns to names that were more semantically transparent. We wrangled the data so that the participants in both data frames were the same. Through this we discovered there was a participant that had completed the experiment more than once and had indicated that Thai was their native language in one of these competitions. Since the participant indicated that they Danish was their native language in all other completions, we decided to include this participant in our analysis. Before we could include this participant, we needed to exclude all other completions apart from the first one and code their native language to Danish. We did the same operations for the sessions 2 and 3 csv file. There were 20 participants in total, 8 in the /f/-/v/ group (voicing priming), 7 in the syllable-final /s/-/z/ group (position priming), and 5 in the /f/-/th/ group (control group).

First, we looked at the baseline data of how well all participants collectively did in session 1 to get an idea of how difficult the task was. We put the stimulus (/s/ and /z/) on the x-axis and mean accuracy on the y-axis. The boxplot below illustrates this data. We chose to use boxplots to visualize our data because they lend themselves well to categorical predictors. You lose some information about the spread in your data in boxplots, however, we were not too concerned with spread in this project.

Chart, box and whisker chart

Description automatically generated

The scattered dots illustrate the distribution of data points. In reality they are not scattered of course, because the x-axis is not continuous, however it makes it easier to see their distribution. The blue triangles are the mean accuracies. For their discrimination to be at the level we would expect of a native English speaker, it should be above 90%, and it therefore seems that the task was difficult, however they are able to perceive the difference since their discrimination is not at chance level either. It looks like from the plot that the participants were better at identifying /z/ than /s/. To confirm this observation, we created a subset of the data and fitted a linear model in R using the lm() function from the broom package (Robinson, Hayes, and Couch 2021). Table 1 shows the output of this model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term | Estimate | Std. Error | t value | p value |
| (Intercept) | 0.70833 | 0.03964 | 17.868 | <2e-16 |
| Stimulusz | 0.08500 | 0.05606 | 1.516 | 0.138 |

Table 1: Output of model where session was a predictor of accuracy

Chart, box and whisker chart

Description automatically generatedNotice that the very low p-value for the intercept does not reflect significance as it is not a comparison. Since the predictor is categorical and we have chosen to use treatment coding, the accuracy of “s” responses is in the intercept and the accuracy of “z” responses is at 1. Therefore, the estimate for the intercept is the accuracy of “s” responses and the estimate for the “z” responses is the amount you move up the y-axis for a one unit change on the x-axis. The output of the model shows that there was no significant difference between the accuracy of “s” and “z” responses. This is surprising as we expected the participants to perform better for the /s/ stimulus and this finding therefore suggests that the participants generally were very proficient in English.

Next, we plotted the data from all three sessions. To do this we first needed to merge the two data frames which we did using the full\_join() and left\_join() functions. The boxplot above visualizes the complete dataset where session is on the x-axis, accuracy is on the y-axis, and contrast is the filler. To dots in the second session in the control group are outliers, however we kept them in for full transparency because there were fewer participants in this group. Therefore, it is also spurious to do statistics and conclude anything general from this dataset. Looking at the plot, it seems like the position priming group was more accurate in session 1 compared to the other two groups. It also looks like that there is a difference in accuracy between sessions 1 and 3 for both the control group and the position group. However, both of these differences are in the unexpected direction. The control group gets better over time and the position priming group gets less accurate over time. Lastly, the voicing priming group seems to be more accurate in the intervention session than the other two groups, which we in fact expected because /f/ and /v/ are Danish sounds. When the boxplots are long it tells us that the data points are widely distributed and it therefore seems that most of the tasks were quite difficult. Only the /f/-/th/ and /f/-/v/ contrasts were easy tasks. However, it is important to remember once again that the control group only included 5 participants, and since there is an outlier, the boxplot only covers 4 participants. The boxplot for the /f/-/v/ contrast on the other hand covers 7 participants and it is therefore somewhat safer to assume that this reflects a robust perceptual pattern. We created another boxplot, only of the sessions 1 and 3 data, which is seen in figure 3. It also seems that there was a difference between sessions 1 and 3 for the control group and the position priming group from figure 3.

To confirm the observations made from visual inspection of the boxplot, we created a subset of the data and fitted a linear model using the same function, however this time there were two categorical predictors and an intervention. Session (two levels: sessions 1 and 3) and contrast (three levels: voicing priming, position priming, control) were predictors of accuracy. Table 2 shows the output of this model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Estimate | Std. Error | t value | p value |
| (Intercept) | 0.706667 | 0.072306 | 9.773 | 2.09e-11 |
| GroupVoicing Priming | 0.001667 | 0.092173 | 0.018 | 0.986 |
| GroupPosition Priming | 0.124286 | 0.094671 | 1.313 | 0.198 |
| SessionSession 2 | 0.093333 | 0.102256 | 0.913 | 0.368 |
| GroupVoicing Priming:SessionSession 2 | -0.128750 | 0.130352 | -0.988 | 0.330 |
| GroupPosition Priming:SessionSession 2 | -0.148095 | 0.133885 | -1.106 | 0.276 |

Chart, box and whisker chart

Description automatically generated

Table 2: Output of model where session and contrasts interacted to predict accuracy

The output of the model shows that there was no significant difference between the session 1 data between any of the groups. The output further shows that there was no interaction and therefore neither a significant difference between groups in the session 3 data. The observations regarding this part of the data were therefore not confirmed. To confirm the observations about the intervention session, we took another look at the data. We created a new boxplot, which is seen in figure 4, which shows sessions 1 and 2.

Chart, box and whisker chart

Description automatically generated

The observation that the voicing priming group was more accurate in session 2 than the two other groups seems to hold true. We therefore fitted a linear model identical to the former except the session predictor’s levels were now “session 1” and “session 2”. Table 3 shows the output of this model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Estimate | Std. Error | t value | p value |
| (Intercept) | 0.706667 | 0.073549 | 9.608 | 3.21e-11 |
| GroupVoicing Priming | 0.124286 | 0.096298 | 1.291 | 0.2055 |
| GroupPosition Priming | 0.001667 | 0.093757 | 0.018 | 0.9859 |
| SessionSession 2 | 0.016667 | 0.104014 | 0.160 | 0.8736 |
| GroupVoicing Priming:SessionSession 2 | -0.126190 | 0.136186 | -0.927 | 0.3607 |
| GroupPosition Priming:SessionSession 2 | 0.231250 | 0.132592 | 1.744 | 0.0902 |

Table 3: Output of model where session and contrasts interacted to predict accuracy

The output of the model first of all confirms what we already know, namely that there was no difference between groups in session 1. However, there was neither an interaction and therefore neither a difference between groups in the session 2. The observations made from figures 3 and 4 were therefore not confirmed.

**6 Critical Evaluation**

The results of the analysis do not confirm many of the expectations we had going in. As mentioned above, we set out to test the effect of priming and reproduce earlier findings that we assumed in our study design. We did not find an effect of priming. There are many potential reasons for this. One explanation could be that the contrasts we chose for the intervention session were not ideal to elicit a priming effect. This is highly probable especially in the light of the performance of the participants in position priming group in the intervention session. We chose this contrast because of Eger and Bohn’s (2015) finding, however we were unable to reproduce it here. The voicing priming contrast was perhaps neither ideal because English /v/ and Danish /v/ differs slightly, and seeing as the contrast was English, it might not have been optimal to elicit a priming effect. Another explanation is that phonetic priming is not possible. The literature on this topic reports varying results about the effect of phonological priming and Durvasula and Parrish (2018) argue that it is not possible. This is an interesting possibility for future research. Lastly, the number of participants could pose a problem for the results. There were 20 participants in total and only 5 in the control group which unfortunately included an outlier. Therefore, it would be spurious to conclude that these results reflect any robust perceptual patterns, and it is therefore entirely possible that we would find different patterns with new and more participants.

These results suggest that native Danish speakers need more exposure to English speech sounds in order to improve perception, which further suggests that a more immersive environment should improve your second language skill level. Therefore, it would be beneficial for children in schools to speak with native English speakers in order for them to become highly proficient and ultimately become more understandable in their second language.

**7 Conclusions**

This project shows that we cannot improve native Danish speakers’ perception of syllable-initial /s/-/z/ through priming. We have shown this through a detailed description of the study design, procedures, and data analysis and we provide all the necessary materials needed to reproduce our findings. Therefore, we take part in the steps towards higher reproducibility in linguistics.

**Table 4: Software metadata**

|  |  |  |
| --- | --- | --- |
| **Nr** | **Software metadata description** |  |
| S1 | Current software version | R (version 4.1.1) and RStudio (version 1.4 1717) |
| S2 | Permanent link to executables of this version (your Github repo URL) | *example : https://github.com/combogenomics/DuctApe/releases/tag/DuctApe-0.16.4* |
| S3 | Legal Software License | *List one of the approved licenses, e.g. Creative Commons 4.0; see Week 6 lecture recordings for more* |
| S4 | Computing platform / Operating System | macOS Monterey |
| S5 | Installation requirements & dependencies for software not used in class | Praat(Boersma, P., & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.16. Retrieved 3 February 2018.)  PERCY (Draxler, Christoph. 2021. <http://webapp.phonetik.uni-muenchen.de/WebExperiment/>) |
| S6 | If available Link to software documentation for special software |  |
| S6 | Support email for questions | [201708941@post.au.dk](mailto:201708941@post.au.dk) |

# Table 5: Data metadata

|  |  |  |
| --- | --- | --- |
| **Nr** | **Metadata description** |  |
| D1 | First\_session\_17\_11.txt | The participants’ performance in session 1 and participant information |
| D2 | Second\_session\_17\_11.txt | The participants’ performance in sessions 2 and 3 |

**References**

Boersma, P., and D Weenink. 2020. *Praat: Doing Phonetics by Computer* (version Version 6.1.16.).

Durvasula, Karthik, and Alicia Parrish. 2018. “Is There Phonological Feature Priming?” *Linguistics Vanguard* 5 (1): 1–11. https://doi.org/10.1515/lingvan-2018-0041.

Eger, Nikola Anna, and Ocke-Schwen Bohn. 2015. “Picking up the Cues to a New Consonant Contrast: Danish Learners’ Production and Perception of English Word-Final /s/-/z/.” In *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–5.

Horslund, Camilla S., Anne A. Ellegaard, and Ocke-Schwen Bohn. 2015. “Perceptual Assimilation and Identification of English Consonants by Native Speakers of Danish.” In *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–5.

Robinson, David, Alex Hayes, and Simon Couch. 2021. *Broom: Convert Statistical Objects into Tidy Tibbles* (version R package version 0.7.10). https://CRAN.R-project.org/package=broom.

Shannon, Robert V., Angela Jensvold, Monica Padilla, Mark E. Robert, and Xiaosong Wang. 1999. “Consonant Recordings for Speech Testing.” *The Journal of the Acoustical Society of America* 106 (6): 71–74. https://doi.org/10.1121/1.428150.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the Tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Winter, Bodo. 2019. *Statistics for Linguists: An Introduction Using R*. New York: Routledge.