Cimiano P., Pietsch C., & Wiljes C. (Eds.) (2021). Studies in Analytical Reproducibility: the Conquaire Project. Bielefeld. https://doi.org/10.4119/unibi/2942780

7 | Reproducing experiments on early verb understanding in infants

Vidya Ayer[1], Christian Witte[1], Philipp Cimiano[1], Katharina J. Rohlfing[2], Iris Nomikou[3]

1 – Semantic Computing Group Faculty of Technology & Cognitive Interaction Technology Excellence Center (CITEC), Bielefeld University
2 – Psycholinguistics, Faculty of Arts and Technology, Paderborn University
3 – Department of Psychology, University of Portsmouth

Abstract

In this chapter, we describe an effort to reproduce the main result of the paper *"Evidence for early comprehension of action verbs"* by Nomikou et al. [1]. The study aimed at investigating the ability of 9-and-10-month old infants in understanding verbs. The study in question followed a so called *preferential looking paradigm* consisting in investigating the ability to understand the meaning of words by testing whether infants look longer at a related stimulus compared to an unrelated stimulus. As a method to track looking time proportions to the target stimulus, an eye tracker was used. Data were collected from 9- to 10-month old infants who were presented with paired-picture trials while listening to corresponding verbs. The infants saw two images on the screen side by side, each one from a different context category (CARE or PLAY). One of the pictures was related to the verb in question, while the other image was a confounder. The percentage of time that infants looked at the matching picture, before and after having heard the verb, was recorded across participants, computing a difference score. In case the difference was positive, this was taken as evidence of understanding the meaning of the verbs. The study could only find a positive difference for 10-month olds but not for 9-month olds, showing that the ability to understand the verb in question emerges between 9 and 10 months. In close interaction with the authors of the original paper we rewrote the analysis scripts which were used by the authors to refine the results during a second iteration of reviewing in response to requests by reviewers. Overall, we could reproduce the central results of the study. This case represents a case of *full analytical reproducibility.*

The data and scripts for the paper described above can be found at `https://gitlab.ub.uni-bielefeld.de/conquaire/psycholinguistics`.

Keywords

Psycholinguistics, Language learning, Verb Understanding, Infants, Eye Tracking

7.1 Introduction

The Psycholinguistics research group at Paderborn University is concerned with investigating language development in young children. Its main research interest is how children acquire the meaning of words to reveal links between language and cognitive development and to analyze early meanings as building blocks for conceptual and linguistic thinking.

There is a debate on the question whether nouns are acquired before verbs. In contrast to nouns, which can be easily singled out by holding an object or pointing to it, verbs are relational since they combine agents and their actions with some objects. In consequence, verbs have a more complex semantic structure. However, it is also possible that early use of nouns is cumulative as it also binds together situational elements. For example, an infant might say *ball* but relate this noun to the action of rolling [2].

In the study reproduced as part of this work titled *'Evidence for early comprehension of action verbs'* [1], the research group studied 9- and 10-month-old infants' understanding of verbs using a technique similar to the one used by Bergelson et al. [3], except that verbs were used in place of nouns. Following the conceptual development approach proposed by Mandler [4], the hypothesis was that infants must conceptualize situated actions early in their development to get concepts about objects off the ground. Early concepts, thus, will entail the role of objects, i.e., what the objects do and what is done to them [4], which provides a solid basis for the acquisition of verbs. Thus, the hypothesis was that children at a younger age, as found so far, will understand verbs that are drawn from their everyday life contexts. Instead of using dynamic pictures that refer to verbs, static object pairs were used and parents were asked to utter the relevant verbs.

In this work, we aim at reproducing the main result of the paper mentioned above, i.e. that demonstrated early verb understanding by showing that infants tend to look longer at the correct target picture once their parent uttered the corresponding verb. The study found a developmental difference between 9- and 10-month olds though: 9-month-olds were not able to reliably demonstrate verb understanding. With respect to early semantic development, as visualized by the target looking times, the data suggests that on hearing a verb, the infants can associate it to object stimuli related to the verb. This is in line with the researchers argument claiming that action concepts can be evoked in object perception. Furthermore, the results complement research proposing that children learn language by building relations and drawing from rich visual concepts [5].

7.2 Methods

Here, we describe the methods used for the experimental settings in the original experiment.

7.2.1 Experimental settings and data acquisition pipeline

The study in question followed a so called *preferential looking paradigm* consisting in investigating the ability to understand the meaning of words by testing whether infants look longer at a related stimulus compared to an unrelated stimulus. As method to measure target looking times, an eye tracker was used. Data were collected from 9- to 10-month old infants who were presented with paired-picture trials. The infants saw two images on the screen side by side, each one from a different context category (CARE or PLAY). One of the pictures was related to the verb in question, while the other image was a confounder. These images were shown for a total of 9.5 seconds. Within the first 3s of each trial, parents heard a beep before they heard a sentence that they were asked to reproduce. Then, a second beep prompting them to begin repeating the sentence. While the parent was saying the target verb, the experimenter pressed a key on a wireless keyboard to mark the precise moment at which the verb was perceivable to the infant. This mark was logged into the data. An attention getter, i.e. a 3s clip featuring colorful animated shapes accompanied by different sounds, appeared after each trial. The experiment lasted 5 minutes. The entire visit of the infants to the lab lasted 45 minutes.

Because of individual differences in the production of the target phrase by the parent, the post-target analysis window extended from 367 to 4.500 ms after the onset of the spoken target word. To calculate the onset of the target word, the recorded time-stamp of the keyboard key press was used. A Python script was used to split the looking times into two periods: before and after the uttered verb. The dependent variable, namely, word comprehension, was thus operationalized by a difference between the proportion of target looking upon hearing the target word (367 to 4.500 ms post keyboard keypress) minus the proportion of target looking before hearing the word (from when pictures were displayed until just before the keyboard keypress). This way, a difference value was obtained that could be positive or negative. If the value was positive, it indicated increased looking at the target object by the infants after hearing the verb, thus demonstrating their understanding of the target word.

7.2.2 Methods applied to analyze the data

The raw eye-tracking data were filtered using python scripts according to predefined areas of interest (AOIs). Then total gaze durations at the AOIs were calculated and subsequently the script took into account a specific timestamp generated by a key press of the keyboard and calculated the gaze durations before and after the keypress as well as the proportions of gaze at the target or distractor AOIs before and/or after the keypress. These calculations were formatted in a table and used for further calculations. These included before-after difference scores for each of the two presented instances of each pair of stimuli, with the two difference scores being subsequently averaged. These difference scores were then used in a series of statistical tests: t-tests, ANOVAs and binomial tests. For details, the reader is referred to the original publication [1].

In a subsequent review round of the submitted manuscript, various versions of the initial script were produced in collaboration with the Conquaire project to repeat the analysis using a fixed time window for the inclusion/exclusion of data points. This was requested by the paper reviewers. To address this comment, three new versions of the scripts were created with varying window durations, the changes incurred were assessed by comparing the results of a sample of data files and the usage of the script with a 4500ms time-window was selected to re-run the analysis and all the statistical tests.

During the creation and implementation of the scripts, both initially and in the second round of analysis, there was intensive collaboration between members from the psycholinguistics group and the Conquaire team. This was necessary to check for errors in the scripts. For this, random manual calculations were performed on the raw data and then compared with the results produced by the scripts to test for accuracy. In some cases, multiple iterations were needed until the systematicity in the discrepancy between script and manually calculated results was discovered and corrected.

7.2.3 Main Results

Using the process detailed above, the scripts produced tables of variables ready for statistical analysis. A mixed, between, and within-subjects ANOVA was used with AGE (9 months vs. 10 months old) as the between-subjects variable, and TIME (before vs. after the word was spoken) as the within-subjects variable. There was a significant AGE x TIME interaction effect $F(1, 46) = 5.687, p < .021, \eta = .107$. Since an independent-samples t-test indicated significant differences between the 9 and 10 months olds, the data were treated in separate groups. Additionally, a linear regression was calculated with the increase in looking times at the target as the dependent variable and infants' age in days as an independent variable. The regression model did not attain significance, suggesting that the change in performance was not linear, $F(1, 46) = 2.23, p = 0.142$.

7.3 Analytical Reproducibility

Computational reproducibility experiments were conducted with the Psycholinguistics research group at Paderborn University at the paper publishing stage to modify the data analysis scripts and produce results, then implement visualizations with Pandas and matplotlib that was later stored in GitLab under continuous integration. To facilitate team-collaboration on porting and refactoring the code, the python scripts and extracted (TSV format) files for data analysis are available at the following Git repository: `https://gitlab.ub.uni-bielefeld.de/conquaire/psycholinguistics`.

Primary Data

The data in the git repository include the images seen by the infants on the screen, the recordings heard by the parents, the eye-tracking data and the 3s attention getter clip featuring colorful animated shapes moving to different sounds that appeared after each trial. Excel sheets with information identifying participants were not uploaded to the GIT repository due to privacy protection issues.

Analysis Data

The python scripts and extracted data (TSV) files for analysis are stored in the **data_output** folder on gitlab. The research data structure (in the TSV and Excel) files are described below: The TSV files are stored in the "data_output" folder within the subdirectory folders, viz. "tables_3500", "tables_4000" and "tables_4500" for the three time windows. For example, to protect the identity and ensure the infant participants' privacy, filenames are anonymized and named as "VP20_output.tsv" etc.. In each file, the various columns such as "Left_before", "Left_After", "Right_Before", "Right_After", "Fixation_Direction", etc.., contain the measurements for each participant (VP). Within the same TSV document, starting from approximately line 26, another header line contains a new set of measurements titled: Before, After, Bef_Aft_Tar, Target, Dis (before), Dis (after), Bef_Aft_Dis, T-D, T-D(B-A).

7.3.1 Data Workflow Lifecycle

The research data workflow lifecycle diagram in Figure 7.1 explains the sequence of the research data processing and tasks for this project. The research project used Free & Open Source Software (FOSS), which increased the prospect of cross-platform availability of processing tools as Python programming language and visualization packages (like Pandas, Matplotlib) are freely available for multiple platforms.

The old data analysis scripts, written in Python version 2.x, were ported to version 3.6 for program maintenance due to end-of-life for Python version

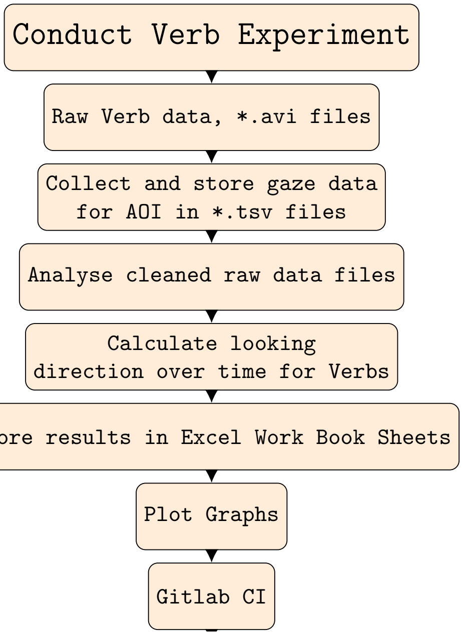
Conduct Verb Experiment
Raw Verb data, *.avi files
Collect and store gaze data for AOI in *.tsv files
Analyse cleaned raw data files
Calculate looking direction over time for Verbs
Store results in Excel Work Book Sheets
Plot Graphs
Gitlab CI

Figure 7.1: Data Workflow

2.x. Refactoring the old scripts from a complex mass of conditional loops, into a simplified modular callable program, was undertaken to ease program maintenance.

The main restructuring changes that were introduced are:

- Most conditional loops were refactored into modular methods. Breaking the code apart into more logical components creates semantic units that are clear and reusable.

- A dict to store the vertical area of interest for each avi file.

- Introduced a class that acts as a wrapper for the dict (which stores the result of one avi file (AOI)) and other methods that can handle the logical componentization.

- A sliding time window to compensate for missing data points - this short time window allows searching for the next fixation data. Three time windows: 3.5ms, 4.0ms and 4.5ms (experimentLength = 3500/4000/4500) were used.

Two Excel sheets stored the analysis results **results_simple_difference_score.xlsx** and **results_simple_target_distractor.xlsx** while the analysis data is stored in tab-separated value (TSV) files.

7.3.2 Reproducibility Results

Once the analysis script was ported to Python-3.6, it was possible to analyze the data and reproduce the results described in the paper as described in section 7.2.3 above. Figure 7.2 shows the percentage of looking times to the matching image for the different verbs, averaged across all subjects including all ages (both 9-month and 10-month olds). The verbs in question were: *'bauen'* (engl. build), *'fahren'* (engl. ride), *'lesen'* (engl. read), *'sitzen'* (engl. sit), *'anziehen'* (engl. dress), *'baden'* (engl. bathe), *essen* (engl. eat), *'schlafen'* (engl. sleep). Figure 7.3 shows the percentage of looking times to the matching image for the different verbs, averaged across all subjects for 9-month old infants only; Figure 7.4 shows the corresponding average looking times for 10-month old infants. Finally, Figure 7.5 shows the percentage of looking times averaged over all verbs and subjects, comparing the average for 9-month old infants vs. 10-month old infants. Within the 9-month-old infant group, on average, the infants spent 51.1% (SD = .056, MIN = 36.7%, MAX = 63%) of their looking time on the target object before the target word was spoken and 49.6% (SD = .063, MIN = 35.4%, MAX = 62%) of their looking time on the target object after the word had been spoken. Within the 10-month-old infant group, on average, these infants spent 43.4% (SD = .095, MIN = 25.9%, MAX = 60.1%) of their looking time on the target object before the target word was spoken and 49.6% (SD = .081, MIN = 31%, MAX = 64.1%) of their looking time on the target object after the word had been spoken. We could thus reproduce the main results of the original paper, showing a positive difference between percentage of looking time to target image after the corresponding verb was spoken minus the proportional looking time to the target before the verb was spoken for 10-month olds. For 9-month olds, this difference was on average negative, showing a lack of verb understanding.

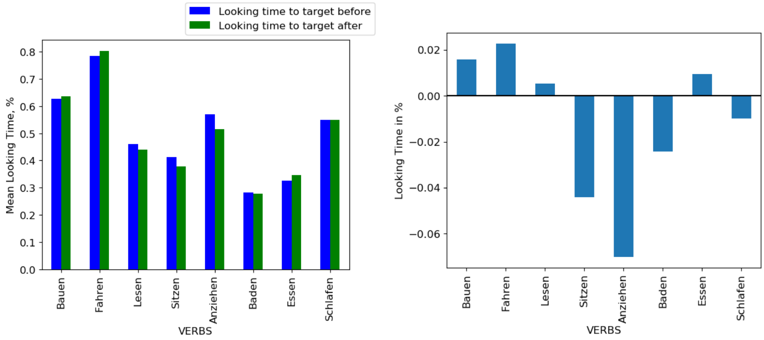

Figure 7.2: Looking times in percentage at matching image before and after utterance for all eight verbs averaged over all subjects (both 9-month and 10-month olds); Right: Difference in looking times for both 9 and 10-month olds

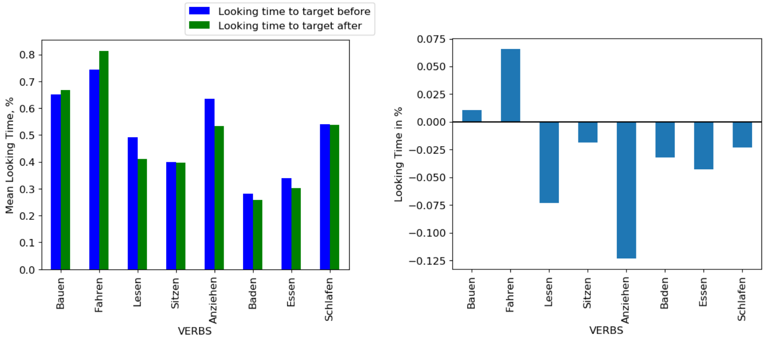

Figure 7.3: Left: Looking times in percentage at matching image before and after utterance for all eight verbs averaged over all subjects (9-month olds); Right: Difference in looking times (After-Before) for 9-month olds

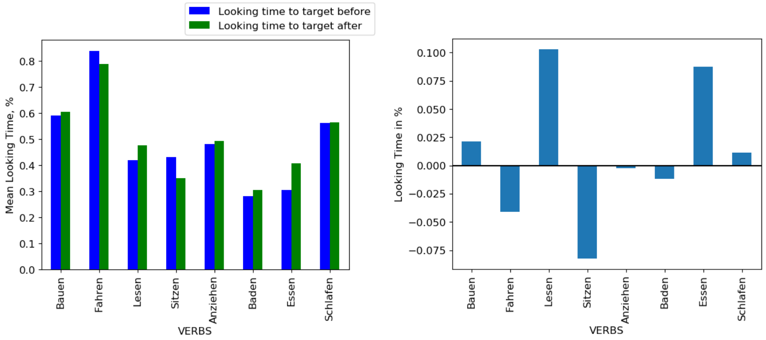

Figure 7.4: Left: Looking times in percentage at matching image before and after utterance for all eight verbs averaged over all subjects (10-month olds); Right: Difference in looking times (After-Before) for 10-month olds

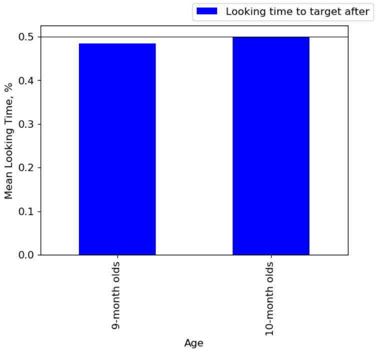

Figure 7.5: Average of percentages of looking times to target averaged over all verbs and subjects comparing 9-month and 10-month infants

7.4 Summary of computational reproduction experiment

In this reproducibility experiment, we were able to reproduce the main result of the study published by Nomikou et al. [1]. This was possible as the data and Pyhton scrips used to analyze the data were available. We engaged in this reproducibility experiment while the paper was in a second round of reviewing and considered the comments of the reviewers to adapt the Python program to allow for different time windows in the analysis. Overall, the results could be reproduced independently. The data and the Python script are available in a git repository for re-use and validation by third parties. This represents a case of *full analytical reproducibility.* Both the derived data capturing the looking times of each subject as well as the script for analysing the data are available in the Git repository, therefore supporting reproduction.

7.5 Conclusion

In this paper, we describe the successful reproduction of the computational analysis phase of a study investigating the early understanding of verbs by 9-month and 10-month-old infants. The reproduced study adopted a preferential looking time paradigm and conducted a so called paired-picture trial in which a verb under investigation was semantically associated to one of two pictures shown, the target picture, and another picture acting as a so called confounder. Using an eye tracker, the difference between proportion of looking times at the matching image before the verb was spoken compared to looking times after the verb was spoken was measured. As a result, the study showed positive differences for 10-month olds, which was operationalized as a measure of early understanding of the verbs. For 9-month olds, in contrast, the study was not able to reliably demonstrate verb understanding. The analytical pipeline that was used to generate results for publication was developed jointly between researchers of the Psycholinguistics group in Paderborn and researchers working in the Conquaire project. The derived data from the experiments (looking times) as well as the Python script are available for further re-use and correspond exactly to the version that was used to generate the published results. In this case, we thus have an example of *full analytical reproducibility*, with the analyses being repeatable by others as a result of the Conquaire project.

Acknowledgments

We would like to acknowledge the support of Lukas Biermann and Fabian Herrmann for helping with implementation of the scripts and data analysis.

References

[1] Iris Nomikou, Katharina J. Rohlfing, Philipp Cimiano, and Jean M. Mandler. Evidence for early comprehension of action verbs. *Language Learning and Development*, pages 64–74, 9 2018.

[2] Katherine Nelson. Concept, word, and sentence: Interrelations in acquisition and development. *Psychological review*, 81(4):267–285, 1974.

[3] Elika Bergelson and Daniel Swingley. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258, 2012.

[4] Jean M Mandler. On the spatial foundations of the conceptual system and its enrichment. *Cognitive science*, 36(3):421–451, 2012.

[5] Iris Nomikou, Malte Schilling, Vivien Heller, and Katharina J. Rohlfing. Language-at all times. *Interaction Studies*, 17(1):120–145, 2016.