Plan Overview

A Data Management Plan created using DMPonline

Title: The effect of grammatical complexity and verb frequency in the acquisition of English using the Sentence Repetition Test

Creator: Javier Aguado-Orea

Principal Investigator: Javier Aguado Orea

Contributor: Julian Pine

Affiliation: Sheffield Hallam University

Template: SHU Template

ORCID iD: 0000-0002-2311-5295

Project abstract:

This study has two main aims. First, to design and validate an experimental tool with typically developing children acquiring English as a first language. Second, to get a reliable estimation of three main factors: (1) the potential effect of the relative frequency of verbs, (2) the potential effect of verb inflection, and (3) the potential effect of sentence length. The main method adopted is the Sentence Repetition Test. This technique consists of asking children (and adults) to repeat sentences with varying degrees of complexity. It is possible to play with at least three key factors. First, sentence length. Second, the relative frequency of some elements, like verbs or arguments, in colloquial speech. And third, different grammatical features, like the complexity of the subject or the object of the sentence. The long-term focus is the analysis of the learning problems explaining Developmental Language Disorder (DLD) across different languages, but this specific study will only recruit typically developing (TD) children acquiring English as a first language.

ID: 110967

Start date: 01-01-2023

End date: 01-12-2024

Last modified: 09-11-2022

The effect of grammatical complexity and verb frequency in the acquisition of English using the Sentence Repetition Test

Data Collection

What data will you collect or create?

A final dataset will be produced with the combination of two sources: the parental report (CDI questionnaire) and an audio recording. Parents of participating children are initially asked to complete the CDI scale (perception of their child's language skills). This set of information includes the following personal details: gender (or undisclosed), date of birth (in months and years format), number of siblings (up to three), and any potential language-related concerns. In the event that parents report any concerns, they must fill out an open question. This response will be assessed for potential exclusion from the study since we are only interested in typical developing languages. Participants also must include their email address to be able to progress to the second stage of the study. All participants are assigned a random numerical code (RNC) between 1 and 999,999. For each participant, a csv file is generated with the scores of CDI using a combination of two numerical codes separated by an underscore line, and followed by the 'cdi' code (e.g., 0308_7623_cdi.csv): 1) Four numbers for the child age in years and months (e.g. 0308 for a 3 year 8 months old child), and 2) The last six numbers of the RNC.

Two separate documents are then generated to summarise the information for all participants: one with the email address and the RNC (named SRT_Eng_TD_RNC.csv), and one with the CDI scores and RNC (named SRT_Eng_TD_CDI.csv). These documents are stored in two separate folders on Q-Drive. On completion of the data collection, SRT_Eng_TD_RNC.csv is deleted.

Parents are contacted by email within the next three months, to arrange their participation in the second stage of the study. They are sent an email from a university account (@shu.ac.uk). If they do not reply to this initial email in one week, a second and final reminder is sent. If parents have not replied after a further week, all records corresponding to this participant are removed from the study. Their records are also removed if, after being contacted, parents do not want to continue taking part in the study. If they express their wish to take part, the Zoom interview is arranged.

The main source of data for the present study consists of a dataset of audio recordings and their corresponding transcriptions. All meetings organised over Zoom will be password protected and scheduled by invitation only. The recording will not be taken using Zoom's online system, but with a physical digital audio recorder attached to the receiving computer. Any personal information will be removed. During the recording, a parent will be present with the child, but remain silent during the main part of the study. No video or still images will be collected. Any accidental recording of video footage will be discarded immediately. The audio recordings will be stored on the institutional space (Q Drive). For each specific participant, two types of data will be generated. A digital audio file covering only the time comprehended between the first and final trials, and a plain text file with its corresponding transcription.

Audio files will be named in the following using the combination of two numerical codes separated by an underscore line, and followed by 'snd' (e.g., 0308_7623_snd.aif) 1) Four numbers for the child age in years and months (e.g. 0308 for a 3 year 8 months old child), and 2) The last six numbers of the RNC.

All speech data will be transcribed using the CHILDES system (MacWhinney, 2000) and named following the same criteria (age_RNC) followed by 'txt'. It is customary to use the suffix .cha for transcripts adopting the CHILDES system (e.g., 0308_7623_txt.cha)

All files will be associated with the RNC assigned to every participant, so they can be matched with the CDI scores and information about gender and age. Within 72 hours, parents will receive a further email message acknowledging their participation and reminding them that they can remove their participation if they wish so, by sending a message within the following seven days. If no message has been received, the audio file and transcription, matched with the CDI scores, will be uploaded to the OSF repository, and no further disposal of data is contemplated. The file with the email addresses will be deleted at the completion of data collection (i.e., when the expected number of participants has been reached).

In sum, participants are offered the disposal of data at two points in time: after completing the CDI questionnaire, and after taking part in the audio recording. If they do not request data removal, completely anonymous versions of the sound recordings, transcriptions, and CDI scores will be made available using the Open Science Framework (OSF) under the main researcher's account. For each participant there will be three files: 1) the raw CDI scores, 2) the sound file, and 3) the transcription. For all participants, there will be one file with the global CDI score, persona information including gender and age, and the result of the analyses of the transcribed files.

How will the data be collected or created?

1. Design.

This is a 2x3 experimental design

The first factor (within groups) is person agreement (singular vs plural).

The second factor (within) is sentence length with three values: short subject and object (SVO), long subject (SsVO) and long object (SVOo).

The relative frequency of verbs in the corpus is entered as a continuous factor.

1. Participants

Three groups of participants will be recruited for the study. A small sample of adults to validate the stimuli, and a larger number of

children to get the main results. And finally, one of the parents of these children, to complete an online questionnaire.
This study aims to recruit in total around 92 people distributed in the following way: a) 12 English-monolingual adults, b) 40 English-monolingual child participants per sample of language, and c) one of their parents per sample of language too. A similar number of participants from both genders will be recruited. Children will be between 2 and 6 years old.

1. Materials and procedure

The study is run online, using a combination of tools structured in three stages.
Stage 1. Stimuli validation with adult participants
Stage 1a) After reading the conditions of participation and being explained the purpose of the study, adults are asked to complete a brief (about 3 minutes long) online questionnaire to report any potential history of language-related issues. Then, they proceed to stage 1b
Stage 1b) An online interview, using Zoom, is arranged with adult participants. During this interview, the screen of the researcher is shared, and a series of videos including the target sentences (described below in further detail) played through the speakers of the participant device. Participants are asked to repeat a series of sentences after having heard them spoken to them via voice recording. The sound is recorded with a conventional digital recorder and stored in a password-protected online space in Sheffield Hallam University (q drive).
Stage 2. Parental reports
Parents interested in the study are asked to fill an online questionnaire, where they can read the main purpose of the study and accept the ethical implications if they wish. Then, one of the parents is asked to complete a small questionnaire as a preliminary assessment of the child's language skills. This questionnaire is based on a publicly available version of the Communicative Development Inventories (CDI) developed by Hamilton et al., (2000). The CDI scale consists of a list of words that parents have to mark as either comprehended or comprehended and produced by the target children. Completion takes about three minutes.
Stage 3. Child participation
An online interview, using Zoom, is arranged with parents. During this interview, the screen of the researcher is shared, and a series of videos including the target sentences (described below in further detail) played through the speakers of the participant device. Children are asked to repeat a series of sentences after having heard them spoken to them via voice recording. The sound is recorded with a conventional digital recorder and stored in a password-protected online space in Sheffield Hallam University (Q-drive).
Sentences for the repetition study
Both participating adults and children are first presented with three relatively simple sentences for an initial round of practice:

- The cat is big
- This is a red car
- The boy with a hat

Then, the actual trials are presented in a pseudo-randomised order (a fully randomised order is not used to avoid harder sentences during the first 5 trials). They consist in 24 sentences with three levels of varying difficulty. Eight shorter sentences (e.g., the boy jumps a wall). And 16 longer sentences with either longer subjects (e.g., the boy with a teddy blows a balloon) or longer predicates (e.g., the girls keep a car with a key). Subjects are always either "the boy/s" or "the girl/s" (gender and number is fully counterbalanced). Participants watch a cartoon that matches the meaning of the stimulus as the sentences are played to increase their interest in the study.
Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a British communicative development inventory. *Journal of Child Language*, *27*(3), 689–705. https://doi.org/10.1017/S0305000900004414

Documentation and metadata

What documentation and metadata will accompany the data?

A detailed data map will be created for each dataset.
For the CDI raw data there will be four columns [word (the target word), u (understands), us (understand and uses), n (don't)] and 41 rows (one per item).
Transcriptions follow the CHILDES standard.
The main data file for all analyses is SRT_Eng_TD_CDI.csv
It includes the global CDI score per participant and the resulting analyses of the Sentence Repetition Test. These scores will be included in the data map file:
SRT_Eng_TD_CDI_datamap.txt

Ethics and Legal Compliance

How will you manage any ethical issues?

The project will be reviewed by the SHU ethics committee. Information sheets and consent forms will be used to ensure that informed consent is gained that allows for the preservation and sharing of the anonymised data.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

The principal investigator (Javier Aguado-Orea) will own primary data collected. Data will be made available under the CC-BY licence (see https://creativecommons.org/licenses/by/3.0/), allowing for immediate downloading and re-use, subject to appropriate acknowledgement.

Storage and Backup

How will the data be stored and backed up during the research?

We will use the University's networked Research Store for all master copies of our data. Data is backed up automatically on a daily basis, and can be fully recovered in the case of accidents. All backups are securely kept on two remote locations for a period of 90 days. Access to all folders is restricted to researchers, students and external partners working on the project. At project close down relevant data relating to this project will be securely archived, and all data will be deleted from the Research Store.

How will you manage access and security?

All data will be temporarily stored in Q-Drive, using password protection, before they are made publicly available.

Selection and Preservation

What data are of long-term value and should be retained, shared, and / or preserved?

the audio file and transcription, matched with the CDI scores, will be uploaded to Sheffield Hallam University Research Data Archive (SHURDA) and the OSF repository, and no further disposal of data is contemplated. All data (raw and analyzed) will be deposited in SHURDA at the end of the research project. The data will be retained in the archive for a period of 10 years since the last time any third party has requested access to the data. When depositing the data, no further changes to data formatting will be required as all necessary actions will have been conducted as the research progresses.

What is the long-term preservation plan for the dataset?

All 'raw' data (with appropriate documentation), and the analyzed data will be deposited with the Sheffield Hallam Research Data Archive (SHURDA) and made available after the embargo period has expired. This approach to open access will ensure the legacy of the project by enabling follow-up and/or longitudinal studies to be compared with these initial raw data sets.

Data Sharing

How will you share the data?

Suitably prepared transcripts and supporting documents such as codebooks will be deposited with the Sheffield Hallam University Research Data Archive (SHURDA) and OSF and made immediately available to all registered users under a CC-BY licence.

Are any restrictions on data sharing required?

We will deposit and share our data at the end of the project without any delay. Any research outputs that are published will contain a statement that refers to the underlying datasets and how these datasets can be accessed; any restrictions to access will be outlined and justified in this statement. The raw anonymized data and the data collection methodologies will be made available on a Creative Commons with Attribution (CC-BY) or equivalent license.

Responsibility and Resources

Who will be responsible for data management?

The Principal Investigator (Javier Aguado-Orea) is responsible for each data management activity.

What resources will you require to deliver your plan?

Transcription of audio files require the use of CLAN software, distributed free via TalkBank.org.
OSF is free of charge.