NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

FACULTY OF HUMANITIES
SCHOOL OF LINGUISTICS

PROJECT FOR LINGUISTIC DATA: QUANTITATIVE ANALYSIS AND VISUALISATION COURSE

THE INTERRELATIONSHIPS BETWEEN PART OF SPEECH OF THE WORD AND WORD ACQUISITION OF INFANTS AND TODDLERS

Vladislava Smirnova, 191

Advisors:
O.N. Lyashevskaya, Candidate of Sciences
I.V. Schurov, Candidate of Sciences

MOSCOW
2020

Contents

Hypothesis

This study has 2 hypotheses to prove. First is that children memorize words of closed class after acquisition of particular amount of words from opened class. Children assimilate common nouns earlier than verbs and adjectives, when closed class items are scarcer in their speech. Other hypothesis of the study is the suggestion that frequency of words in adults' speech and the vocabulary size of children have strong correlation. According to the previous research on this topic (Hansen, 2016), we could also expect that adult frequency would be a significant predictor of age of acquisition within some classes of words. The author could possibly add supplementary task to this paper and construct a prediction model for part of speech learning.

Research design

Data

The dataset includes two files. The first table consists of information about 732 Norwegian words with translation to English, average children' age of acquisition, average vocabulary size of children, lexical categories and frequencies of words in the Norwegian Web Corpus and child-directed speech (abbreviated as "CDS".). The second table includes measures of how frequently each word is used in Norwegian both on the internet (as observed in the Norwegian Web as Corpus dataset) and when an adult is talking to a child. The data was collected by Pernille Hansen, PhD in linguistics at MultiLing, University of Oslo. Most relevant factors, which would be studied in this paper: word class, frequency of words in adults' speech and vocabulary size of children.

Statistically formulated hypotheses

H0: mean ages of acquisition of different classes are similar.

H1: mean ages of acquisition of different classes have variations

H0: there is no correlation between frequencies of words in parents' speech and the vocabulary size of children, the true correlation coefficient R is 0.

H1: there is correlation between frequencies of words in parents' speech and the vocabulary size, the true correlation coefficient R is not 0.

H0: there is no correlation between frequencies of words in parents' speech and age of words acquisition, the true correlation coefficient R is 0.

H1: there is correlation between frequencies of words in parents' speech and age of words acquisition, the true correlation coefficient R is not 0.

Model

Statistical analyses of this paper would be performed with R and additional package tidyverse. The author would use ANOVA test for understanding learning of different parts of speech and classes of words. Correlation would be implemented as a method of understanding connection between the frequencies of words in adults' speech, the vocabulary size of children and age of acquisition within some lexical categories. For the additional task of predicting which class would be learnt by children of each age Random Forest classifier could be possibly used.

Data collection method

The main data consists of 732 words included in the Norwegian adaptation of MacArthur-Bates Communicative Development Inventories form, a parental questionnaire demonstrated to give a valid and reliable measure of early lexical development. The Norwegian CDI norms consist of data from 6500 monolingual Norwegian children. Important thing about CDI: their norms are typically cross-sectional, and can thus not be used to determine when each child acquired each word, only which words she or he currently produces (or understands). This limitation was circumvented by calculating the age in months where at least 50% of the children are reported to produce each word. It is commonly referred to the data as a word's age of acquisition.