Detection of a Specific Language Impairment (SLI)   
in children's speech

Project Proposal

Ostyakova Lidiya

Specific Language Impairment is a condition that affects roughly 7% of 5-year old children. It is characterized by a lack of language ability in comparison to your peers but with no obvious mental or physical disability. Children with SLI are more likely to speak in short sentences and often have difficulty with the morphemes -ed, -s, be, and do. Moreover, children with SLI are more likely to need support from the examiner. The main goal of this project is to research how SLI diagnosis correlates with certain speech characteristics, gender, and age of a child.

**Hypotheses**

The first hypothesis of this study is that the diagnosis depends on certain characteristics of children’s speech:

* *Number of Repetitions*
* *Number of Fillers*
* *Total Number of Words*
* *Number of Do's*
* *Total Number of Words spoken by the examiner*
* *Total Number of Sentences*

Moreover, it’s important to find out the correlation between SLI and non-linguistic characteristics of a speaker:

* Gender
* Age
* Country (variable *Corpus*)

**Data**

Data on this topic was taken from **Kaggle** [https://www.kaggle.com/dgokeeffe/specific-language-impairment#all\_data\_R.csv]. The dataset consists of narratives from a child attempting to complete a wordless picture task. It includes information from three different corpora obtained via the CHILDES project: *Conti-Ramsden 4*, *ENNI*, *Gillam*. The first corpus consists of samples from British adolescents, the second one is from Canadian children aged 4 to 9, and the third one is from U.S. children aged 4 to 12. There are 1163 samples overall.

**Data Collection Method**

For this project, the data from Kaggle will be reduced and not all the variables will be used. However, if it is necessary, additional information from the Internet will be taken into consideration for an interpretation of the chosen characteristics. Since the data include three different corpora, the collection method for all of them will be described.

***Conti-Ramsden 4***

The corpus contains transcripts of a storytelling task based on Mayer’s wordless picture book “Frog, Where Are You”. The children first viewed the picture book in their own time before being prompted to retell the story in the past tense. If the children started telling the story in the present tense the interviewer would prompt them with the phrase “What happened next?” in order to attempt to revert them back to the past tense. If they failed to start to retell the story in the past tense after two prompts no further prompts were made.

## ***ENNI***

Each child was presented with two wordless picture stories with one more complicated than the other. The examiner held the book and turned the page after the child appeared to be finished telling the story for a particular picture.

## ***Gillam***

The Gillam dataset is based on another tool for narrative assessment known as “The Test of Narrative Language (TNL). The TNL consists of four storytelling tasks, the first is a recall of a script based story, the rest being wordless picture books. The TNL appears to be an intermediary in difficulty compared to ENNI.

**Research design**

**Variables**

There are more than 20 variables presented in the dataset, however, only the following ones will be taken into consideration:

|  |  |
| --- | --- |
| **Variable** | **Type** |
| **Group** (0 for typically developing children, 1 for language-impaired ) | binary  **(TARGET)** |
| Number of Repetitions | numerical |
| Number of Fillers | numerical |
| Total Number of Words | numerical |
| Number of Do's | numerical |
| Total Number of Words spoken by the examiner | numerical |
| Total Number of Sentences | numerical |
| Country | categorical |
| Gender | binary |
| Age | numerical |

The information about the country is absent in the data and will be taken from the variable *Corpus* because it is necessary for checking one of the hypotheses and, therefore, should be more demonstrative in the dataset. This variable is not considered in the existing studies on the topic of SLI.

The variable *Gender* is categorical on Kaggle, but it will be changed to a binary variable for convenience. The variables *Total Number of Sentences* and *Total Number of Words* will be compared in order to find out the best way of comparing such a speech characteristic. They are likely to be presented as a new variable in the data.

Statistical analysis will be performed with R and additional packages such as tidyverse, ggplot2, vcd, ggfortify, ca, stats, party. Working with the data will be performed with the use of pandas and python3.

The first stage of the project is a visualization and searching for a balance between the corpora in the dataset. Then various statistical tests will be conducted to assess the significance of some of the variables and test the hypotheses:

|  |  |  |
| --- | --- | --- |
| Null hypothesis | Alternative hypothesis | Test |
| There is no dependency between a group and characteristics of children’s speech (for each of them) | There is a dependency between a group and characteristics of children’s speech (for each of them) | T-test |
| There is no dependency between a group and the age of a speaker | There is a dependency between a group and the age of a speaker | T-test |
| There is no dependency between a group and the gender of a speaker | There is a dependency between a group and the gender of a speaker | Chi-squared test |
| There is no dependency between a group and a country of a speaker | There is a dependency between a group and a country of a speaker | Cramer’s V |

As the next step, logistic regression will be used to analyze the significance of each linguistic and non-linguistic parameter. Moreover, it is planned to study odds concerning only non-linguistic parameters. Finally, there will be a conclusion and an interpretation of the results.