R(ead)(B)abbling

TA: Lucas Reisch

Team: Puran Zhang (pz75), Jiangjie Man(jm2559), Pengcheng Zhang(pz84), Xiuyan Xin(xx254)

1. Work Distribution

Each team member is scheduled to share the work equally. Because of the nature of our project, members have participated in all tasks together. All of us have brainstormed the project idea, coded and contributed to aesthetic quality. In terms of project management, we took turns to do the planning and control the risks. Specifically (since report 3):

Puran Zhang (pz75):

- Wrote out the storyline;
- Built regression model using *regression.min.js*

Jiangjie Man(jm2559):

- Preprocessed data of words in the story text into different P-O-S tag
- Implemented words randomly flying when user scroll the page

Pengcheng Zhang(pz84):

- Implemented scroller using *ScrollMagic.js*
- Performed dynamic effect for generating the regression lines when user scroll the page

Xiuyan Xin(xx254)

- Processed the data of baby`s comprehension level and word classification based on category

- Completed two graphs: baby comprehension heap map and word classification bubbles

2. Project Description

Our project focuses on presenting progressing process of babies perceiving English vocabulary and providing analysis in related aspects such as the regression line of perceiving by word-of-speech, the category of the words, the relation between comprehension and mother`s education level and etc. We brainstormed a lot of ideas, and in pursue of achieving a smoother and aesthetically pleasing experience, we enable words animation by scrolling which can go back and forth seamlessly. And the form of scrolling is creative in the aspect of giving intuitive impression of the baby`s growth.

To better use the dataset, even one dataset may be used in different ways which will be shown in the following part.

2.1 Description of the Data:

Our data comes from *wordbank* and contains 3 files, namely By-Child Summary Data, By-Word Summary Data and Full Child-by-Word Data. The links is as below:

http://wordbank.stanford.edu/analyses?name=admin_data

Because of the complexity of data, we decided to limit the scope to just the English language, and babies from 16 to 30 months old. Specifically there are total of 3 raw data files:

- By-Child Summary Data
    - Java is used to group the rows by the mother education level(mom_ed) and baby`s age. And take the average amount of comprehension level for the visualization of heatmap.
- By-Word Summary Data

    - We applied basic NLP analysis and used nltk package from python to tag out the P-O-S tag of words in our data.
    - From there, we used javascript and d3 to nest the pos tags and used regression.js (external library) to apply cubic polynomial regression on the relationship between age vs learning rate within different pos tags.
    - Java is used to group the words by category for the visualization of word bubbles.
- Full Child-by-Word Data

For the story.txt, we used common_word.py to sort out overlapped words in the story and in our dataset. In this way, we find the most matching one among 5 files and used that for this visualization.

2.2 Technical Guide:

For the "word-flying" feature, words appears cumulatively according to specific month the user scrolls to, for example, if the baby are 16 month right now, we look for the probability at 16 month in the regression model for each POS tag. Then we generate random uniform value between 0 and 1 for each word in the story, if the value is lower than the learning probability at the given month, we assume the word has been learned by the child and thus make the word "fly" to where it should be in the text.

3. Description of Mapping from Data to Visual Elements

We use scaleOrdinal() and scaleQuantile() to scale features into different colors such as part-of-speech, the comprehension level. For the visual design part, we use the *** font to create the sense of vigour and new-born. We also split the view into halves and place descriptive and displaying part separately. We chose similar color for the words, the regression line and the squares in the heatmap to achieve consistency. And we chose soft colors for the background of displaying which would make people feel comfortable.

4. The story

Our data visualization tells a story of how the babies from 16 to 30 months learn the vocabulary as well as the world. For the data analysis part, we found that the rate of perceiving is quite similar between different word-of-speech, however noun is the most quickly perceived when comparing to others, and the function words come along and listed number two. Predicates and verbs are the slowest one for babies to learn.

Secondly, when it comes to categorizing the words more concretely, we found that the action words, food and drinks, descriptive words and household are listed top four, which is quite understandable because they are closely related to the environment for babies aged between 16 to 30 months.

Last but not least, when we look into mother`s education we found the comprehension is in proportion with the education, that is, the higher the mother achieves academically, the better the babies comprehend vocabulary. This may be interpreted in the way that the mothers with higher academic level are more likely to consider early education important, thus pays more attention to developing children`s comprehension.