EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

PHILOSOPHISCHE
FAKULTÄT
Seminar für Sprachwissenschaft

Data Science for Linguists
Summer 2025
Johannes Dellert

Assignment 02: Exploring Vocabulary Acquisition Data with NumPy and Seaborn

handed out: 30 April, 14:00
to be submitted by: 7 May, 12:00

In this sheet, you will explore data about the ages at which English-speaking children acquire words. The data were collected by [Kuperman et al., 2012] through web-based crowdsourcing for their study of the connection between age of acquisition and lexical decision speed. The data were released in the shape of the file `Kuperman-BRM-data-2012.csv`, which is available as part of the materials for this assignment.

Task 1: Loading a Dataset into NumPy

Create a new Jupyter notebook and import `numpy`. Load the contents of the second, fourth, and sixth column of the CSV file into a NumPy array, not forgetting to skip the header (you cannot work with mixed data yet). Each row now contains age-of-acquisition data for one English word. The first column will be the number of participants with known age of acquisition for the word, the second column the average age of acquisition, and the third column will be a frequency count in the SUBTLEX-US corpus. Use standard functions to explore the NumPy array a little in order to make sure that the data make sense, and you imported everything correctly. Use filtering and `numpy.isnan` in case you are running into trouble with NaN values. Do the summary statistics (averages, maxima, etc.) yield any surprises?

Task 2: Visualising Word Frequencies

We start our exploration of the data with a simple analysis of the word frequencies contained in the dataset.
a) Normalise the last column of your array in place in order to turn raw frequencies into relative word frequencies. (Hint: use universal sum, and broadcasting for division by the result)
b) Use Seaborn to plot a smooth density estimate of the relative word frequencies. Create a second plot based on the logarithms of the raw frequencies (a universal function is all you need!). What do you notice? Do you think something might be wrong with the data?

Task 3: Vocabulary Sizes at Different Ages

Answer the following questions based on your NumPy array, explaining what you did in Markdown cells:
a) How many English words do children acquire in the first four years of their lives?
b) What percentage of tokens in the corpus would a nine-year-old child understand?
(Hint: filter by age of acquisition, then add up)
c) How many of the top-2500 most frequent words (B1 level) would we expect a five-year old native speaker to know? (Hint: use `argsort()` to sort by last column, select last 2500 entries, then filter by age)

Task 4: Analysing and Visualising Vocabulary Acquisition

After investigating the development of vocabulary sizes, we now shift the perspective towards analysing the dynamics of vocabulary acquisition.
a) By how many points does the percentage of understood tokens grow in each year? What ages are the peak years of increase in comprehension? (Hint: list comprehension, `np.diff()`)
b) By how many words does the lexicon of a child grow in every year of its life? Is there a peak and a tendency? (Hint: `np.where()` with two conditions in list comprehension)
c) Create graphs of both the absolute and relative growth of the child's vocabulary at each age, and use the graphs to validate your results from a) and b).

Task 5: Using Seaborn to Investigate a Suspected Frequency Effect

We would expect that words which are encountered more often are acquired more quickly. Later in the course, we would now start to build a model in order to quantify the strength of this frequency effect. Instead, we will just discover a way of computing rank correlations in pure NumPy, and practice some more Seaborn, in order to gain an impression of the strength of the effect.

a) Compute the Spearman rank correlation between age of acquistion and negative log frequency. This can be done through multiple application of the `argsort()` method on arguments to the `np.corrcoef()` function which computes the Pearson correlation coefficient. You might perceive this as not entirely straightforward - can you explain what is happening in each step of your solution?

b) Create a joint plot of the logarithm of the ages of acquisition against negative log frequency. If you use the variant which includes a linear regression fit, can you read and interpret the results of that regression? Does it fit with what you were able to conclude from the rank correlation?

References

[Kuperman et al., 2012] Kuperman, V., Stadthagen-Gonzalez, H., and Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44:978–990.