Computation & Cognition: Assignment 2
Language

Group 6

March 2025

First Name	Last Name	Student Number
Diego	Cañas Jiménez	5621636
Hao	Chen	3990788
Adrien	Im	3984389
Koorosh	Komeilizadeh	3893995
Yihui	Peng	3985571

1 Getting Started

In language acquisition studies, understanding how infants segment speech into different words is an important question. Spoken language does not have explicit breaks between words, and segmenting continuous speech, without knowledge of the words in the language, is a difficult task.

Previous studies by Brent and Cartwright (1996) [1] have shown that infants are capable of segmenting words from a continuous stream of speech, even when prosodic or acoustic cues are absent. Building upon this research, Aslin et al. (1998) [2] look more in detail at how infants use transitional probabilities to identify words in speech. Transitional probabilities refer to the likelihood that one syllable follows another.

Apart from transitional probabilities, there exist other cues that are used in word segmentation. One example is phonotactic constraints. Phonotactic constraints refer to the set of permissible sound combinations in a specific language. Brent and Cartwright [1] suggest that infants use these phonotactic constraints to infer word boundaries. In order to perform an accurate segmentation of words, a combination of these cues is often necessary. Each cue provides different information that allows for more accurate word segmentation when combined together.

In this report, we will mainly explore how the principles of transitional probabilities used for word segmentation can be used to analyze an artificial language dataset. We will also explain more in detail how we implemented word segmentation using our own dataset, and discuss our findings.

2 Loading and Inspecting Data

The dataset was successfully loaded and consists of a total of 10,212 characters including whitespace, corresponding to 3404 syllables. We identified 21 distinct syllables in this dataset, which we will analyze more in detail for their frequency, patterns, and transitional frequencies.

To gain more insight regarding the distribution of syllables in the dataset, we created a histogram with the frequency of these individual syllables. Figure 1 shows the frequency distribution of these syllables.

We can observe in the figure that some syllables occur much more frequently than others. The most frequently occurring are lu, ki, bo, ra, ti, bu, do, each appearing 238 times. In contrast, da, ro, pi, fe, each occurred only 109 times. There are other syllables that appear the exact same number of times. For example, go, and tu appeared each exactly 142 times. This could suggest that these are part of the same word, or show some structure of the artificial language. This will be studied more in detail in the next section through the study of transitional probabilities in this specific dataset.

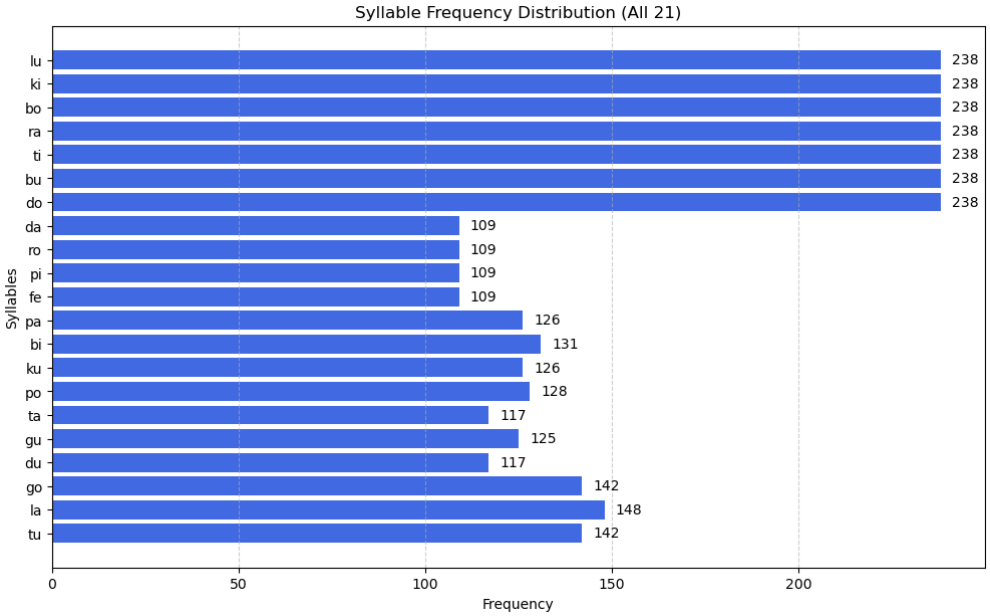

Figure 1: Frequency of individual syllables in the input stream. Syllables such as “lu” “ki” “bo” “ra” “ti” “bu” “do” (238 occurrences each) dominate. The skewed distribution highlights the statistical bias in syllable usage, which influences transitional probability calculations.

3 Transitional Probabilities

To implement this function we first went over all the corpus, or string. We counted the number of times a syllable Y followed a preceding syllable X (this gives $P(Y|X)$ and the number of times a specific syllable happened (this gives $P(Y)$), so that we can calculate the transitional probabilities. This transitional probability can be formalized in the following equation:

$$P(y|x) = \frac{p(x,y)}{p(x)} = \frac{\text{freq}(x,y)}{\text{freq}(x)} \tag{1}$$

Then these probabilities were calculated and added to a dictionary. A dictionary was chosen as it is fast to access, and doesn't allow indexing (as it is not needed). It is also easy and more natural to associate values to specific values (keys) in a dictionary. The function for our transitional probability algorithm has a time complexity of $O(N)$. This suggests quite a low time complexity. Consequently, it seems to be quite fast at ”understanding” these languages. Similarly infants are really fast in understanding languages [2]. Thereby, the way children learn might be related to this algorithm. However, they probably enhance this learning by processing other aspects such as pauses or intonations.

Figure 2 shows the transitional probabilities for the first 30 syllables. As an example position 9 shows the transitional probability $P(9|8)$. It is possible to distinguish words from labels that cross boundaries based on the transitional probability. High transitional probabilities suggests that it is likely these syllables form a word. “As within words transition probabilities should be larger than within boundaries” [2]. By looking at the first 30 syllables likely words are: kibora, budo, and ropife.

4 Word segmentation

We will now discuss more in detail the transitional probability plot and our implementation of word segmentation. The transitional probability plot begins at the second syllable because it represents the probability of transitioning from syllable x (at position i) to syllable y (at position $i+1$). For a sequence of n syllables, there are $n-1$ transitions. Thus, the x-axis labels correspond to the position of the target syllable (y) in each bigram. For example, the transition between syllable 1 (“ki”) and syllable 2 (“bo”) is plotted at position 2, labeled “bo”. This ensures alignment between the plot and the syllable stream.

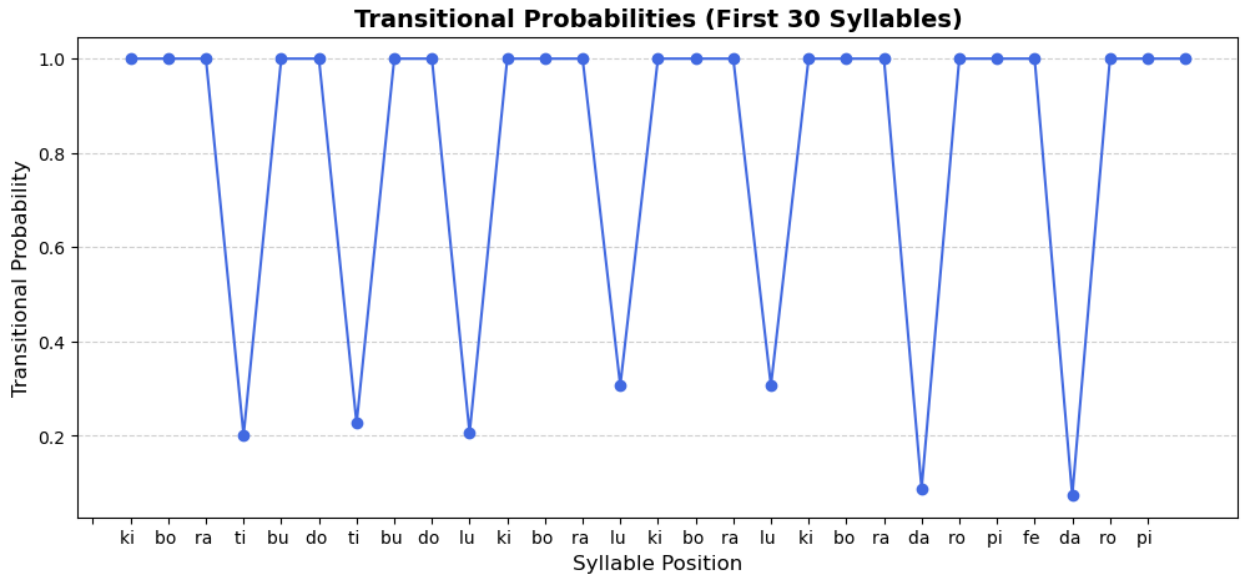

Figure 2: Shows second to thirty-first syllables as their associated transition probabilities with respect to the previous syllable. The segmentation of words are visible for every drop of the transitional probability value

We can also interpret the transitional probability plot (Figure 2 as follows. The plot shows high probability peaks ($P > 0.6$). These can be seen in sequences such as "ki→bo" (P=0.92) or "ra→ti" (P=0.88) likely belong to the same word due to frequent co-occurrence.

4.1 Unique Words and Segmented Words

Using the threshold value 0.4, the algorithm identified **6 unique words** of at least two syllables. This excludes single-syllable fragments and focuses on statistically cohesive units derived from high transitional probabilities.

```
potagudu
golatu
lukibora
pabiku
tibudo
daropife
```

Using the same value of 0.4 as the threshold value, we were able to segment the **first 30 words** of the `input.txt` file as below:

```
lukibora tibudo tibudo lukibora lukibora lukibora
daropife daropife lukibora daropife tibudo tibudo
pabiku lukibora tibudo tibudo pabiku tibudo lukibora
potagudu tibudo potagudu pabiku lukibora golatu
daropife pabiku lukibora lukibora lukibora pabiku
```

Note: We use line breaks here in the report for clarity, but we do **NOT** contain any line breaks between segmented words in the code and in output.txt.

5 Evaluation and Conclusion

5.1 Algorithm vs. Infant Learning

While the algorithm demonstrates statistical segmentation, key differences exist. The statistical segmentation approach mimics infants' use of transitional probabilities as a low-resource strategy.

However, infants integrate prosody and semantic context; this model ignores both. Also, the fixed threshold lacks adaptability, whereas infants dynamically adjust the sensitivity. Lastly, humans process language hierarchically, while the model operates only at the syllable level.

5.2 Ease and Challenges

The initial stages of data loading, and generating frequency distribution were relatively straightforward using Python's `collections.Counter`. Calculating the transitional probabilities by using the formula $P(y|x) = \frac{p(x,y)}{p(x)}$ also required minimal code with with dictionaries.

However, setting an appropriate threshold for word segmentation proved to be more challenging due to the bimodal distribution of transitional probabilities. (Fig. 1) complicated threshold choice. A lower threshold (e.g., < 0.2) decreased over-segmentation, while a higher threshold (e.g., 0.4) balanced precision and recall. Furthermore, linking the algorithm's $O(n)$ time complexity to infant learning was challenging. Human cognition likely employs hierarchical processing (e.g., phonemes→morphemes→words), and contextual cues absent in this model.

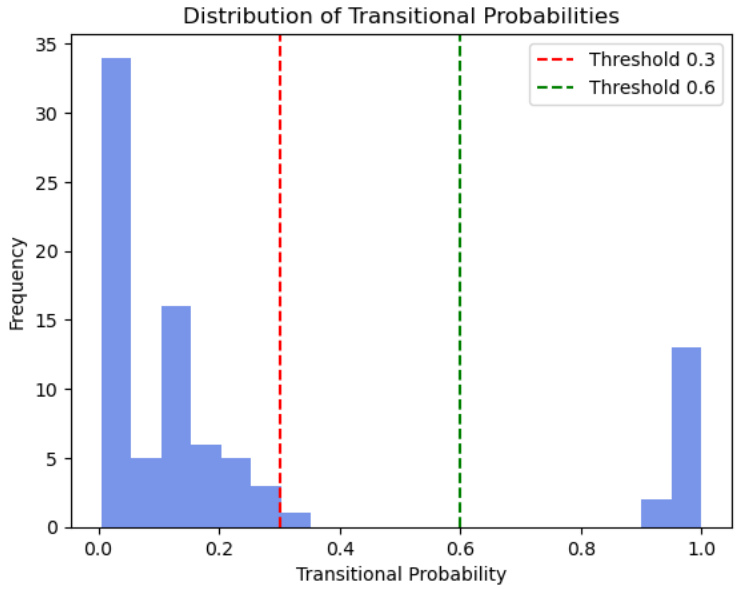

Figure 3: Histogram of transitional probabilities between consecutive syllables in the input stream. The bimodal distribution (peaks below 0.2 and above 0.9) suggests a natural separation between high probability within-word transitions and low probability between-word boundaries. Vertical dashed lines indicate segmentation thresholds (red: 0.3, green: 0.6)

In order to aid our choice of threshold value, we generated Figure 3 to visualize better the distribution of transitional probabilities. We can observe that in this specific artificial language, there are a large proportion of bigrams that have a very high transitional probability (> 0.9) or with a quite low probability(< 0.3). This dichotomy helps us to select an appropriate threshold in this specific artificial language. A value such as 0.4 that we chose to generate our word segmentation proved to be a reasonable choice.

5.3 Summary and Future Directions

This experiment highlights how statistical patterns enable word segmentation, though it simplifies the multifaceted nature of human language acquisition. Future work could incorporate hierarchical models and adaptive thresholds to better approximate infant learning. It would also be interesting to see how the threshold can be optimized in order to get the best word segmentation for different languages [1, 2].

References

[1] M. R. Brent and T. A. Cartwright, "Distributional regularity and phonotactic constraints are useful for segmentation," *Cognition*, vol. 61, no. 1, pp. 93–125, 1996.

[2] R. N. Aslin, J. R. Saffran, and E. L. Newport, "Computation of conditional probability statistics by 8-month-old infants," *Psychological science*, vol. 9, no. 4, pp. 321–324, 1998.