Determining Core Words for Mandarin-Speaking Children

Written by: Morgan Wood

Client: Hsiao-Ting Su

15 November, 2023

Abstract

This report aims to help provide guidelines for selecting core words for Mandarin-speaking children between the ages of 4 and 6. Using data provided, this analysis

1. Provides recommendations for choosing the thresholds of commonality and frequency for selecting core words,
2. Provides a list of core words for children between the ages of 4 and 6 as a whole, as well as lists of core words for each age group,
3. Provides graphical depictions of the relationship between commonality and frequency of core words, and
4. Tests to see if the commonality of words are the same between different age groups; this analysis provides evidence that the core words between 4-year-olds and 6-year-olds are not the same.

1 Overview

Determining core words for a set of individuals is an important first step in developing Aided Augmentative and Alternative Communication devices (Aided AAC) tailored for those individuals. Aided AAC are either electronic or non-electronic devices used as an alternative to traditional verbal communication. Examples of Aided AAC are communication boards or speech generating devices.

The goal of this report is to display a set of potential core words for Mandarin speaking children between the ages of 4 and 6 to aid in the creation of Aided AAC tailored to their needs.

To determine core words, multiple text files containing transcribed stories told by children are deconstructed into a list of words. In our setting, a word can by a single Chinese character or multiple Chinese characters.

For this report, I consider four different groups of children each with n children within the group and a total number of w words spoken. The groups considered are listed below.

- All Children (Ages 4 through 6) (n = 127, w = 58773)
- Children of Age 4 (n = 40, w = 17893)
- Children of Age 5 (n = 50, w = 22787)
- Children of Age 6 (n = 37, w = 18093)

For each group of children, two statistics are computed for each distinct word spoken. First, if a word is spoken by $c \cdot 100\%$ of children in the group, the word is assigned a commonality c. The frequency that a word is spoken is also looked at. From this we can obtain a relative frequency r which will be the number of times per 1000 words the word is spoken. The total frequency of each word is also reported.

Within this report, a word is considered core if at least 30% of children spoke the word and the word is spoken with a relative frequency of at least 0.5 times per 1000 words.

2 Choice of Commonality and Frequency Threshold

I begin this report with a comment on the choice of commonality and relative frequency threshold for declaring a core word.

While commonality and frequency represent two different statistics of the data, they are related. For example, if a word has commonality 1, then this requires that each individual in the study spoke the word. This implies that the word is spoken with a cumulative frequency at least equal to the number of individuals n. More generally, if a word has a commonality of c, the relative frequency must be at least

$$1000 \cdot \frac{c \cdot n}{w}$$

where w is the total number of words spoken.

In fact, if we require a commonality threshold of $c = 0.3$, this forces a relative frequency r of at least the following for each group.

- All Children (Ages 4 through 6): $r \geq 0.6482569$

- Children of Age 4: $r \geq 0.6706533$
- Children of Age 5: $r \geq 0.6582701$
- Children of Age 6: $r \geq 0.6134969$

Thus, unless a frequency threshold surpasses the value above, this threshold will not influence the selection of core words. If it is desired for frequency to play a role in the selection core words, I suggest selecting a frequency threshold strictly larger than the values listed above.

Regardless of this result, for the remainder of this report, I continue with the commonality and frequency thresholds $c = .3$ and $r = .5$, respectively, because this was requested. This implies that core words were actually only chosen with the condition that the commonality was at least 30%.

3 Core Words

Below, I give a table that previews the first 6 core words for children in the combined group. The core words listed below are in order of decreasing frequency. In total, there were 122 core words selected. The complete list of core words can be found on GitHub in the excel file "core_words_processed.xlsx"[1] under the "Core Words for Combined Group" tab.

Preview of Core Words Across All Ages

	Composite.frequency	Relative.frequency.per.1000.words	Commonality
我	2656	45.19	0.99
的	2437	41.46	0.99
個	1917	32.62	0.98
是	1507	25.64	0.98
有	1466	24.94	0.98
就	1165	19.82	0.96

Similarly, core words for other groups can be found. Below is a preview of core words for 4-year-olds, 5-year-olds, and 6-year-olds, which had a total of 121, 125, and 126 core words selected, respectively. It is worth noting that the number of core words increases with each age group. This matches the intuition that older children will have a broader vocabulary. Also of interest is the observation that one child in the 6-year-old group did not use the word "我" which corresponds to the English word "I".

Preview of Core Words for 4-Year-Olds

	Composite.frequency	Relative.frequency.per.1000.words	Commonality
的	856	47.84	1
我	839	46.89	1
個	642	35.88	1

Preview of Core Words for 5-Year-Olds

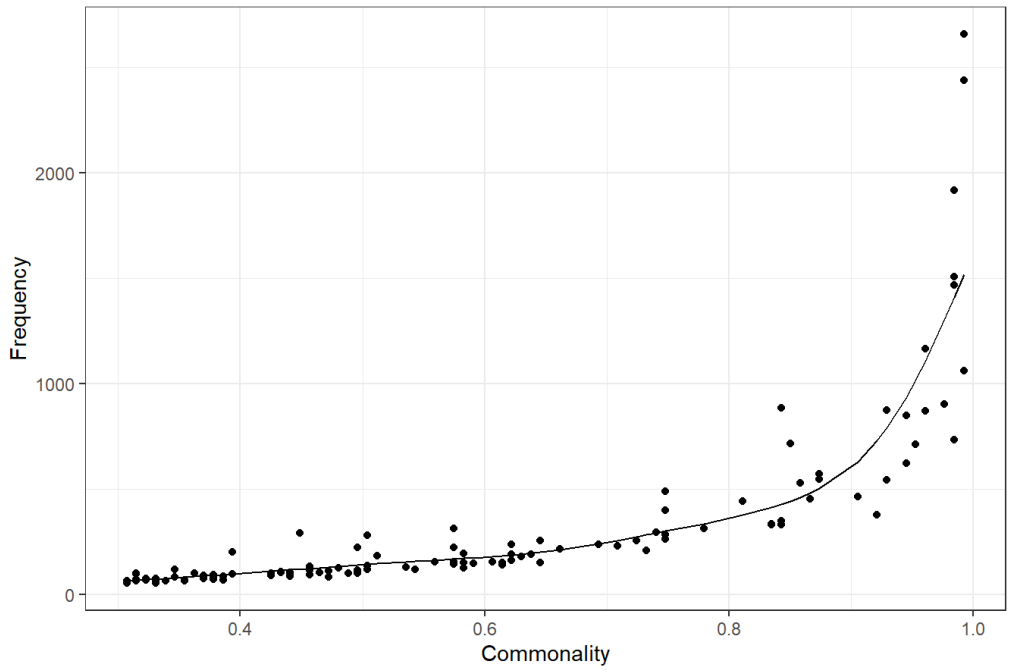
	Composite.frequency	Relative.frequency.per.1000.words	Commonality
我	960	42.13	1
的	910	39.94	1
個	734	32.21	1

Preview of Core Words for 6-Year-Olds			
	Composite.frequency	Relative.frequency.per.1000.words	Commonality
我	857	47.37	0.97
的	671	37.09	0.97
個	541	29.90	0.95

The complete list of each of these charts can be found also in the excel file “core_words_processed.xlsx” under the corresponding “Core Words for X-Year-Olds” tab.

4 Relationship Between Commonality and Frequency

Next, I explore the relationship between commonality and frequency. To provide one possible summary of this relationship, a smoothing spline was fit to the set of core words (as defined by the combined group). This can be found below. As expected, we see a generally increasing relationship. Graphically, we also see increased variability in the relationship as the commonality increases.



A smoothing spline is fit to the relationship of commonality and total frequency of core words in the combined group of children.

The plot above can be used to summarize the relationship graphically as well as to predict the frequency of a core word given its commonality.

Smoothing splines require choosing a parameter that controls the “wiggliness” of the line. In this analysis, the parameter was chosen by visually selecting a curve that best fit the data without appearing to overfit. The chosen curve provides a balance between flexibility and overfitting. However, the smoothing parameter can also be chosen using cross validation. Choosing the parameter using this method results in a “wigglier” line that using intuition appears to me to suffer from over-fitting.

Other potential models are explored in the Appendix. Namely, an exponential model is fit as well as a discontinuous piecewise function. These models have the advantage of having a fitted curve with a closed-form equation. However, they do not fit the shape of the data as well as the smoothing spline displayed above.

Alternatively, a smoothed scatterplot is also presented in the Appendix.

5 Core Words Between Different Age Groups

Next, we look to see if the commonality of possible core words change between different age groups. To do this, we consider all words that are determined to be a core word in the 4-year-old, 5-year-old, or 6-year-old group. Below is a preview of a chart which catalogs the groups in which each word is classified as a core word. The complete charts can be found in the excel file “core_words_processed.xlsx” under the “Core Words Comparison” tab.

Preview of Core Words for Different Age Groups						
	X4.Year.Olds	X5.Year.Olds	X6.Year.Olds	Composite.frequency	Relative.frequency.per.1000.words	Commonality
我	x	x	x	2656	45.19	0.99
的	x	x	x	2437	41.46	0.99
個	x	x	x	1917	32.62	0.98
是	x	x	x	1507	25.64	0.98
有	x	x	x	1466	24.94	0.98
就	x	x	x	1165	19.82	0.96

In the above chart, if a word is considered core within a certain age group, this is represented with an “x” in the corresponding cell. This chart also includes the frequency and commonality in the combined group and orders each word by decreasing frequency.

For each of the words considered core for any age group, we conduct a statistical test to see if the commonality of the word significantly differs between two different age groups. This is done though a z-test for the difference of proportions (or equivalently a chi-squared test for independence).

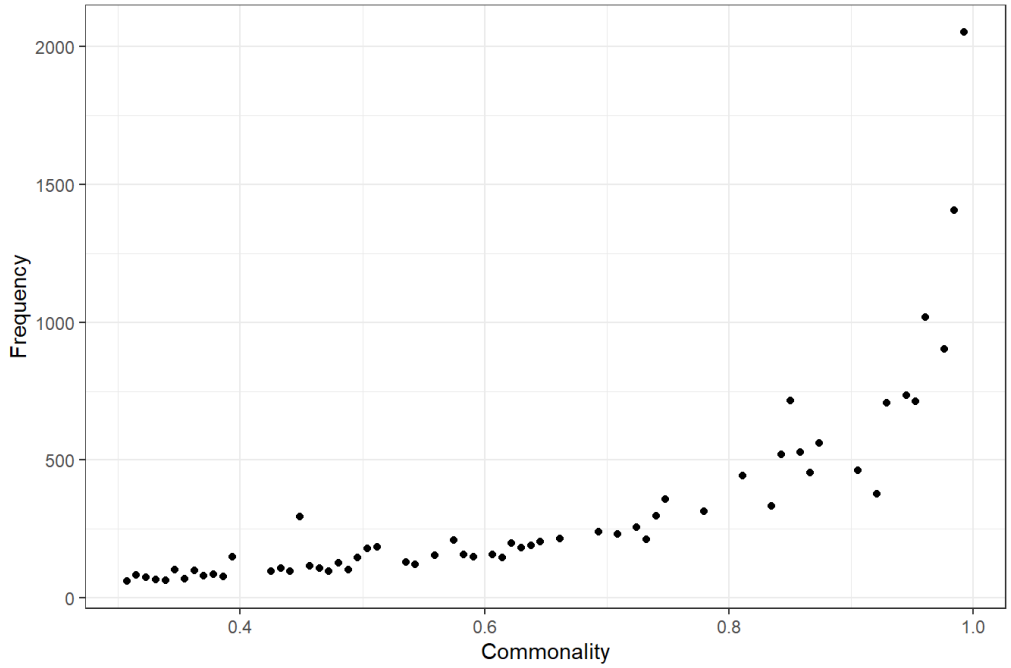
After accounting for multiple tests using the Bonferroni correction, we find no significant difference between the commonality of any core word in the age group of 4 verses the age group of 5. Similarly, we find no significant difference between the age group of 5 and the age group of 6.

We do find a significant difference in the commonality of at least one word between the 4-year-old age group and the 6-year-old age group. Specifically, the word 她 (meaning ‘she’ or ‘her’) is significantly more common in the 6-year-old age group than in the 4-year-old age group with a commonality of 62% verses 13%. This suggests that the core words between these two age groups are significantly different.

In conclusion, our analysis suggests that creating different Aided AAC for 4-year-olds and 6-year-olds may be advantageous, especially if considering the commonality threshold of 0.3 and frequency threshold of 0.5.

6 Appendix: Other Models of the Relationship Between Commonality and Frequency

We begin with an alternative scatterplot representation that smooths the relationship by averaging the frequency of words with the same commonality.



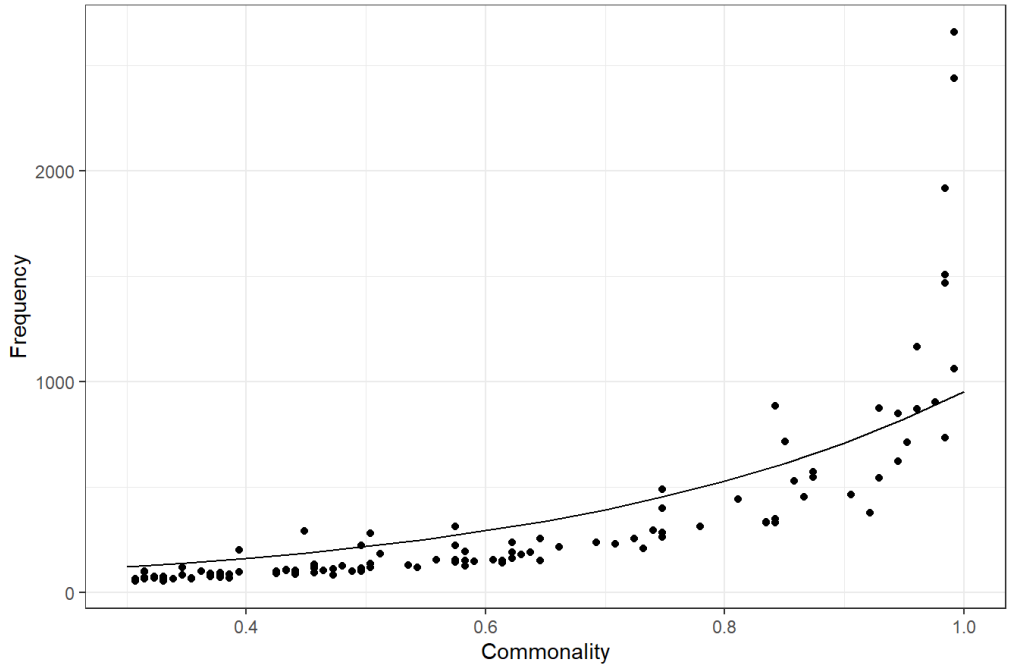
A smoothed scatterplot created by averaging the frequency of core words with the same commonality in the combined group of children.

In addition to the smoothing spline presented in the main report, we also explored other models for the relationship between commonality and frequency. Below are plots corresponding to two other potential models followed by a brief discussion on each model.

We first consider an exponential model. One advantage to this model is that the line of best fit can be expressed in a closed-form expression. In fact, fitting an exponential curve to the relationship between commonality and frequency we see can obtain the prediction

$$\text{Estimated Frequency} = 50.46 \cdot 18.88^{c}$$

for a word with commonality *c*. This curve is displayed below. Unfortunately, this curve, while easy to interpret, does not fit the shape of the relationship well.



An exponential model is fit to the relationship of commonality and total frequency of core words in the combined group of children.

We next considered a piece-wise smooth model based on intuition. It is possible that words can be described as either “very common” or “not very common”. The very common words can be thought of as the English equivalent of “I”, “is”, etc. The very common words correspond to the words on the far right-hand side of the scatterplots.

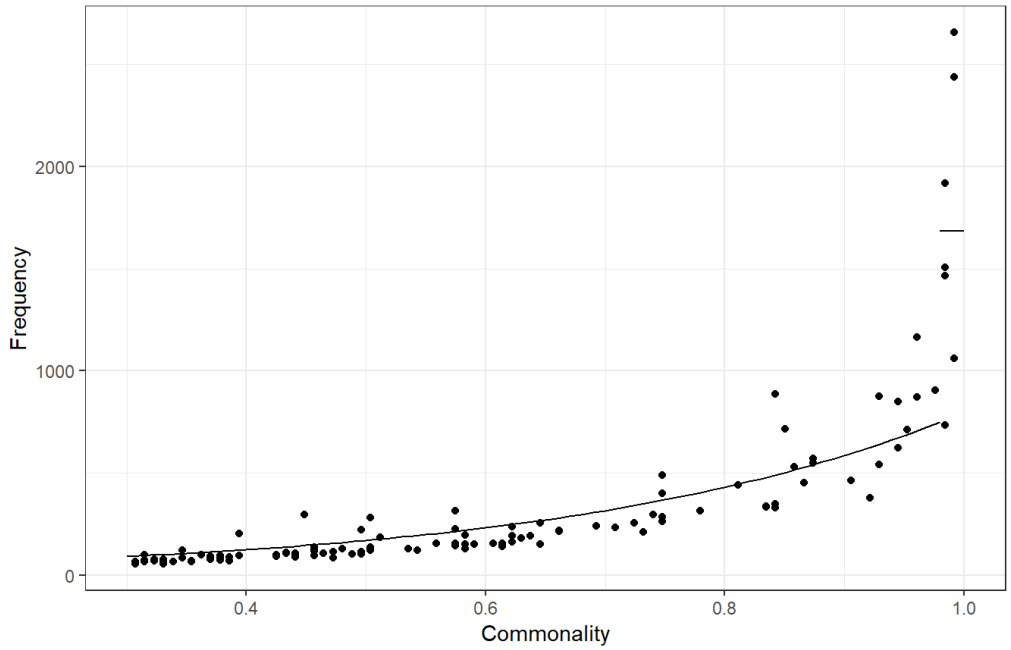
These very common words appear to have a different distribution than the other words. For example, the variance in the very common words is much higher. There is also much less of a clear pattern. One possible hypothesis is that not very common words fit an exponential distribution while very common words are randomly distributed around some common average frequency.

To fit a model using the above description, the process involves (1) determining a threshold for classifying words as very common, (2) fitting an exponential model for the not very common words, and (3) establishing an average frequency for very common words.

Below is the curve that minimizes squared error which uses a threshold of 0.98 commonality. This model has the same benefit as the exponential model in that the relationship can be expressed in closed-form, while having a much closer fit to the true data.

$$\text{Estimated Frequency} = \begin{cases} 36.22 \cdot 22.02^{c} & \text{if } c \leq .98, \\ 1682.57 & \text{if } c > .98, \end{cases}$$

for a word with commonality c.



A piecewise model is fit to the relationship of commonality and total frequency of core words in the combined group of children. For words below .98 commonality, an exponential model is fit. For words above .98 commonality, the average frequency is displayed.

One potential drawback to the above curve is that the model performance is very sensitive to the choice of threshold. Also, words that have commonality either just to the left or the threshold or just to the right of the threshold will have drastically different estimates of frequency which is not ideal.

Because of the drawbacks of the above models, I believe the smoothing spline given in the main report is a better choice for representing the relationship between commonality and frequency.

---

1. Link: https://github.com/smithmor/Aided_AAC_Analysis/blob/7cab44371c38a43652a1a81eacb7862b4a6128c9/core_words_processed.xlsx (https://github.com/smithmor/Aided_AAC_Analysis/blob/7cab44371c38a43652a1a81eacb7862b4a6128c9/core_words_processed.xlsx