

Preliminary Vocabulary Frequency Findings for Mandarin Chinese AAC Treatments

Ming Chung Chen – National Chiayi University

Katya Hill – University of Pittsburgh

Tianxue Yao – Carnegie Mellon University

Abstract

In this paper, we describe the preliminary results of a study to identify the vocabulary frequency of native Mandarin Chinese speakers during a dyadic conversation. The language samples based on twelve (N=12) participants were analyzed to report the total number of words, total number of different word roots (TND), and number of spoken words used to make up 50%, 60%, 70% and 80% of the sample. Reported are the top 10 most frequently used MC words for each participant. The results will provide a high frequency or core vocabulary for MC to use for augmentative and alternative communication (AAC) interventions.

Research Description

Introduction

Vocabulary selection and organization has been central to making AAC decisions for any language (Baker, Hill, Amato, & Menna, 2000; Baker, Musselwhite, & Kwasniewski, 1999). Studies of the most frequently occurring words from language samples of different populations have been used to support pre-stored vocabularies for English (Balandin & Iacono, 1999; Beukelman, Jones, & Rowan, 1989; Hill, 2000; Stuart, Beukelman, & King, 1997) The language samples revealed how a relatively small number of words (approximately 250-400 words) occur consistently across English speakers. Statistically, the high frequency words or “core” vocabulary make up 80-85% of the total words used in a language sample.

Current frequency studies based on word usage patterns during face-to-face conversations are not available for Mandarin Chinese (MC). MC studies (Hwang, et al., 2004; Tseng, 2004) have reported MC character or “unit of meaning” frequencies, but not word frequency (Liu & Zachary, 2006). Word frequency is critical to selecting a core vocabulary for AAC interventions, because high frequency words are common to topics, situations, environments, and activities. Individuals who rely on AAC need access to the core words in order to build language competence and optimize communication.

For AAC interventions, a critical clinical question focuses on identifying the vocabulary that will result in the best fluent communication. Simply translating the “gloss” (printed word) associated with the graphic symbols used in the United States to provide a “set of symbols” representing the vocabulary of a language, e.g. French, German, Spanish, or in this case, Mandarin Chinese does not support vocabulary selection and access needed for communication competence. The development and introduction of most graphic symbol-based AAC interventions starts with identifying high frequency vocabulary and defining the core vocabulary of a natural language (Baker & Chang, 2006).

The purpose of this investigation was two fold: 1) to demonstrate reliable transcription and word segmentation methods for MC, and 2) to identify the vocabulary frequency used by native MC speaking adults during dyadic conversations. Specifically, we sought to determine the frequency and commonality of vocabulary (words) used by MC speakers. The establishment of a high frequency word usage list can be used to select a core vocabulary for an augmentative and alternative

communication (AAC) system and to support AAC treatment decisions.

Method

1.1 Participants

Twelve adults (six male and six female) whose native language was Mandarin Chinese were recruited for the study. The average chronological age of the participants was 39.8 (SD=18.4) with a range from 26 to 67 years. They all had a minimum of a high school diploma. No disabilities were reported.

1.2 Procedure of language sample collecting

The language sampling procedure was typical of previous studies harvesting conversational samples in closed environments (Graham & Hill, 2007; Stuart, et al., 1997). An interview was conducted in the Performance and Testing Lab at the University of Pittsburgh. The language samples were collected after the participants became accustomed to talking under the condition of wearing the microphone. The investigators started with introductory comments, explained the basic procedure, assured them that at any time they may stop the procedure, and offered that they may request a break during the one (1) hour if necessary. No specific topics were prompted for the conversation, but topics were self-selected, spontaneous, and independent of the environment. The conversation was conducted exclusively in Mandarin Chinese.

1.3 Instrumentation

A digital voice recorder equipped with a tie-clip omnidirectional microphone was used to record the language sample. The recorded voice file was exported into a laptop with Microsoft Windows XP (Mandarin version). The Mandarin Chinese version of Microsoft Word (Version 2003) was used to transcribe the audio tapes. In addition, an online word segmentation system developed by the Institute of Information Science and the Institute of Linguistics of Academia Sinica in Taiwan (<http://ckip.iis.sinica.edu.tw/CKIP/wordsegment.htm>) was used to segment the sentence into words.

1.4 Data analysis

The middle 20 minutes of each conversation were transcribed by the third author and a graduate student from the Department of Communication Science and Disorders at the University of Pittsburgh. Both transcribers are native MC speakers. We selected the sample for transcript to purposefully exclude introductory and closing remarks. We want the participants to be relaxed and comfortable with the surroundings and not thinking about the vocabulary they were using or could use while carrying on a conversation.

Inter-rater reliability was calculated as the number of characters in both transcripts of a participant divided by the number of characters in the first transcript multiplied by 100. The percentage agreement based on 16% of the samples was 92% (range 90-94%).

Each participant's transcript was automatically segmented to generate the *word* list using the CKIP system. This system segments words using criteria such as word length, morphemes, and probability (Ma & Chen, 2003). The results from the CKIP segmentation step were saved and the word and its grammatical category were imported into SPSS 16.0 for data analysis. A text file of the word list for each participant was calculated. The frequency of each word was counted and the words were sorted by frequency. The cumulative percentage for word frequency was tabulated also to show the number of different words that represented a given percentage of the language sample.

2. Results

2.1 Number of words used in the dialogue

The Results of calculating the words spoken by the twelve participants during the 20 minute conversation using automated tools are shown in Table 1. The average number of total words was 2329.9 (SD=536.6.) with a range from 1567 to 3225 spoken words. The total number of different word roots (TND) used by the participants ranged from 380 to 718 word roots with a mean of 539.8 root words (SD=113.8)

The data revealed that 50% of the TND consisted of an average of 33.9 words (range of 29-39) for the participants. An average of 57.2 words (range 44-75) made up 60% of the total words used by participants. Finally, an average of 164.6 words (SD=33.7) was found to make up both 70% and 80% words used in the language samples.

2.2 Example for the high frequent words

The results of the top ten high frequency words by participants were shown in Table 2. As the table indicated, these top 10 words made up 22% to 29% of the words used by the participant in the 20 minute conversation. The results show that there is high overlap among the participants. Also, none of the top 10 words are nouns.

subject	Total words	TND	50%	60%	70%	80%
1	2596	644	38	64	113	199
2	2445	640	33	59	107	206
3	3225	642	35	56	98	170
4	2744	615	36	59	98	175
5	2273	518	36	56	89	152
6	1709	453	34	54	90	158
7	2875	585	35	57	100	171
8	2233	430	29	46	74	124
9	2741	718	39	63	118	226
10	1618	415	33	53	88	148
11	1567	380	29	44	71	117
12	1933	437	30	75	75	129

3. Conclusions

Our preliminary findings support the earlier attempts to identify MC vocabulary using automated approaches or MC databases. This early data analysis shows that a relatively small number of words (less than 200 words in this sample) will make up 80% of the spoken words used in conversation. These findings are similar to the vocabulary frequency studies in European languages and demonstrate that high frequency words or core vocabulary is consistent across different speakers and topics. The MC core vocabulary can be used as the foundation to support building language fluency using an AAC system.

Table 2.
Top 10 high frequency words/units of meaning used by the participants.

sub	Top 10 words/units of meaning*	%
1	的(DE)、個(GE)、我(I, me)、那(that)、是(to be)、就(just, then)、了(LE)、很(very)、一(a, one)、這(this)	29.2
2	我(I, me)、的(DE)、是(to be)、個(GE)、那(that)、就是(it is)、對(right, yes)、有(to have)、就(just, then)、啊(ah)	24.7
3	的(DE)、個(GE)、這(this)、他(he, him)、是(to be)、就(just, then)、那(that)、種(kind)、一(a, one)、你(you)	28.8
4	的(DE)、個(GE)、是(to be)、我們(we, us)、你(you)、我(I, me)、那(that)、這(this)、有(to have)、一(a, one)	25.3
5	就(just, then)、啊(ah)、的(DE)、去(to go)、說(said)、是(to be)、對(right, yes)、我們(we, us)、個(GE)、他們(they, them)	22.5
6	的(DE)、對(right, yes)、啊(ah)、是(SHI)、有(to have)、就是(it is)、那(that)、我(I, me)、個(GE)、就(just, then)	29.3
7	我(I, me)、的(DE)、就(just, then)、是(to be)、都(all, both)、啊(ah)、你(you)、要(want)、我們(we, us)、啦(LA)	24.1
8	的(DE)、是(to be)、我(I, me)、你(you)、他們(they, them)、那(that)、他(he, him)、對(right, yes)、有(to have)、個(GE)	28.4
9	我(I, me)、的(DE)、就(just, then)、然後(then)、是(to be)、去(to go)、個(GE)、那(that)、要(want)、很(very)	23.1
10	的(DE)、是(to be)、有(to have)、也(also)、嗯(um)、對(right, yes)、啊(ah)、那(that)、個(GE)、你(you)	26.6
11	的(DE)、我(I, me)、就(just, then)、那(that)、你(you)、是(to be)、就是(it is)、有(to have)、嗯(um)、不(not)	26.7
12	的(DE)、嗯(um)、對(right, yes)、我(I, me)、會(will)、是(to be)、個(GE)、那(that)、就是(it is)、就(just, then)	28.9

*capitalized letters in parentheses indicate pronunciation of MC character; English word for MC character(s) provided in parentheses were appropriate.

Reference

- Baker, B., Musselwhite, C., & Kwasniewski, K. (1999). *Literacy, language, and minkspeak: Core vocabulary is the key*. Presentation for the Summer Seminar on Literacy and AAC. Durham, NC.
- Baker, B. & Chang, S. (2006). A Mandarin language system in augmentative and alternative communication (AAC). *International Journal of Computer Processing of Oriental Languages*, 19, 225-237.
- Baker, B., Hill, K., Amato, J., & Menna, D. (2000, August). *Do we liberate individuals by teaching wide context specific vocabulary?* In Proceedings of the 9th International Society for Augmentative and Alternative Communication Biennial Conference(pp 727-729). Washington DC: ISAAC.
- Balandin, S. & Iacono, T. C. (1999). Wusses, and whoppas: Core and fringe vocabularies of Australian meal-break conversations in the workplace. *Augmentative and Alternative Communication*, 15, 95-109.
- Beukelman, D. R., Jones, R., & Rowan, M. (1989). Frequency of word usage by nondisabled peers in integrated preschool classrooms. *Augmentative and Alternative Communication*, 5, 243-248.
- Graham, K. & Hill, K. (2007, September). *A Pilot Study Comparing AAC Vocabulary Usage Patterns Based on User Experience*. Poster presented at the 2007 Clinical AAC Research Conference. Lexington, KY.
- Hill, K.(2001). *The development of a model for automated performance measurement and the establishment of performance indices for augmented communicators under two sampling conditions*. Ph.D. Thesis, University of Pittsburgh, Pittsburgh, PA.
- Hwang, M. Y., Lei, X., Ng, T., Bulyko, I., Ostendorf, M., Stolcke, A., et al., (2004, December). *Progress on mandarin conversational telephone speech recognition*. In Proceedings of the Fourth International symposium on Chinese spoken language processing. (pp. 1-4). Hong Kong: Chinese University of Hong Kong.
- Liu, C. & Zachary, S. (2006). Developing a core vocabulary for a mandarin Chinese AAC system using word frequency data. *International Journal of Computer Porcessing of Oriental Languages*. 19, 285-300.
- Ma, W. Y., & Chen, K. J. (2003, July). *Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff*. Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing(pp. 168-171). Sapporo, Japan.
- Stuart, S., Beukelman, D. R., & King, J. (1997). Vocabulary use during extended conversations by two cohorts of older adults. *Augmentative and Alternative Communication*, 13, 40-47.
- Tseng, S. C. (2004, December). *Spontaneous mandarin production: Results of a corpus-based study*. In Proceedings of the Fourth International Symposium on Chinese Spoken Language Processing (29-32). Hong Kong: Chinese University of Hong Kong.