Sino-US English Teaching, June 2025, Vol. 22, No. 6, 189-195
doi:10.17265/1539-8072/2025.06.001

A Corpus-Based Study on the Use Frequency of Large Numerals by Mandarin-Speaking Children

YANG Meiling, DU Ailin
Northeastern University, Shenyang, China

The usage characteristics of the large numerals in child language reflects both linguistic and cognitive development. The present study systematically examined Mandarin children's naturalistic use of large numerals "bai", "qian", and "wan" and compared it with adult usage patterns based on a self-constructed corpus. The results revealed a significant preference for the use of "bai" over both "qian" and "wan" among children and adults alike. This shared pattern suggests that children's acquisition of large numerals is strongly shaped by adult language input. This study addresses a critical gap in research on the acquisition of large numerals by Mandarin-speaking children, providing valuable insights for the broader theoretical framework of cross-linguistic numerical cognition.

Keywords: large numerals, child language, use frequency, corpus-based study

Introduction

Numerals serve as a key instrument for children not only to comprehend the world around them, but also to manage their daily lives, for example, recognizing how many toys they have, following instructions like "three more bites" at mealtime, or using classroom numbers to find their way around school. In these everyday situations, children do more than simply recite number words; they must connect abstract numerical concepts, such as quantity and order, with real-world contexts. Through this process, children come to understand, apply, and effectively communicate numerical concepts with others. Consequently, a key question arises regarding how children acquire the meanings of numerals in early childhood.

Extensive research has been conducted on children's acquisition of number words, with the majority of studies concentrating on the acquisition of small numerals. For instance, Wu and Xu (1979) analyzed the speech development records of children from birth to three years olds and they found that most children are generally able to comprehend the basic meanings of the numerals "one" and "two", but they do not yet have a clear grasp of "ten". Wynn (1992) found that children's acquisition of small numbers follows a gradual, stepwise progression. At around two and a half years old, children can understand "one", but struggle with other numerals. Over the next few years, they progressively master numerals like "two", "three", and "four". By age five, children appear to grasp the internal logic of numeral systems and are able to accurately use numbers like "five" and beyond.

Acknowledgments: This study is funded by Liaoning Provincial Social Science Planning Fund Project "The Development of Children's Large-Number Representation and the Role of Parental Linguistic Input: A Cross-Linguistic Study of English and Chinese" (Project Fund No. L22CXY003) (本文系辽宁省社会科学规划基金项目"英汉对比视域下儿童大数量表征发展及父母语言输入的影响研究"(基金编号：L22CYY003)的阶段性研究成果).

YANG Meiling, Ph.D., lecturer, Foreign Studies College, Northeastern University, Shenyang, China.
DU Ailin, undergraduate, Foreign Studies College, Northeastern University, Shenyang, China.

These studies have primarily focused on understanding how young children learn and use basic number words, often overlooking the complexities involved in the acquisition of larger numerals.

Compared with small number words, large number words, such as “bai” (“hundred”), “qian” (“thousand”), and “wan” (“ten thousand”) involve a greater number of quantity units and more complex numerical relationships. Moreover, unlike small numerals like “one” or “two”, which are easily linked to tangible objects in the physical world, larger numerals often represent quantities that extend beyond immediate perceptual experience. Together, these characteristics make the acquisition of large numerals particularly challenging for young children. Yet, few studies have explored children’s acquisition of larger number words, such as “bai”, “qian”, and “wan”. Among the limited existing research, most studies use number line estimation tasks to examine how children comprehend the relative magnitude of these large numerals (Xu et al., 2023). While such approaches are valuable for probing underlying magnitude representations, they reveal little about how children actively employ these terms in spontaneous communication. Furthermore, the factors that influence children’s acquisition of large numerals remain largely unexplored. Existing research has notably highlighted the role of parental speech in children’s vocabulary acquisition. A large number of studies indicate that the frequency of words in parental input is a strong predictor of children’s vocabulary learning (e.g., Goodman, Dale, & Li, 2008). In the field of early numerical abilities, numerous studies have shown a robust correlation between parents’ mathematical language input and children’s number skills. For example, Huang (2024) found that higher-frequency exposure to parents’ input provides more opportunities for children to form mappings between number words and their meanings, accelerating the acquisition of early numerical concepts. However, few studies have examined how children’s use of large numerals is shaped by variations in parental input.

Given the limitations of prior research, the present study aims to examine the use frequency of large numerals “bai”, “qian”, and “wan” among Mandarin-speaking children based on a self-constructed corpus. By incorporating a comparative analysis with data from their parents or caregivers, this study also seeks to explore the potential influence of input from older speakers on children’s acquisition of large numerals. Specifically, the study aims to answer the following research questions:

1. What are the similarities and differences in the frequency distribution patterns of large numerals, such as “bai”, “qian”, and “wan” between Mandarin-speaking children and their parents or caregivers?

2. What factors might explain the observed similarities and differences in the use of these numerical terms between children and their parents or caregivers?

Method

Data Collection

As one of the most authoritative child language corpora in the world, the Child Language Data Exchange System (CHILDES) provides essential data for the study of children’s language development (Wen & Hu, 2001). The present analysis draws on Chinese sub-corpus available in CHILDES. It contains transcripts of naturalistic conversations between Mandarin-speaking children and their parents or caregivers, thereby providing a solid foundation for examining the use of large numerals in real-life communicative contexts.

Specifically, to construct a corpus suitable for quantitative analysis, we used these large numeral terms “bai”, “qian”, and “wan” as search keywords to extract naturalistic speech samples from children aged three to six years and their adult caregivers. The retrieved data were first carefully screened to ensure data quality. During this process, datasets containing missing critical information, such as children’s age or the number of participants,

were excluded. The final corpus consisted of data from 79 children and their caregivers, with the average age of the children being 59.82 months (*SD* = 14.65).

Research Procedure

Based on the self-constructed corpus, the study proceeded in following steps:

First, the raw frequencies of large numerals, namely, "bai", "qian", and "wan", were systematically extracted from both the child and adult sub-corpora using the FREQ function of the CLAN data analysis program. For example, to extract occurrences of "bai" in children's speech, we employed the commands "freq + t*CHI + s'百'.cha"/"freq + t*CHI + s'hundred'.cha". Here, "freq" presents the token frequency of a specific keyword, "t*CHI" restricts the search to child speech, and "s'百/hundred'" defines the search term. Similar procedures were applied to extract the frequencies of "qian" and "wan". The raw frequencies were then standardized as the number of occurrences per 10,000 words. Finally, chi-square tests were conducted to examine whether there were significant differences in frequencies of large numerals between children and adults.

Results and Discussion

Frequency Distribution Patterns

We analyzed the frequency of large numerals "bai", "qian", and "wan" in both the child and adult sub-corpora. Detailed results are presented in Table 1.

Table 1

Frequency Distribution of "Bai", "Qian", and "Wan" in Child and Adult Speech

	Large numerals	Raw frequency	Normalized frequency
Child	"Bai"	118	4.74
	"Qian"	26	1.04
	"Wan"	21	0.84
Adult	"Bai"	140	3.42
	"Qian"	16	0.39
	"Wan"	21	0.51

As indicated in Table 1, both children and adults demonstrate a high degree of consistency in their use of large numerals. First, compared with numerals "qian", and "wan", "bai" is the most frequently occurring large numeral in both child and adult speech. This finding indicates a distinct preference for the use of "bai" across both age groups. This shared preference may suggest a potential developmental trajectory in children's acquisition of large numerals. It is possible that "bai" functions as a foundational conceptual anchor. It allows children to initially construct its meaning through experiences and subsequently provides a cognitive basis for acquiring larger and more abstract numerical concepts. This progression reflects the stepwise and sequential nature of children's numerical cognition, similar to the acquisition pattern of small number words.

Second, both children and adults exhibited low frequencies in their use of the numerals "qian" and "wan". The similar low frequency of "qian" and "wan" across age groups lies in their shared linguistic identity as "high-level large numerals." Their frequency of use is primarily shaped not by individual age or cognitive developmental stage, but rather by universal pragmatic rules inherent to the linguistic system, such as the specificity and abstractness of usage contexts. In everyday communication, caregivers and other adults more frequently employ the numeral "bai" in a variety of familiar contexts, such as referring to prices (e.g., "one

hundred yuan"), describing ages (e.g., "one hundred years old"), or using common idiomatic expressions. In contrast, "qian" and "wan" occur less often in daily discourse because they often denote larger, more abstract quantities that are less immediately relevant to a young children's daily experiences. This asymmetrical input shapes children's acquisition process through frequency-dependent learning mechanisms. Repeated exposure to "bai" in meaningful and interactive situations enhances its perceptual salience and cognitive accessibility, making it easier for children to recognize, recall, and use this numeral productively.

Third, the observed similarities in the usage patterns of large numerals between adults and children were confirmed through statistical tests. Chi-square tests results revealed that there is no significant difference between children and adults in the usage frequency of large numerals: "bai" ($\chi^2 = 2.25$, $p = 0.13$), "qian" ($\chi^2 = 2.98$, $p = 0.08$), and "wan" ($\chi^2 = 0.01$, $p = 0.94$). These findings suggest that children's use of large numerals may be heavily influenced by adult language input. In naturalistic settings, adults often incorporate large numerals into everyday activities, such as mentioning prices while shopping, stating scores during games, or reading storybooks with page numbers. Moreover, adults often provide corrective feedback regarding children's numerical expressions during daily interactions. Through processes, such as imitation, repetition, and feedback incorporation, children gradually calibrate their numerical language production to align more closely with conventional adult usage. In addition, this finding suggests that the use of large numerals is shaped by specific pragmatic constraints. Across age groups, speakers tend to favor numerical units that minimize cognitive load and are better aligned with the pragmatic demands of the context. The numeral "bai" appears to have high functional utility in daily life. In contrast, numerals, such as "qian" and "wan" are more selectively used in specialized contexts, such as mathematical tasks or formal descriptions.

To gain deeper insights into children's use of large numerals, we analyzed the developmental trends in the usage frequencies of "bai", "qian", and "wan" across different age groups. Children were divided into six age groups ranging from three to eight years old. For each of these groups, both the normalized and cumulative frequencies of each numeral were calculated. The results are displayed in Figures 1 and 2.

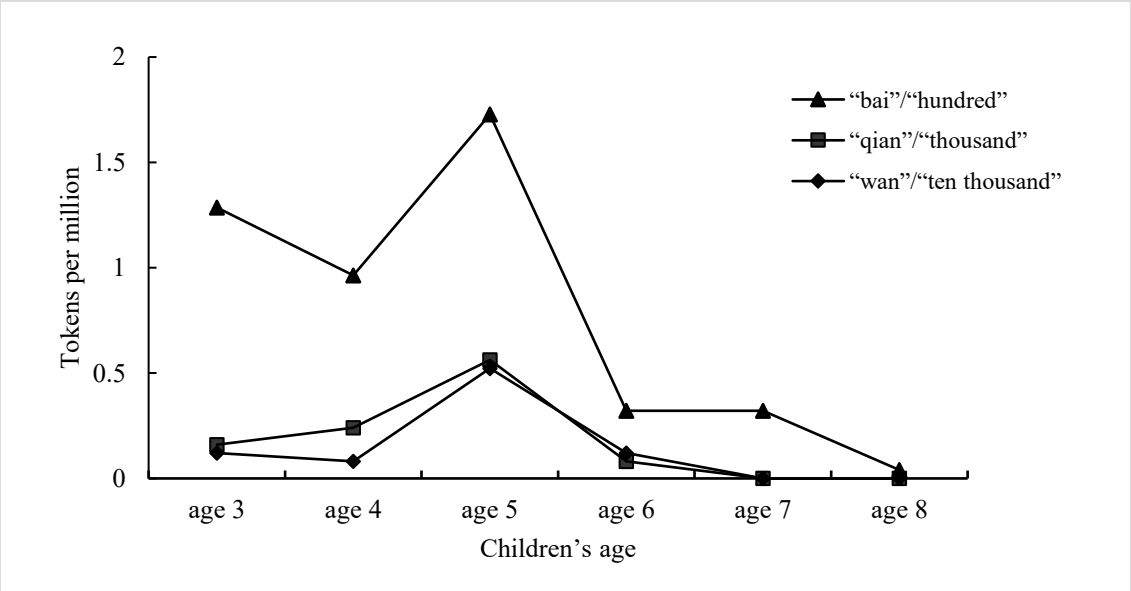

Figure 1. Normalized frequency distribution of the large numerals "bai", "qian", and "wan" in Mandarin-speaking children.

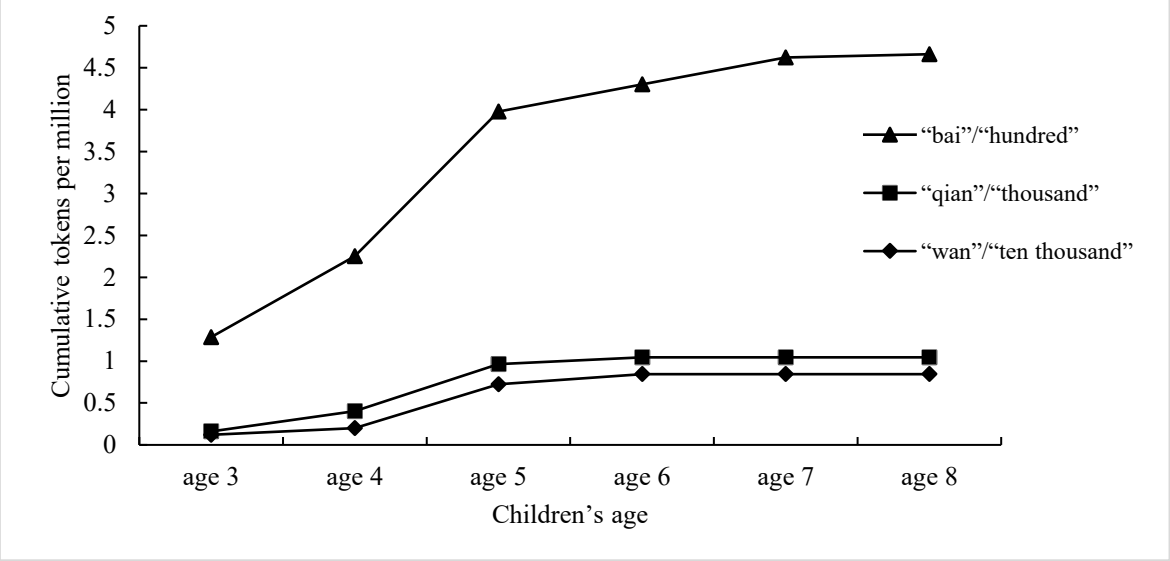

Figure 2. Cumulative frequency distribution of large numerals "bai", "qian", and "wan" in Mandarin-speaking children.

Figure 1 illustrates the normalized frequency distribution of the large numerals "bai", "qian", and "wan" among children aged three to eight. As shown in Figure 1, for all numerals, children's usage frequency exhibits a decline between the ages of three and four. After this initial decline, their usage frequency of these numerals begins a steady increase, which peaks around the age of five. Following the peak at age five, children's usage frequency of the three numerals gradually declines again as they grow older. The findings from this study reveal the fluctuating, non-linear nature of children's usage of large numerals over time. Van Geert's (1991) dynamic systems theory offers a compelling framework for interpreting this finding. He emphasizes that language acquisition is a complex, non-linear process shaped by interactions among the learner, the environment, and cognitive development. The observed patterns of fluctuation in numeral usage in the present study may reflect the continuous reorganization of cognitive systems as children move through different stages of mathematical development. In other words, as children progress through different stages of mathematical development, their understanding of numerical concepts, including large numerals, becomes more complex and sophisticated.

Figure 2 presents the cumulative frequency distribution of "bai", "qian", and "wan" among children. The results demonstrate a period of rapid increase in the cumulative usage of large numerals before the age of five, which is followed by a marked slowdown thereafter. This developmental trajectory is consistently observed across all target numerals, namely, "bai", "qian", and "wan". This finding is consistent with previous research on English-speaking children's usage of large numerals (Miller, Smith, Zhu, & Zhang, 1995). Moreover, this finding reinforces the notion that ages three to five represent a critical period for rapid development in both cognitive and linguistic abilities. From a neurocognitive perspective, this period of rapid growth coincides with significant developments in prefrontal cortex function and increased connectivity among parietal, temporal, and frontal regions associated with numerical processing (Ding et al., 2008). Chen (2015) pointed out that such neural maturation likely supports the transition from approximate, non-symbolic number sense to exact, language-mediated numerical cognition. Finally, the observed frequency hierarchy, with "bai" being the most frequent, followed by "qian" and "wan" in both children and adults, suggests that the acquisition of large numerals is not solely driven by conceptual complexity but is also profoundly shaped by input frequency and real-world salience.

Overall, the findings of this study regarding the usage patterns of large numerals in children and young adults reveal that large numeral acquisition is best characterized as a dynamic process driven by the interaction between inherent neurocognitive maturation and socially situated learning experiences. Future research incorporating longitudinal data is needed to more precisely quantify the relative contributions of inherent cognitive abilities, environmental input frequency, and neurodevelopmental changes to the acquisition of large numerals.

Conclusion

This study represents the first systematic attempt to investigate differences in the frequency patterns of large numerals, including “bai”, “qian”, and “wan”, in the speech of Mandarin-speaking children and adults. The findings provide important insights into how large numerals are used across different age groups and suggest key developmental patterns in children’s numerical language acquisition.

First, in terms of usage frequency, both children and adults exhibit a strong preference for the numeral “bai” compared with “qian”, “wan”, and “bai”. This pattern suggests that compared with “qian” and “wan”, “bai” is the most familiar and frequently used large numeral in everyday speech for both groups, likely due to its more common presence in both everyday contexts and educational settings.

Another key finding from the study is that no statistically significant differences were observed between children and adults in the overall distribution of the numerals “bai”, “qian”, and “wan”. This indicates that while the specific frequencies of large numerals differ in terms of usage across different age groups, the general patterns of numeral preference remain consistent. Both children and adults seem to prioritize “bai” as a core part of their numerical lexicon.

A particularly interesting aspect of the results is the dynamic trajectory of children’s use of large numerals. Specifically, a period of rapid growth occurs between the ages of three and five, during which children exhibit a sharp increase in the production of all three numerals. This rapid expansion aligns with other developmental theories, which suggest that ages three to five represent a critical period for cognitive and linguistic growth, particularly in relation to mathematical cognition. Notably, this consistent trajectory across all target numerals substantiates the view that numerical cognition develops along a dynamic and non-linear pathway. These findings further reinforce the theoretical proposition that the acquisition of large numerals is driven primarily by domain-general cognitive reorganization, rather than through item-specific memorization processes.

Altogether, this study systematically investigates the usage patterns of large numerals in Mandarin-speaking children and their caregivers. The findings not only delineate the normative developmental pathway for large numeral learning in Mandarin-speaking children but also highlight the interplay between cognitive readiness, environmental input, and conceptual complexity in the emergence of numerical abilities.

References

Chen, Y. H. (2015). The developmental characteristics and psychological mechanisms of the magnitude representation and number concept in children. *Acta Meteorologica Sinica, 31*(1), 21-28.

Ding, X. Q., Sun, Y., Braass, H., Illies, T., Zeumer, H., Lanfermann, H., & Fiehler, J. (2008). Evidence of rapid ongoing brain development beyond 2 years of age detected by fiber tracking. *American Journal of Neuroradiology, 29*(7), 1261-1265.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language, 35*(3), 515-531.

Huang, H. H. (2024). Math talk in play contexts: Relations between parent and child math language and early math skills. *Early Childhood Education Journal*, 1-11. Retrieved from https://link.springer.com/article/10.1007/s10643-024-01783-w

Miller, K. F., Smith, C. M., Zhu, J., & Zhang, H. (1995). Preschool origins of cross-national differences in mathematical competence: The role of number-naming systems. *Psychological Science, 6*(1), 56-60.

Van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review, 98*(1), 3-53.

Wen, Z. J., & Hu, H. L. (2001). Developing and utilizing the world's largest children's corpus—CHILDES. *Foreign Language Teaching and Research, 45*(5), 374-377.

Wu, T. M., & Xu, Z. Y. (1979). A preliminary analysis of the speech development records of children from birth to three years old. *Acta Psychologica Sinica, 24*(2), 153-165.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology, 24*(2), 220-251.

Xu, X., Chen, C., Wang, L., Zhao, M., Xin, Z., & Liu, H. (2023). Longitudinal relationship between number line estimation and other mathematical abilities in Chinese preschool children. *Journal of Experimental Child Psychology, 228*, 105619.