8-2007

Early vocabulary development in English, Mandarin, and Cantonese : a cross-linguistic study based on Childes

Shuxia Liu

EARLY VOCABULARY DEVELOPMENT IN ENGLISH, MANDARIN, AND CANTONESE: A CROSS-LINGUISTIC STUDY BASED ON CHILDES

SHUXIA LIU

Abstract

Early language development is an exciting topic in the field of child language acquisition. Only a limited amount of cross-linguistic studies has attempted to investigate the similarities and differences in child language development across different languages. In this thesis, I present a study based on English, Mandarin and Cantonese corpora extracted from the Child Language Data Exchange System (CHILDES, MacWhinney, 2000). I investigated the lexical compositions of certain lexical categories (nouns, verbs, and adjectives) in children and their caregivers' vocabularies across eight different children age groups ranging from 13 to 60 months. ANOVA, frequency analysis, and cluster analysis were used to analyze the data. The development trajectories of lexical diversity and complexity of children's speech were also analyzed by two novel techniques: D-measure and the Mean Length of Utterances. My research clearly shows that (1) in all the cultures, children's early language development exhibits roughly similar patterns: an increasing diversity in lexicon and increasingly complicated speech patterns emerge as a function of time, and children's vocabularies become more similar to those of their parents over time; and (2) culture variations in children's linguistic input have strong influences on their language output, which is reflected in the noun vs. verb ratio and the varying percentages of nouns, verbs, and adjectives in the total words children are able to speak in the three cultures.

I certify that I have read this thesis and find that, in scope and quality, it satisfies the requirements for the degree of Master of Arts.

Signature

Dr. Ping Li, Thesis Advisor

Signature

Dr. Jane Berry, Committee Member

Signature

Dr. Elizabeth Crawford, Committee Member

EARLY VOCABULARY DEVELOPMENT IN ENGLISH, MANDARIN, AND CANTONESE: A CROSS-LINGUISTIC STUDY BASED ON CHILDES

By

SHUXIA LIU

M.A., Tianjin Normal University, Tianjin, China, 2004

B.A., Tianjin Normal University, Tianjin, China, 2001

A Thesis

Submitted to the Graduate Faculty

of the University of Richmond

in Candidacy

for the degree of

MASTER OF ARTS

in

Psychology

August, 2007

Richmond, Virginia

Acknowledgments

I would like to express my gratitude to my supervisor, Dr. Ping Li, for his guidance and advice on this project. I would also like to thank the other members of my committee, Dr. Jane Berry and Dr. Beth Crawford for the valuable comments and assistance that they provided at all levels of the research project.

In addition, I would also like to thank Mr. Michael West, whose proofreading efforts made this thesis more readable. I must also acknowledge Chung Lung Chan, a willing and able undergraduate research assistant in our lab, whose help on the Cantonese portion made the thesis more complete.

Appreciation also goes out to my family for the constant support they provided me throughout my entire life. Thanks to my parents, my two brothers and my sister.

Last but not least, a special thanks to my husband, Dr. Xiaowei Zhao, without whose love, encouragement and technical assistance, I would not have been able to complete this thesis.

Early Vocabulary Development in English, Mandarin, and Cantonese: A Cross-Linguistic Study based on CHILDES

Language ability is probably the most important feature that distinguishes human beings from other species. The processes by which children acquire their native language or languages have attracted the attention of researchers in a variety of areas such as linguistics, psycholinguistics and cognitive science. The language acquisition problem has been rated as one of the three topics of greatest interest in the field of psycholinguistics (Aitchison, 1998). Many researchers believe that the acquisition of a first language must be considered one of the most important achievements of early childhood. Generally speaking, after preliminary sound practice stages such as cooing and babbling, children start to produce their first words around the age of 12 months. Then, on average, at around 18 to 20 months of age, many children show a rapid increase of their vocabulary size. This dramatic acceleration in learning rate is often referred to as *vocabulary spurt*, or *naming explosion* (Dromi, 1987; Goldfield & Reznick, 1990; Li, Zhao, & MacWhinney, 2007). For English speaking children, by the age of six years old, when they enter elementary school, evidence shows that they have frequently already grasped around 14,000 words (Carey, 1978). This indicates that children are learning, on average, approximately nine or ten new words per day during this period. This is an extremely fast rate of language acquisition when compared to the painfully slow rate of language acquisition experienced by adult learners.

This rapid rate of early language development has made many scholars believe that there must be a "language instinct" that is innately coded in the human genome (Chomsky, 1968; Pinker, 1994). However, some researchers believe that the child's linguistic environment including language input (the speech a child hears during daily life, or so-called *child-directed speech*, see Foster-Cohen, 1999, MacWhinney, 2000) also plays a critical role in language development (Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996; Tomasello & Slobin, 2005). In light of this theory, a variety of empirical research, including both longitudinal and cross-sectional studies, has been conducted to investigate the characteristics of children's early speech (language output), the linguistic input they receive, and the relationship between these two variables.

In every culture, adults use hundreds of thousands of words to communicate with each other. The child's lexicon also includes a considerable number of words that belong to various grammatical categories (Hart & Risley, 1995). In addition, children often experience a wide variety of differences in the language input that they receive from their environment**.** It is interesting to note that, despite the wide variation of input they receive, children (at least those with a shared culture) often show similar patterns in their early acquired vocabularies. For example, it is generally the case that a child's language comprehension occurs ahead of speech production, and the vocabulary that a child is able to comprehend is much larger than the child's spoken vocabulary (Reznick & Goldfield, 1992). Another interesting finding is that the early words of English-speaking children often include some common nouns in reference to objects with solid functional and physical properties like *ball, box, bubble, car*; and words that

describe people or things in their immediate environment, such as *daddy, mommy* and *baby* (Bates, Dale, & Thal, 1995). These "referential style" words are often learned first by children, and more-challenging verbs and adjectives will join their vocabulary later. Closed-class words[1] become frequent even later (Bates, Marchman, Thal, Fenson, Dale, Reznick, Reilly, & Hartung, 1994). It has been found that the early vocabularies of English-speaking children display proportionally more nouns than words in other lexical categories. From a summary of previous research shown on Table 1 (according to data extracted from Sandhofer, Smith, & Luo, 2000, and Goldfield, 2000), we see that the proportion of nouns compared to other lexical categories is universally quite large in words produced by children at early ages. For example, if we look at the famous MacArthur-Bates Communicative Development Inventories (CDI, Dale, & Fenson, 1996), in the 680 word list of words toddlers are able to produce, 53% of those words are nouns (362), 18% are action words (123), and 9% are descriptive/adjective words (63). Again, some investigators believe that this phenomenon reflects the influence of the linguistic input (such as word frequency) that children received, while others believe that children's early language acquisition has a sort of universal "Noun Bias" (Gentner, 1982; Caselli, Bates, Casadio, Fenson, Fenson, Sanderl, & Weir, 1995).

---

[1] "Closed-class words" are those words that serve to express grammatical relationships with other words within a sentence, such as pronouns, prepositions, conjunctions, auxiliary verbs. There are a relatively few and fixed number of these words in any given language and their numbers do not increase as quickly as what are known as "open" classes of words, such as nouns and verbs.

Table 1. Proportion of word types in different categories of English-speaking children's lexical production

Studies	**Nouns**	**Verbs**	**Modifiers**
Stern (1924)	78%	22%	0%
Nelson (1973)	65%	13%	9%
Benedict (1978)	50%	19%	--
Goldfield (1986)	48%	16%	--
Dale & Ferson (1996)	53%	18%	9%

In the early vocabulary development of children, beyond the common features described above, there are also large individual differences found among different children (Bates et al., 1994; also see Chapter 6 in Foster-Cohen, 1999). For example, some children start to speak much earlier than others; others begin speaking relatively late in life. Many children undergo a clear and steep vocabulary spurt, while the lexical development of a few others follows a curve that demonstrates a consistent and smoothly increasing rate (Bates et al., 1995; Ganger & Brent, 2004; Thal, Bates, Goodman, & Jahn-Samilo, 1997). These variances may reflect the differences in children's external linguistic inputs (e.g. the occurrence frequency of words) and their internal features such as personality and cognitive development (Foster-Cohen, 1999).

English might be the most widely investigated language in the field of child language development. Thus, some rules and patterns found in the vocabulary development of English-speaking children have been generalized to other languages and these patterns were often thought to be universal rules that occur across cultures. Some of

these generalizations have been widely accepted and hold up to scientific scrutiny. For example, the asymmetry between lexical comprehension and production is a phenomenon that occurs widely in all cultures (Benedict, 1979). However, debates still occur concerning the generalizability of other patterns. For example, due to the predominant proportion of nouns in the early vocabulary of English-speaking children (Gentner, 1982; Caselli et al., 1995), some investigators (Gentner, 1982) suggested that "Noun Bias" in early language acquisition is indeed a universal phenomenon that can be observed across all cultures; there are some perceptual and cognitive factors in children that support nouns over words in other categories. Specifically, Gentner proposed a so-called "*natural partitions hypothesis*." Gentner believes that there is a natural conceptual distinction between "concrete concepts" (nouns) and "predicative concepts" (verbs); and nouns are conceptually and perceptually more basic and simpler than other words, thus making nouns easier for children to grasp at a young age when compared to other grammatical categories. However, other investigators disagree with this argument, especially when the research goes deep into linguistic contexts other than English. A highly compelling counter-view rejecting the "Noun Bias" theory was given by Twila Tardif. In a series of works, Tardif and her colleagues argue that Mandarin (Putonghua)-speaking children produce more verb **tokens** and less noun **types** than English speakers. Tardif and her colleagues insist that "Nouns are not always learned before verbs" in Chinese (Tardif, 1996, 2006; Tardif, Shatz, & Naigles, 1997; Tardif, Gelman, & Xu, 1999). In some other Asian languages, such as Japanese and Korean, children are also found to be using verbs earlier than English speakers (Clancy, 1985;

Choi, 1997). The categorical terms "types" and "tokens" are two important concepts in lexical analysis. If a text is 100 words long, we say it has 100 **tokens**. However, within that text, there may be many words that are repeated, and as a result there could be only 30 different words in the text – in which case, we say there are 30 word **types**. The **type vs. token ratio** (**TTR**) is an important criterion that linguists use to evaluate lexical diversity in lexical analysis.

From the above discussions, I find that there are both similarities and differences in children's early lexical development across cultures. Investigating these similarities can help us to understand the possible universal cognitive mechanisms that underlie children's language acquisition. Investigating these differences can also help us to understand the influence that different cultures and language environments have on language acquisition. In addition, these differences and similarities can help us to address the "nature or nurture" debate that continues to arise in psycholinguistics (Pinker, 1994; Elman et al., 1996; Aitchison, 1998). Unfortunately, the majority of previous language development studies have focused on one language only, and the variations in methodologies used in these studies make it relatively difficult to make comparisons across languages and investigations. Because different investigators use different criteria for nouns in their studies, this will influence the extent to which they discover a "Noun Bias" in the languages that they investigate (see discussion in Tardif, et al., 1999). Also, as shown in the next section, there are only a handful of cross-linguistic studies of the differences in early lexical development among languages that use the same criteria to discriminate different lexical categories.

In this study, I follow the approach of previous cross-linguistic studies; and pay attention to the proportion of different lexical categories in the vocabularies of children and their caregivers at different developmental stages. My research is a corpus-based study rooted in the so-called CHILDES (Child Language Data Exchange System, MacWhinney, 2000). CHILDES is a standard computerized exchange system for the linguists and psychologists to share their child language corpora online. After more than two decades of construction, it has become the largest online database of child language in the world. In this study, I have focused on three languages: English, Mandarin and Cantonese[2]. In the future, I hope to expand this list to include other languages, such as Spanish and Japanese. My hypothesis is that, for children at different developmental stages, I would expect to see differences in their lexical compositions. Basically, the vocabularies exhibited by the children will become increasingly sophisticated and complex in direct proportion to their vocabulary diversity and concept complexity. In addition, it is expected that this will be a universal tendency across many languages. However, across different languages, there must also be differences, and the differences between languages that are similar in structure (such as Mandarin and Cantonese) will be smaller and less pronounced than those between languages that are quite dissimilar in their structures (such as Mandarin and English). Finally, the presence and absence of certain patterns (e.g. strong, weak or non-existent

---

[2] Whether Cantonese should be considered as a distinct language in the Chinese language families or as a mere dialect of the Chinese language is still an ongoing debate.

Noun Bias) in children's early vocabularies in different languages may reflect the similar features in their language input (*child-directed speech*) in different cultures.

In the remaining sections, I will first give a brief review of some cross-linguistic studies that have been conducted. I will then discuss the methodologies that I used in this research, especially the CHILDES database, including the CHAT transcript system, and the CLAN programs. From CHILDES, two sets of data were extracted. The first set is based on a small-scale but evenly balanced data; an ANOVA analysis of Noun vs. Verb ratio and a D-measure analysis of lexical diversity in children speech were conducted. The second set gives us a full-scale picture of the lexical development in the three languages. It is based on a large data set which includes all the available files within the appropriate age range (13-60 months) in the selected corpora of the three languages in CHILDES. A series of exploratory analyses were conducted on the data set, including a cluster analysis and a frequency analysis of lexical compositions of different word categories. The procedures and results of the analyses of the two studies will be discussed. The final part is the conclusion and proposal for the direction of future studies.

1. Previous cross-linguistic studies

As the original and the most famous work supporting the theory of "Noun Bias," Gentner's study in 1982 was a cross-linguistic research based on six languages: English, German, Turkish, Japanese, Kaluli and Mandarin. The author collected the lexical composition data of a total of 16 children from these languages, and came to the

conclusion that "*the proportion of nominals solidly outweighs the proportion of predicate terms.*" However, there were some methodological issues in Gentner's study which caused later theorists to rethink and challenge the "Noun Bias" theory. For example, the sample size for each language in this study ranges from two to four children, which might not be large enough to guarantee the reliability of the conclusions reached. In addition, as Tardif (1996) declared, in Gentner's study there were two different data collection methods (maternal reports and naturalistic observations) applied to different languages, instead of a single, unified method. Also, the potential biases inherent in the two methods, which are each somewhat susceptible to subjective interpretations of human observers, may have affected Gentner's final results. In an effort to limit and avoid any biases caused by data collection methods, Tardif's study of ten Mandarin-speaking children's speech examined the children's lexical use through naturalistic observations only (Tardif, 1996).

Another work supporting the "Noun Bias" was conducted by Caselli et al. (1995). This cross-linguistic study included only two languages, English and Italian; but with a much larger sample size. In particular, this study was based on the parental reported vocabulary lists of 659 English and 195 Italian infants between 8-16 months of age (at this period, children's vocabularies are relatively small, ranging from 50 to 100 words). Although there are large structural differences between English and Italian (Italian is a pro-drop language), while English is a non-pro-drop language), the authors did not find significant differences between these two languages in the emergence and growth of lexical categories. For both early lexical comprehension and production in these two

languages, common nouns (types) happen more often than words in other lexical categories, thus showing a clear "Noun Bias."

The results of the previous studies are partly verified by another cross-linguistic work based on naturalistic speech samples (Tardif et al., 1997). In this work, the authors examined adult-to-child speech and children's use of nouns versus verbs across three languages: English, Italian, and Mandarin. The speech examples were extracted from six English-, six Italian-, and ten Mandarin-speaking children and their caregivers; and the data included records of naturalistic conversations between parents and children in their homes. To avoid the problem caused by the debate of what to count as appropriate nouns and verbs (see discussion in Tardif, 1996), the authors only examined the use of common nouns (not including proper names) and main verbs (not including other predicate terms) across the three languages. The results are consistent with the studies mentioned above. Italian-speaking children, just like English-speaking children, produce more nouns than verbs in their early vocabularies; but Mandarin-speaking children produce more verbs than nouns (both in type and token). In addition, they found similar patterns in the caregivers' speech. Chinese adults emphasize verbs over nouns when they speak to their children, while English parents use more nouns than verbs. The Italian adults' lexical uses are in the middle of English and Chinese parents, although Italians still present a considerable "Noun Bias." This finding suggests that children's language input might affect their early vocabulary development. However, the authors also suggest that the relationship between children's language input and output is not so direct and causal, and might depend on many factors in the caregiver's

speech, such as the frequency, utterance position, morphological simplicity of words, and even pragmatics.

In light of the concern that many input factors may affect the children's early vocabulary composition, Tardif, Gelman and Xu (1999) further investigated the noun and verb proportions in English- and Mandarin-speaking children's linguistic input and output under different activity contexts. The conversations of 24 English children and 24 Chinese children with their mothers were recorded under three activity situations: book reading, mechanical toy playing, and regular toy playing. The overall results were consistent with previous studies; Mandarin-speaking children used fewer nouns and more verbs than their English peers. However, an ANOVA analysis demonstrated that context also played an important role in the proportion of nouns. No matter which language was studied, children's language input and output were both dominated by nouns when they were under the book-reading context; but not when they were playing with toys.

A similar cross-linguistic research was done by Choi in comparing the use of nouns and verbs in English and Korean (2000). The author looked at adult-to-child speech in English and Korean under two contexts (book-reading and toy-play). Choi found that Korean mothers tended to have a more balanced use of nouns and verbs than English-speaking mothers. In addition, as in Tardif, et al.'s study (1999), for both the English and Korean children, the speech in the book-reading context was dominated by nouns. But in the toy-play context, the Korean mother used more verbs and focused more on actions than the English counterpart.

Sandhofer, Smith and Luo (2000) also paid attention to the parents' input in English and Mandarin. They also found that the Mandarin-speaking parents tend to use more verb tokens than nouns but English-speaking parents tend to use more nouns than verbs – a result consistent with previous research. More importantly, in this work, the authors offered a new method to evaluate the caregivers' speech. Particularly, the frequency patterns of nouns, verbs and adjectives were examined. At different frequency levels, the proportion of nouns/verbs was calculated and compared. The authors drew the conclusion that English and Mandarin have similar frequency patterns for both nouns and verbs. In adult speech in both languages, nouns follow a flat distribution: most noun types are presented with a relatively low frequency; but verbs follow a steep distribution: very few verb types have a very high frequency. The flat distribution of nouns causes them to be more easily learned since different categories *"described by the common nouns are similarly organized"* (p. 578, Sandhofer, et al., 2000). Due to the highly similar frequency patterns, the authors suggested that the children in the two languages may learn the nouns in a very similar way, and the differences of noun-verb ratio in children's productions between the two languages are mainly the result of the simple truth that Chinese children hear more verbs than their English-speaking peers. Another methodological improvement in this study was that the authors drew the lexical composition of English-speaking children's vocabulary at five different developmental stages – from 11 months to 2 years 11 months in age.

2. CHILDES (Child Language Data Exchange System)

Before the widespread use of personal computers and the Internet, linguists who were interested in child language acquisition were generally forced to base their research on more traditional methods, such as biographical documents of babies' word lists (diaries), carefully recorded transcripts of children and parents' speech, etc. Although these methods have played significant roles in child language studies, they have certain obvious limitations. Chief among these is the fact that, with only a pen and notepad, even the most highly trained observer is unable to record every detail of a child's language development.

With the invention of the tape recorder, scientists that studied linguistics were suddenly able to obtain large scale data sets for several subjects at different developmental stages with relative ease. However, in the beginning, researchers had not reached a consensus as to how to transcribe, share and analyze their raw empirical databases. Due to the length limitations for publishable articles, scientists often only reported the higher-level analyses of their raw empirical data, and kept the original corpora in their own hands. Even when some linguists did have the desire to share their original data with other investigators, the fact that different researchers used different coding rules to transcribe their field data made it difficult, if not impossible, for one investigator to make use of another investigator's transcripts. In addition, without a standard coding scheme and standard data-processing techniques, investigators often discovered that they had obtained different results even though they were working from the identical, original data set (see detailed discussions in MacWhinney, 2000).

CHILDES was initially developed in 1984, and since then many researchers have tried to resolve these shortcomings by constructing a standard computerized exchange system. With affordable computer systems, data-processing and data-saving software, investigators can transcribe the original data into computer files that can be easily duplicated, modified and analyzed. With a standard coding scheme and available standard analytical techniques, investigators at different corners of the world are now able to share their corpora and understand them without difficulty. As a result, the universals and differences across different cultures and children can be found and compared using a standard criterion. Finally, with the prodigious growth of the Internet during the last decade, investigators can now easily upload and share their standardized data at the official website of CHILDES (**http://childes.psy.cmu.edu**). The CHILDES system has been a great success, and it is now the largest and most popular online collection of child language corpora. So far, the project has involved around 4,500 participating members, included around 130 corpora, and it has been used in more than 1,500 published articles (see **http://childes.psy.cmu.edu/intro/utalam.ppt**).

The basic objectives of CHILDES have been summarized and reduced to three core goals, according to MacWhinney (2000). He has suggested that, through Childes, investigators are able to:

1. automate the process of data analysis,

2. obtain better data in a consistent, fully-documented transcription system, and

3. provide more data from more kids from more ages, speaking more languages.

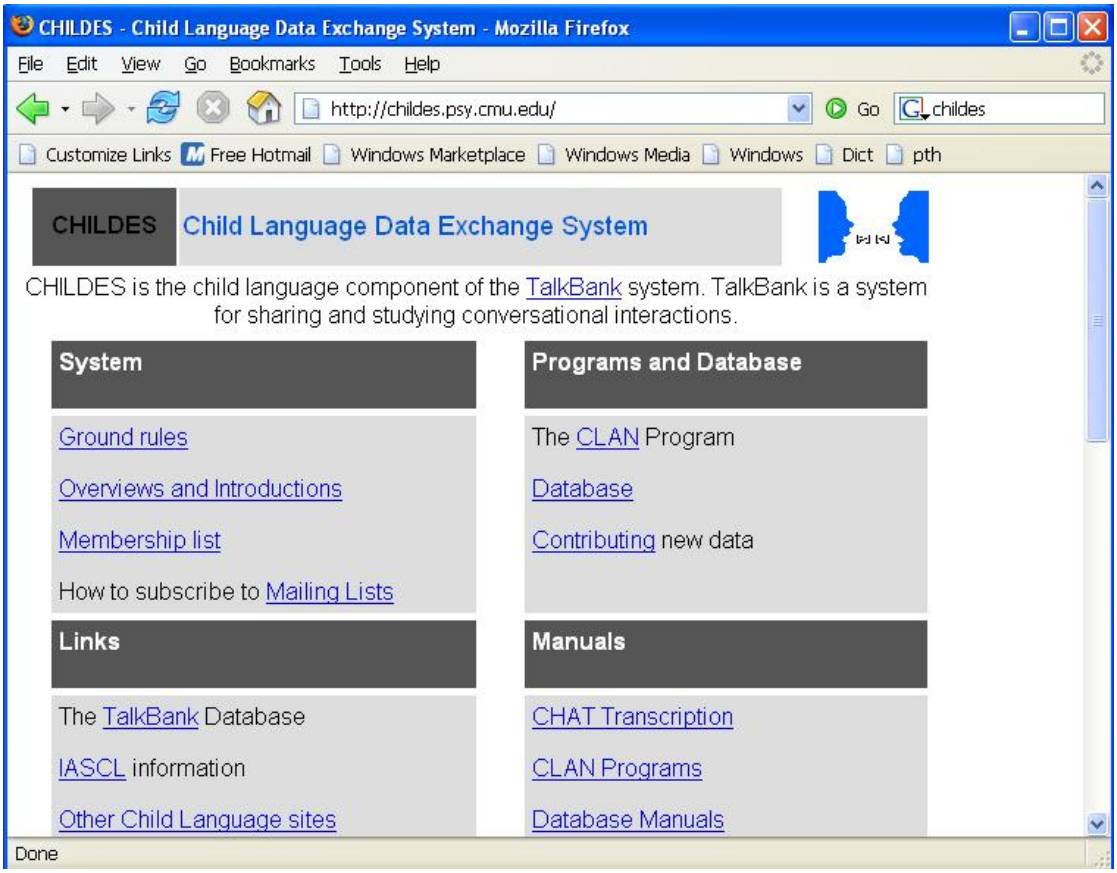

Figure 1. The homepage of the CHILDES project.

To achieve these three goals, three separate but integrated tools have been developed. A program called CLAN was developed to conduct various analyses on the corpus. A transcription and coding system called CHAT was introduced to present and save the data in a consistent format. Finally, an online database system has been developed for scientists to obtain, share and exchange useful child language data. A great feature of the CHILDES database is that it continues to expand as more scientists participate in this project.

2.1. CHAT transcription system

The standard transcription system for the CHILDES is CHAT: which stands for "Codes for the Human Analysis of Transcripts." CHAT helps linguists develop standardized and computerized transcripts of face-to-face conversational interactions (MacWhinney, 2000). CHAT can provide both basic (minCHAT) and advanced formats (midCHAT) of transcription to fulfil the requirements of researchers from the level of beginner to expert. When CHAT rules are strictly followed, the transcript can be expected to be clear, readable and easy to enter, all of which are significant features for a good corpus. A well-coded CHAT format transcript will facilitate the computerized analyses of children and adult speech. The analyses in CHILDES system are conducted by CLAN, which is discussed below.

2.2. CLAN programs

CLAN stands for the "Computerized Language Analysis." CLAN is the official analyzing tool that was specially designed to process data in the CHILDES project. Currently, CLAN includes versions for both Windows and Macintosh operating systems. The CLAN program is equipped with a graphic interface. Through the interface, on the one hand, users can input, see and modify the CHAT files; on the other hand, users can perform a lot of computerized analyses to the corpora, including frequency counts, word and phrase searches, co-occurrence analyses, and morphosyntactic analyses. Here are a few commands in CLAN that I have used in the current research.

FREQ: This command stands for Frequency Analysis. Using this command, researchers can count the number of words appearing in selected files; in addition, the ratio of different words (Types) to the total number of words (Tokens) – the so-called Type- Token ratio (TTR) of words can be reported. This ratio can help us to measure the lexical diversity of a lexical category (Sandhofer et al., 2000).

VOCD: This command can be used to calculate the D-measures of a speech sample. D-measure is a novel method to quantify the vocabulary diversity of speech samples. High values of the D-measure reflect a high level of lexical diversity. Although the calculation of D-measure is still rooted in the TTR method, it avoids the flaw of TTR that TTR value is highly related to the token numbers of samples (McKee, Malvern, & Richards, 2000). Particularly, to calculate D-measure, in a transcript, (1) many randomly sampling word chunks with varying token numbers are extracted; (2) the TTRs of these word chunks are calculated and represented as a curve of their token numbers; (3) then the software finds the best fit of this empirical curve according to a theoretical function of TTR and token size (N), and gets the value of an adjustable parameter – the D-measure, see Formula (1). Since the calculation process of D-measure adequately considers the effects of varying sample sizes, it has been proven to be a more valid and reliable measure of vocabulary diversity (MacWhinney, 2000).

$$\mathrm{TTR} = \frac{\mathrm{D}}{\mathrm{N}}\left[\left(1+2\frac{\mathrm{N}}{\mathrm{D}}\right)^{\frac{1}{2}}-1\right] \qquad \textbf{(1)}$$

MLU: This command stands for Mean Length Utterance. Using this command, researchers can calculate the number of words in a given utterance, and then calculate the

mean or average number of words that are contained within the utterances. This is a very useful tool that enables researchers to measure the average complexity of the speeches that children produce at certain developmental stages.

KWAL: This command stands for Keyword and Line Searching. Using this command, researchers are able to search the data for the words they have specified. More importantly, with KWAL, the lines (context), including keywords, will be extracted. This command is very important to current research. For example, all the utterances produced by a certain participant (e.g. the child) can be extracted and saved in a new file through the KWAL command. Therefore, this function has been a very useful tool to enable us to separate child talk and child-directed speech.

COMBO: This command is the abbreviation for Combination Search. It is a powerful tool to find the specified combination of words in an utterance and can help investigators to find certain complex string patterns. All the data that I analyzed includes the morphosyntactic information (the %mor line) of each utterance. I use this command, along with another command, **MODREP,** to discover whether a certain word (especially a word with ambiguous lexical classifications, such as *paint*) is occurring as a noun or a verb in a given utterance by the co-occurrence of the target word in the main line and its syntactic explanation in the auxiliary %mor line.

CLAN is an extremely powerful and sophisticated tool that enables us to conduct most of the analysis that we perform in computational linguistics. There are also some other effective functions in CLAN, detailed in its manual (MacWhinney, 2000).

3. Methods

In the current research, I investigated the similarities and differences of children's vocabulary development across different languages and different development stages. I looked at English, Mandarin and Cantonese; and investigated the adult-to-child speech and child speech separately. The research included two separate but related studies. The first study was based on a small but well balanced data set in the different situations for the sake of an ANOVA analysis for the noun vs. verb ratios and lexical diversity measure (D-measure) in child vocabulary. The sample size of the second study was larger than the first study. The purpose of the second study was to get a complete picture of the lexical composition in different languages. In particular, I checked certain numbers[3] of the most frequently occurring word types in the children's speech. Then the percentages of four lexical categories: **nouns**, **verbs**, **adjectives** and **others** over total vocabulary were calculated and compared across different languages as well as different developmental stages. In addition, I further determined the vocabulary composition within each lexical category (except the **others** group) by splitting each category into more detailed subgroups. Based on this detailed information, I conducted a cluster analysis to calculate the similarity among different situations.

---

[3] Around 450-500, but the number is smaller for age groups of children younger than 36 months old, due to the fact that children are not able to produce high volume of words at these younger ages

3.1. Corpora for each language

I chose English, Mandarin and Cantonese as my three target languages because each of these languages has all been widely investigated previously and there are a great deal of existing corpora for these languages that can be obtained from the CHILDES database.

The English corpora in my research were extracted from the American English database in CHILDES. This database is vast and researchers are able to use the database to uncover a great deal of data on the acquisition of English in the United States. The corpora include both longitudinal studies and cross-sectional studies investigating the speech of large numbers of children in certain activity contexts (e.g. toy-playing). Most of the corpora are related to children before school age, but a few are from research based on elementary school students. I did not include this latter category in my study because I only wanted to investigate early lexical development. My target corpora included the data sets from Bates, Bernstein-Ranter, Bliss, Bloom 1970, 1973, Bohannon, Brent, Brown, Cornell, Demetras Trevor, Demetras, Feldman, Gleason, Haggerty, Hall, Higginson, HSLLD, Kuczaj, MacWhinney, McCune, McMillan, Morisset, Nelson, New England, Peters, Post, Providence, Sachs, Snow, Suppes, Tardif, Valian, Houten, Van Kleeck and Warren-Leubecker (a total of 34 authors, MacWhinney, 2000). In total, the data included speech from more than 700 children and their caregivers. The age of the children ranged from 13 months old to 60 months old in order to match the data in the other two languages. Those files recording the speech of children out of this age rage were excluded from the analysis, as well as those files without age information.

The Mandarin data that I used was extracted from the East Asian database from CHILDES. This database is smaller than the English data, and includes Tardif's corpora (Beijing, Beijing 2, Context; Tardif, 1993; 1996), Chang (Chang, 1998), Zhou and Zhou2 corpora. It includes speech from around 300 Mandarin-speaking children and their caregivers, and I chose those children with ages ranging from 14 months old to 60 months old. Also, the data includes both longitudinal studies and cross-sectional studies.

The Cantonese data was obtained from two categories of CHILDES. As with the Mandarin data, one part is from the East Asian database, and includes the corpora of CanCorp-33, CanCorp-128 (Lee, Wong, Leung, Man, Cheung, Szeto, & Wong, 1991-1994) and HKU-70 (Fletcher, Leung, Stokes, & Weizman, 2000). The other part is from the Bilinguals database. Here, the corpus (YipMatthews) is from a detailed longitudinal study of five English-Cantonese bilingual children and their parents (Yip, 2005). Since the parents of the children followed the "one-parent-one-language" rule when they spoke to their kids, the corpus has been clearly divided into a Cantonese portion and a Chinese portion by the authors. I then extracted the Cantonese portion into our analysis. The addition of this corpus enabled me to ensure that my Cantonese study had a comparable amount of utterances with the other two languages. In total, the Cantonese data involved the conversations of around 80 children and their caregivers. The ages of the children ranged from 15 months old to 60 months old.

3.2. Different developmental age groups

To investigate the lexical development trajectory of children, it is necessary to examine their lexical composition across time. In addition, adult lexical composition across time can also be checked to help us to determine whether the children's linguistic input and output both follow the same developmental pattern. In particular, we partitioned all the transcripts into different groups according to the age of the children. In Study-I, I included four age levels: 13-24 months, 25-36 months, 37-48 months, and 49-60 months. In Study-II, we split the age range into eight groups to give us a clearer picture of the lexical development of the three languages. The eight age groups are: 13-18 months, 19-24 months, 25-30 months, 31 -36 months, 37-42 months, 43-48 months, 49-54 months, and 55-60 months. Based on this scale with six months as a unit, I obtained a rough picture of how the distributions of nouns, verbs, and adjectives change as a function of the subjects' linguistic experiences, in both children and adult speech.

3.3. Lexical categories

In Study-I, I paid particular attention to the percentages of three types of words – nouns, verbs and adjectives – over the total vocabulary size. For English, I referred to the MacArthur-Bates Communicative Development Inventories (CDI, Dale & Fenson, 1996) to classify the lexical group to which each word belongs. For Mandarin and Cantonese, those words having exact translations in English were classified according to CDI; but the words unique to the languages themselves were classified according to dictionaries and grammar books (e.g. Lǔ, 2001; Institute of Linguistics of Chinese Academy of Social Sciences, 2002).

In Study-II, I further split each lexical group into additional subcategories according to the semantic properties of these words. In this way, we were able to determine which kinds of nouns/verbs/adjectives are usually acquired early in children's language and we were able to compare the culture similarities and differences more clearly. In particular, I counted the number of words (both types and tokens) of each subcategory, and sorted them according to the descending order in each grammatical group.

Nouns: First, following Tardif et al. (1996, 1999), all proper nouns (like *London, Beijing* and people's names) in each language were excluded from my analysis. Then, according to the original definitions of English CDI, and also considering some specific features of Mandarin and Cantonese, I classified the common nouns into a total of 17 subcategories: abstract concepts, animals, body parts, clothing, color names, food and drinks, furniture and rooms, games and routines, numbers, outside things, people, places to go, small household items, toys, vehicles, words about time, and words about relative locations. Here, abstract concepts, color names, and words of relative locations were three new categories in addition to the original English CDI nominal categories

Verbs: There were a few different methods for classifying verbs in previous studies. For example, verbs can be classified differently, according to whether they refer to a physical motion that involves only an actor, an actor and patient (who receives the action), or either an actor or patient (Sandhofer, et al., 2000). In another work, Lee and Naigles (2005) categorized Mandarin verbs into seven semantic classes. I would like to

apply Lee and Naigle's classification method to my cross-linguistic study. In particular, verbs can be classified as:

(1) Basic motions: e.g. *stand, sit, open* in English.

(2) Internal feeling or communication: e.g. *love, miss, say, tell* in English.

(3) Bodily processes or care: e.g. *eat, drink, wear* in English.

(4) Creation/performance: e.g. *build, draw, write* in English.

(5) Auxiliary verbs: e.g. *may, am, are, can* in English.

(6) Light verbs[4]: like *gan4*(do), *you3* (have) in Mandarin.

(7) Others that can not be classified in the above groups.

The auxiliary verbs/ helping verbs were excluded from my analyses since they are usually classified as closed-class words without very strong semantic properties.

Adjectives: Adjectives were organized into various semantic subcategories. Following the work of Blackwell (2005), I classified the adjectives into words representing:

(1) Dimension: like *big, tall, deep* etc. in English.

(2) Color: like *red, white, pink* etc.

(3) Value: like *good nice, bad* etc.

(4) Age: like *new, old* etc.

(5) Physical property: like *heavy, soft, slow* etc.

---

[4] A light verb is a verb participating in complex predication (a V+V compound) that has little semantic content of its own, but provides some details on the event semantics. It is often thought of as having less lexical meanings (thus "light") than normal "heavy" verbs, but more semantic meanings than auxiliary verbs. English does not have many of them.

(6) Human propensity: like *crazy, happy, hungry, smart* etc.

(7) Others that can not be grouped in above categories.

3.4. Problems and solutions

There were a few technical problems encountered during the conduct of this research and a few solutions were arrived at in an effort to resolve these problems. The first problem concerned the *Hanyu Pinyin* (Chinese phonetic spelling system) codes in Chinese transcripts. *Hanyu Pinyin* is one of the most important of the Standard Mandarin Romanization systems (Yin & Felley, 1990). Using *Pinyin*, sounds of characters in standard Mandarin are capable of being represented in Roman letters. In the CHILDES database, Chinese corpora are transcribed either in Chinese characters or in the form of *Pinyin*. The benefit of *Pinyin* transcriptions in CHILDES is that the scripts can be easily shown in different computer systems without the installation of any specific fonts for Chinese characters, since *Pinyin* is written in Roman letters. However, a serious problem exists in *Pinyin* transcriptions – there are many more characters than sounds in Chinese. This means that Chinese has lots of homophones, and these homophones will necessarily be represented identically in a system that only records sound, such as *Pinyin*. For example, the sound *zuo4*[5] could mean do/make (做), sit (坐) and seat (座), *shu1* could mean book(书) and younger uncle (叔), and *he2* could mean both *and* (和) and *river* (河). When an investigator evaluates a single, isolated *Pinyin*

[5] Chinese is a typical tonal language. Numbers here indicate the tone of syllable.

"word" without the help of its context in a sentence, it is impossible for the investigator to determine which of several appropriate semantic meanings should be applied. In order to resolve this problem, for those transcripts that were transcribed in *Pinyin* codes, I rewrote the transcript using Chinese characters.

The second problem involved all three of the languages in my study. In every language, there are polysemantic words which belong to different lexical categories when presented in different contexts. For example, in English, *watch* can be either a verb or a noun, depending on the context. Similar examples uncovered in my study included *paint* (Noun/Verb), *like* (Verb/Preposition), *orange* (Noun/Adjective), and *left* (Adjective/Verb) and so on. There were also similar examples in Mandarin, such as *hua4* (画). *Hua4* (画) could be a verb that means "drawing," or a noun that means "picture." Other examples include *dao4* (到, Verb: go /Preposition: to), *shang4*(上, Noun: up position /Verb: go up), and *kai1* (开, Adj: opening/ Verb: open) et al. This problem has long been observed by linguists and contributors of CHILDES project, who in response have attempted to introduce a tool that would disambiguate the meaning of these words under different contexts. This disambiguation tool consists of two commands in CLAN programs: **MOR** and **POST**. The successful use of these automated commands depends on the previously constructed grammar database for each language. Additionally, the rules in the grammar database should be extracted from certain training samples, in which each word had been manually labelled with an appropriate grammar tag. First, investigators use the MOR command to automatically generate a %mor tier for every utterance line in a corpus. Words will be tagged in the tiers with all possible

lexical categories. Then, the Post command is used to disambiguate the %mor line, based on the pre-constructed grammar database. In this way, each word can be tagged with a unique grammatical label. Below are two examples from an English corpus (Tardif) in CHILDES.

***MOT: watch his eyes .**
%mor: v|watch pro:poss:det|his n|eye-PL . (1)

***MOT: it's a watch .**
%mor: pro|it~v|be&3S det|a n|watch . (2)

Here, we see that the words in this mother's speech were all correctly tagged with their lexical categories. In addition, the easily confused word *watch* was correctly labelled as a noun or a verb according to its corresponding linguistic contexts.

In my study, I categorized the words according to following 4 steps. (1) Most transcripts in English, Mandarin and Cantonese corpora have undergone the morphosyntactic analyses and have been added to %mor tiers, the tags of the words have been checked by the authors of the corpora. So I extracted this information and used it as the basis of my analysis. (2) For a few untagged scripts, I conducted the automatic %mor analysis (MOR and POST) on the scripts based upon the grammar database of the three languages, which could be downloaded from the CHILDES website. (3) Although the MOR command is a very powerful tool (95% or more of words can be correctly tagged in English, according to CLAN manual. MacWhinney, 2000), there were still some easily confused words that could not be correctly tagged. For this situation, I used the KWAL

command to find the words, and then tagged them manually according to the context in which these words occurred. (4) For some other easily confused words that do not have enough context information, I simply classified these words according to the most frequently used denotation of them in dictionaries (Institute of Linguistics of Chinese Academy of Social Sciences, 2002; Miller, 1990).

In the corpora, there were a large number of words that were represented by their inflectional forms, for example the plural forms or past tenses. I combined all the regular forms of a noun or a verb as a single word type. Irregular word forms were counted separately. For example, *table* and *tables* were treated as a single noun type; *work, working, worked* were treated as a single verb type; but *teach* and *taught* were counted as two separate verb types. Similarly, in Mandarin, *ma1*(妈, mother) and *ma1ma* (妈妈, mother) were counted as the same type. This was a common method employed in previous studies (see Sandhofer, et al., 2000).

4. Results

4.1. Study-I

4.1.1. Research procedure

In Study-I, I only paid attention to the child's speech in the corpora for each language as shown in Section 3.1. I chose a total of 72 files from the corpora; each file represented the conversation of one child and his/her caregivers. The ages of the 72 children ranged from 17 months to 59 months (M = 37.18, SD = 12.94). The children were split into four age groups as discussed in Section 3.2. In each age group, there

were 18 subjects: six spoke English, six spoke Mandarin, and six spoke Cantonese. Each language included 24 subjects in total. I carefully matched the age distribution of subjects in each language. Table 2 shows the mean and the standard deviation of the age of children across the four age groups in the three languages.

Table 2 Mean and the standard deviation of children ages (in months) in different age groups of Study-I

Languages		13-24 months	25-36 months	37-48 months	49-60 months
English	Mean	21.00	31.33	40.67	55.00
	SD	.89	2.73	2.16	2.83
Mandarin	Mean	21.33	30.83	41.33	50.00
	SD	.52	4.58	3.88	2.61
Cantonese	Mean	20.83	30.5	43.83	54.50
	SD	2.04	2.59	3.13	2.26

The speech of each child was extracted from the 72 files by using the "KWAL" command. I then conducted the (1) "VOCD" command to obtain the D-measure of each child and the (2) "FREQ" command to obtain the vocabulary of every child. In turn, I calculated the noun types vs. verb types (N/V) ratio for each subject. As a result, I have two dependent variables in Study-I. Since the two variables described two unrelated characteristics of child language, I used two ANOVAs (analysis of variance) for the two dependent variables separately, instead of a unifying MANOVA. Each one was a 3 x 4 analysis of variance. There were two independent variables: one had three levels, the

other had four levels. The first independent variable -- language, had three groups: Mandarin, Cantonese and English. The second independent variable was age group and it had four levels as demonstrated before. For the ANOVA for the N/V ratio, I expected to find a significant main effect for each independent variable. In particular, I expected to find that Mandarin and Cantonese speaking children would be more likely to produce verbs than English speaking children. I also expected to find that older children would produce more verbs than younger children. For the ANOVA for D-measure, I also expected to find significant differences in the lexical diversity across different age groups.

4.1.2. Results of Study-I

Table 3 Analysis of Variance for Ratio of noun types and verb types

Source of Variation	SS	DF	MS	F	Sig.
Age	11.89	3	3.96	1.55	.212
Language	19.61	2	9.08	3.82	.027*
Age x Language	13.07	6	2.12	.85	.54
Error	153.89	60	2.57		
Total	352.65	72			

* $P < .05$

A 3 X 4 ANOVA was conducted in the SPSS program to evaluate the effects of language and age on N/V ratio. As shown in Table 3, the analysis results showed a

significant main effect of language, F (2, 60) = 3.82, P = .027. However, there was no main effect for age, F (3, 60) = 1.55, P = .212. Furthermore, no interaction effect between language and age was found, F (6, 60) = .849, P = .537.

Following a significant main effect of language, an LSD post-hoc test for language at the .05 alpha level was conducted, and yielded the following effects. English speaking children displayed higher mean ratio of nouns and verbs (M = 2.19, SD = 2.71) than Mandarin speaking children (M = .98, SD = .34) and Cantonese speaking children (M = 1.23, SD = .55). There was no significant difference in the ratio of nouns and verbs between Mandarin and Cantonese speaking children.

Table 4. Mean ratio for Nouns and Verbs Changes Across Language and Age

Different Age Group (months)	**Mandarin**		**Cantonese**		**English**	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
13-24	1.15	.33	1.46	1.01	3.87	5.30
25-36	.81	.39	1.32	.30	1.86	.72
37-48	1.09	.28	1.17	.26	1.34	1.54
49-60	.86	.30	.96	.20	1.68	.74
Mean in total	.98	.34	1.23	.55	2.19	2.71

The mean ratios of noun types vs. verb types under different situations can also be found in Table 4. It is clear from these results that English speaking children use more types of nouns than verbs, a clear “Noun Bias” is found (2.19), but children in the other two language groups have relatively weak “Noun Bias” (1.23, Cantonese) or even no

"Noun Bias" (0.98, Mandarin). Mandarin and Cantonese are more similar in terms of N/V ratios. The results found here are consistent with previous cross-linguistic studies (Tardif et al., 1997; 1999). In addition, although the difference as a result of age groups is not significant, we can still find some developmental patterns from Table 3. For all three languages, when children are younger than 24 months old, they display a "Noun Bias" with more noun types than verb types. However, as the children age, the N/V ratio becomes smaller, which means that relatively more and more types of verbs have entered the children's vocabulary.

Another separate 3 X 4 ANOVA was conducted in the SPSS program to evaluate the effects of language and age difference on the D-measure, a variable representing lexical diversity. The results of the analysis (Table 5.) indicate both language and age have significant main effects on lexical diversity. For the main effect of age, $F(3, 60) = 19.92$, $P < .01$. For the main effect of language, $F(2, 60) = 9.55$, $P < .01$. However, no interaction effect between language and age was found, $F(6, 60) = 1.04$, $P = .41$.

Table 5. Analysis of Variance for D-measure

Source of Variation	SS	DF	MS	F	Sig.
Age	14042.40	3	4680.80	19.92	.000**
Language	4490.84	2	2245.42	9.55	.000**
Age x Language	1470.91	6	245.15	1.04	.407
Error	14100.79	60	235.01		
Total	223052.45	72			

** $P < .01$

Following the significant main effects of age and language, an LSD post-hoc test at the .05 alpha level was conducted, and yielded the following effects. First, English speaking children reported higher D-measure (M = 62.26, SD = 18.45) than Mandarin speaking children (M = 44.21, SD = 19.11) and Cantonese speaking children (M = 47.21, SD = 18.45). There was no significant difference in the D-measure between Mandarin and Cantonese speaking children. Second, differences between any two age groups are significant, except for the difference between 25-36 months old children and 37-48 months old children. From Figure 2, we can find that the three languages have a similar development pattern in terms of the D-measure. As time goes by, children's speech becomes increasingly diverse, and it further reflects children's development of their language competence compared to the early age groups. In addition, Mandarin and Cantonese's developmental patterns are closer on the figure, reflecting the similarity of the two languages.

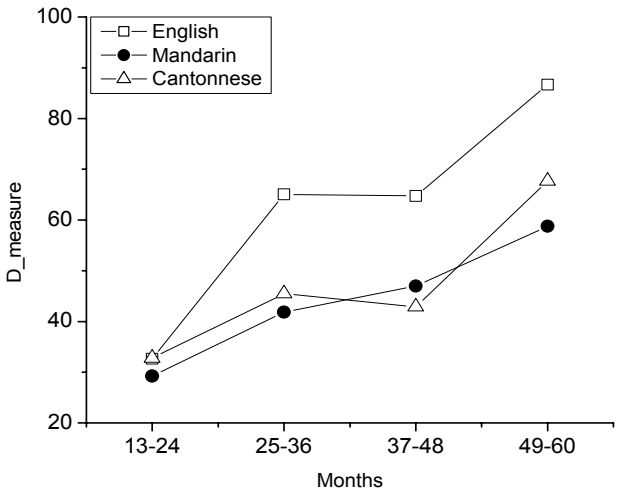

Figure 2. D-measure (lexical diversity) in different languages as a function of age.

4.2. Study-II

4.2.1. Research procedure

The purpose of Study-II is to give us a more complete picture of how languages develop across time in terms of speech complexity and lexical composition. To increase the generalizability, the sample size of the second study was larger than the first, including all the available and age appropriate data files from CHILDES as discussed in Section 3.1. To get more detailed results, I classified the data files with children's age between 13 and 60 months into eight age groups with six months as the scaling unit, as can be seen in Section 3.2. I also obtained both the child speech and adult speech sample in order to investigate the similarities and differences between the language input and output of children in each language. As a consequence in Study-II, I dealt with 48 (3x8x2: languages x development levels x people: adult/children) situations in total.

In Study-II, I first used the command of "KWAL" to extract the child speech and adult speech into separate files for each of the 24 Languages x Age group as described above. I then conducted commands of (1) "MLU" to get the mean length of utterances for each situation; and (2) "FREQ" to get the vocabularies of child and adult speech across different languages and age groups. Then, I calculated the noun vs. verb ratio (in both types and tokens) for each situation.

I also examined the lexical compositions in the vocabularies of the 48 situations. In particular, I checked certain numbers of the most frequently occurred word types in the vocabularies. Next, the percentages of three lexical categories: nouns, verbs, and adjectives over the total number of word types in each vocabulary are calculated. Then

their developmental trajectories along ages were compared across different languages. In addition, I further examined the vocabulary composition within each lexical category by splitting each category into more detailed subgroups as shown in Section 3.3. I investigated the common subcategories in each grammatical category and compared them across ages and languages. There were 30 subcategories for each vocabulary of the 48 situations. The percentages of subcategories over the total word numbers (types and tokens) can be used to describe the lexical composition of each vocabulary. Based on this detail information (treating each subcategories as a variable), I conducted a cluster analysis of the vocabularies of the 48 situations to determine the similarity and difference across language, age, and people.

4.2.2. Results of Study-II

4.2.2.1. Mean length of Utterances (MLU)

From Figure 3, we find that the mean length of utterances of children's speech in the three languages all increase as a function of the children's age (from average about 1.2 words per sentence to around 4-5 words per sentence). This indicates that children's speech becomes increasingly complex with time. In addition, along with the results of lexical diversity in Section 4.1, this gives us a picture of a gradually increasing improvement of the language ability of children across different languages. Another interesting finding from this figure is that although the complexity of adults' speech continues at a level higher than children, they also show an increasing pattern with time.

This indicates that parents tend to speak in simpler forms (shorter sentences) to their children during their early age, a characteristic of child-directed speech.

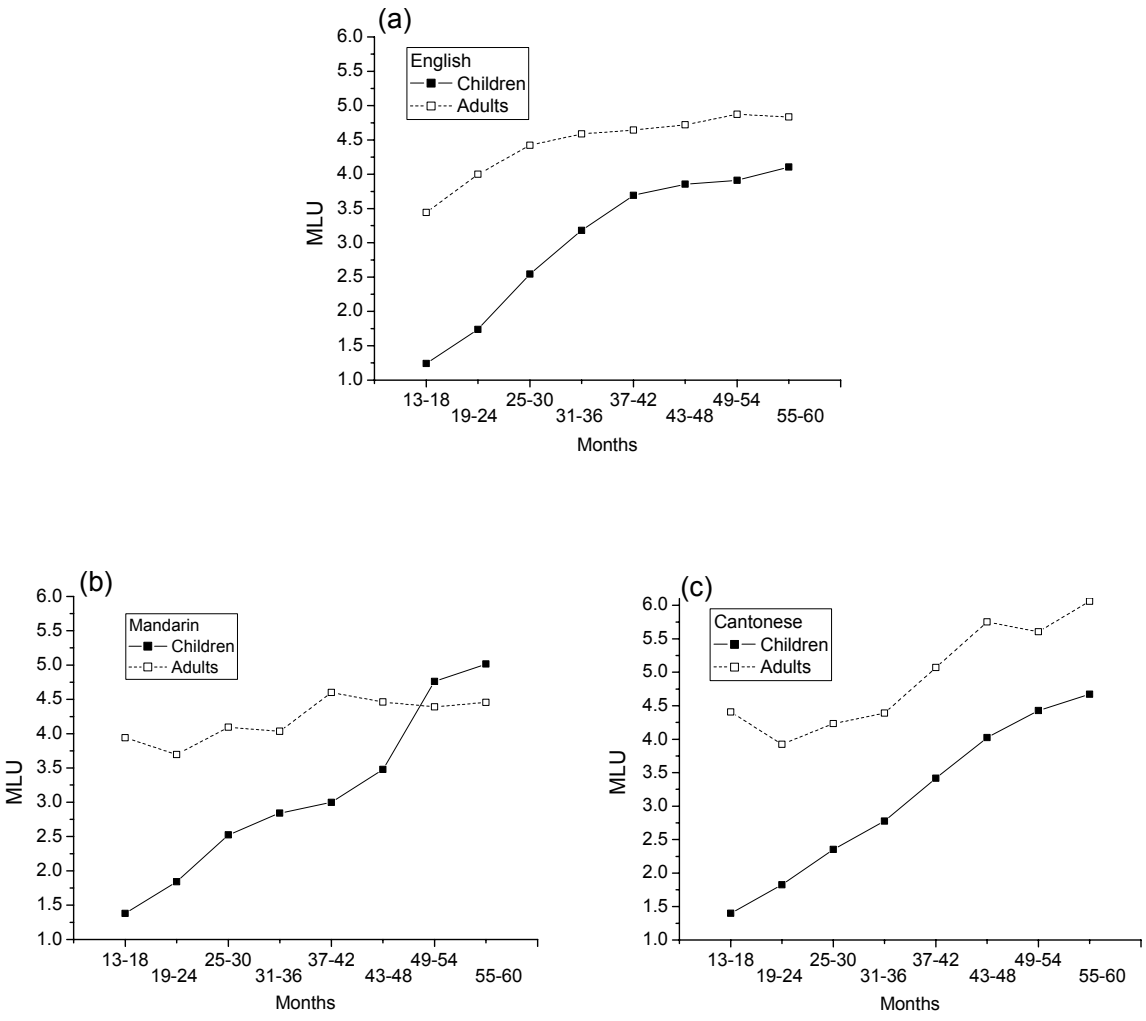

Figure 3. Children's and adults' MLU in different languages as a function of age. (a) English, (b) Mandarin and (c) Cantonese.

4.2.2.2. Noun-verb ratios (in both types and tokens)

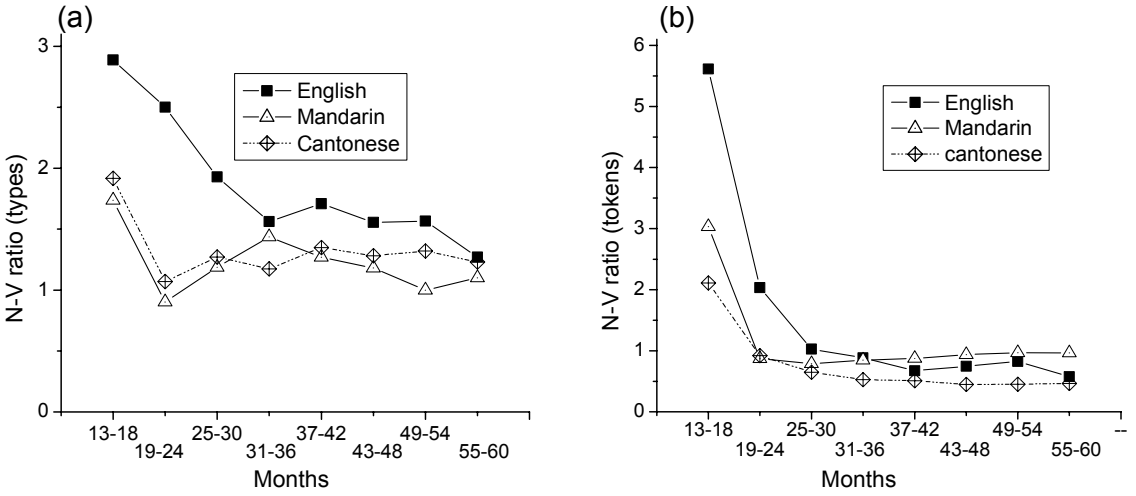

Figure 4. Children's noun-verb ratio as a function of age. (a) Types and (b) Tokens ratio.

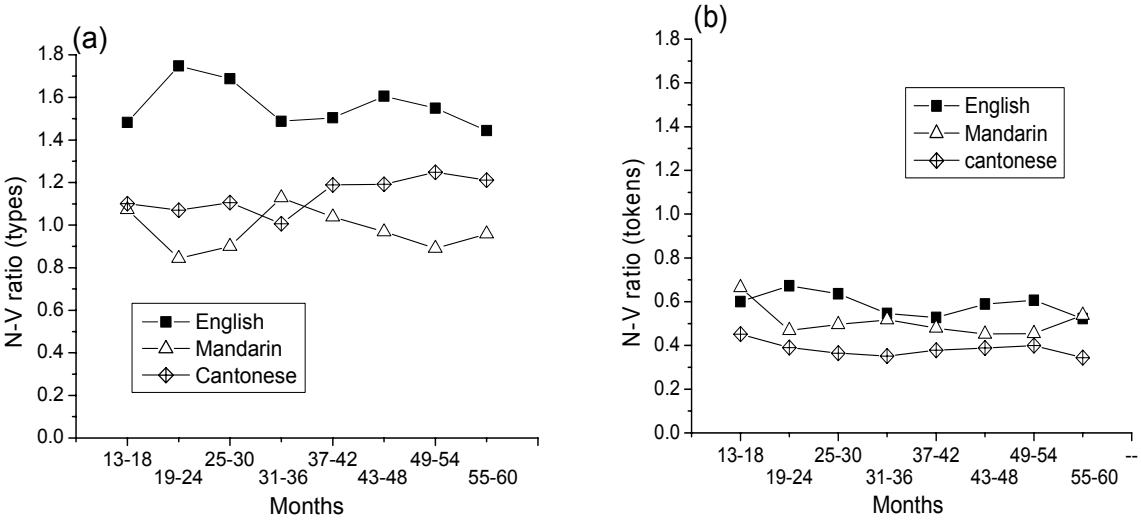

Figure 5. Adults' noun-verb ratio as a function of age. (a) Types and (b) Tokens ratio.

Following the pattern appearing in Study-I, I also calculated the noun types vs. verb types ratio in children's vocabularies, but here I also counted the word tokens.

From Figure 4, we find that the three languages follow a similar developmental pattern in the ratio of nouns vs. verbs. Whether it is calculated in terms of word types or word tokens, and which language is considered, it is clear that there are more nouns than verbs in children's vocabularies at the earliest stage – therefore a clear "Noun Bias" is discovered (n/v ratio is larger than 1 under these situations). But as children age, the "Noun Biases" become weaker by and large, approaching the level of adults vocabularies as shown on Figure 5. In addition, compared with English, the "Noun Biases" shown in Mandarin and Cantonese is much weaker; this difference can be explained by the similar pattern which is also reflected in adults' lexical composition as shown in Figure 5(a). Finally, by investigating the token ratios in Figure 4(b) (also see discussion in Section 4.2.2.3), I further found that English-speaking children have even more verb tokens than noun tokens in their older ages. Considered with the fact that the number of noun types is still larger than that of verb types in these same age groups, we can draw the conclusion that, on average, verbs occur more frequently than nouns. This result is consistent with the findings of Sandhofer, et al. (2000), which indicates that nouns follow a flat distribution: most noun types are presented with a relatively low frequency; but verbs follow a steep distribution: very few verb types display a high frequency. In addition, the superior number of verb tokens over noun tokens is a universal feature of the adults' speech across all the three languages as shown on Figure 5(b); and comparing Figures 4(b) with 5(b), we can clearly see the tendency that children's speech approaches adults' speech.

4.2.2.3. Lexical compositions (in both types and tokens)

In Figures 6, 7, and 8, I drew the percentages of nouns, verbs and adjectives over the total number of words (in both word types and word tokens) in the vocabulary of children in the three languages. The similarity and difference in the developmental tendencies of the lexical compositions in the three languages can be found. Again, the extremely large percentages of nouns at the earliest age of children clearly support an early and strong "Noun Bias" for all three languages. However, the percentages of nouns in the total vocabulary decrease with time, accompanying with the increment of percentages of verbs and adjectives. Once again, Mandarin and Cantonese each present a much weaker "Noun Bias" than English.

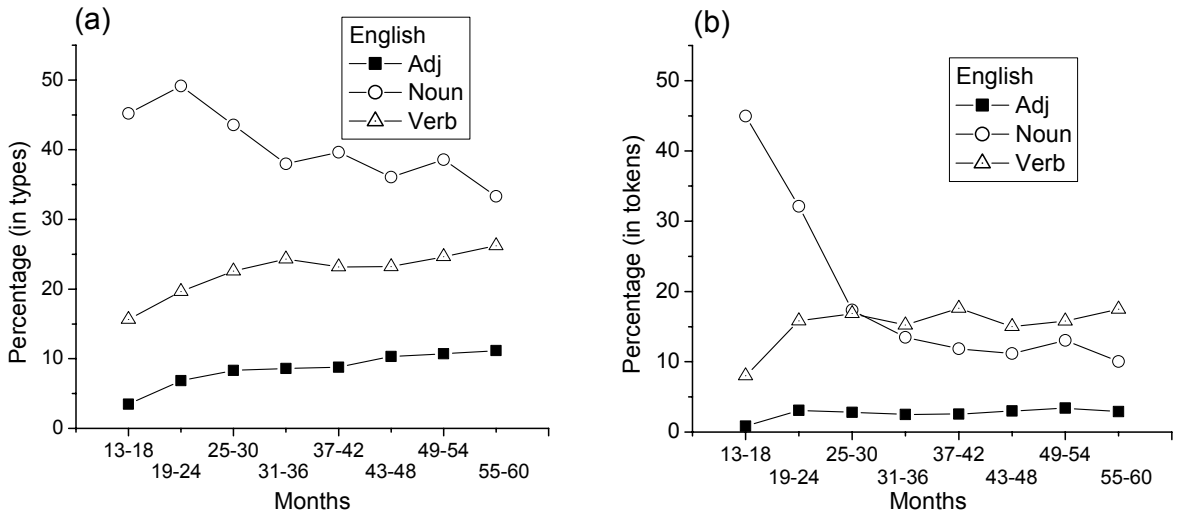

Figure 6. Percentage of nouns, verbs and adjectives at each age group in child vocabulary of English from 13 months to 60 months. Based on (a) word types, (b) word tokens.

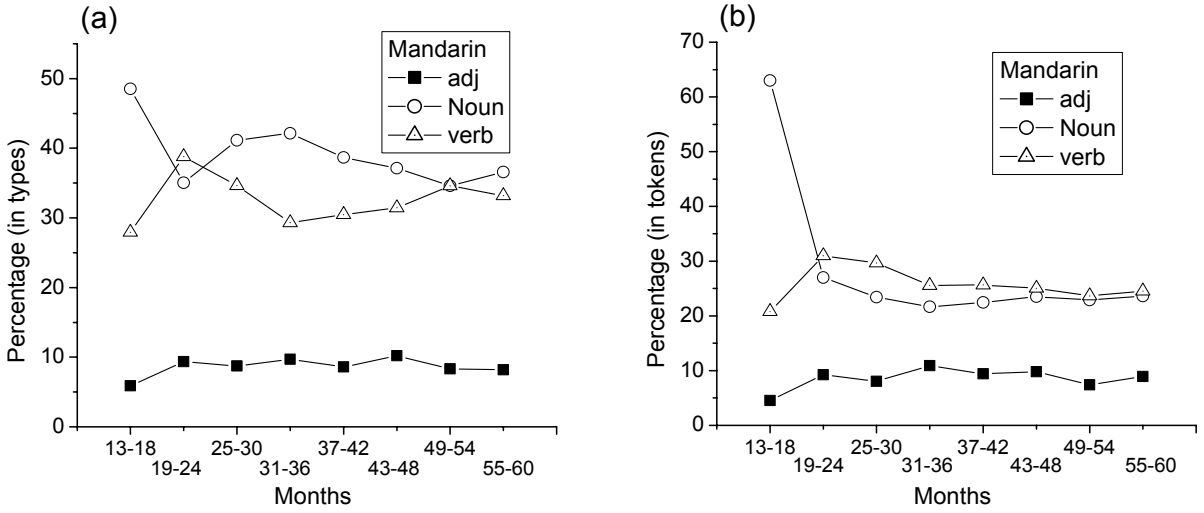

Figure 7. Percentage of nouns, verbs and adjectives at each age group in child vocabulary of Mandarin from 13 months to 60 months. Based on (a) word types, (b) word tokens.

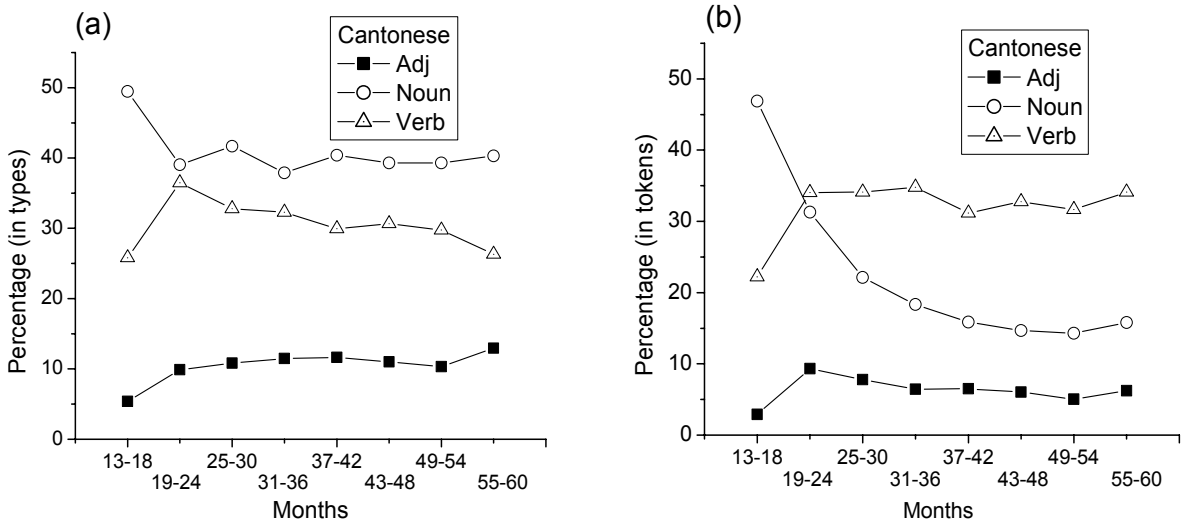

Figure 8. Percentage of nouns, verbs and adjectives at each age group in child vocabulary of Cantonese from 13 months to 60 months. Based on (a) word types, (b) word tokens.

At the end of my research range, when only word types are counted, the number of nouns is still greater than the number of verbs in three languages. Therefore, there are still certain "Noun Biases" that exist, although these may be quite weak, as shown in Figure 7-9.a. When I investigate the lexical compositions based on word tokens, I can still discover the "Noun Bias" at very early stages of children's language development. However, the percentage of noun tokens decrease dramatically when children grow up. Finally, all of the three languages have more verb tokens than noun tokens, a kind of "verb bias" although nouns still have more word types than verbs. Also, as shown in Figure 4.b and discussed in Section 4.2.2.3, this indicates that most verbs occur more frequently than nouns. This type of "flat" distribution of nouns and "steep" distribution of verbs are consistent with Sandhofer et al.'s findings (2000) based on the parents' input.

I also looked into each grammatical category to discover which subgroups of words in nouns, verbs and adjectives are most frequently used. Results indicate both similarities and differences in the type of words and word categories due to culture. For adjectives, the words referring to **dimension** (*large, small* etc.), **value** (*good, bad* etc.) and **physical properties** (*cold, hot* etc.) are among the most frequent words that children speak in the three cultures. For the verbs, the result is not surprising. The verbs referring to **motion (***run, go***)**, **internal feelings and communications** (*want, love, say*) occur most frequently. For nouns, for all three languages, words that refer to **food and drink**, **people**, **toys, animals,** and **numbers** are among the most frequent that children produce. Importantly, these are also the words that children's parents speak most often to them and the things that the children find most often in their daily lives. On the other

hand, our data also indicate differences between cultures, especially in the group of nouns. For example, in Mandarin and Cantonese, the words about **relative locations** (e.g. 上边/upside，下边/downside) and **color names** (e.g. 红色/red color) are sometimes used by children, whereas these words are not common in English. These differences reflect cultural biases in discussing the world that surrounds the child.

4.2.2.4. Cluster analysis.

Based on the detailed lexical composition, I applied a cluster analysis on the 48 situations to check the similarity among the lexicons of them. Cluster analysis is a type of interdependence multivariate statistical technique. The basic objective of this method is to identify the overall structure among a defined set of observations (here are the lexicons of the 48 situations). Particularly, cluster analysis will group observations into a few clusters, and the observations in same cluster are more similar to each other in structure than to observations in other clusters (Hair, Anderson, Tatham, & Black, 1998). The dendrogram, or cluster tree diagram, produced by the cluster analysis is shown in Figure 9. Here, I used a label with eight chars to identify each of the 48 situations: *en* as English, *ch* as Mandarin, *ca* as Cantonese; and numbers 1-8 represented age groups from 13-18 months to 55-60 months. So a label like *cachild2* represented the lexicon of Cantonese-speaking children in the 19-24 months group. This Cluster analysis was based on the Ward method. From Figure 9, I find that the different situations are by and large clustered into three large groups according to language. This means that the lexicons of people speaking the same language have similar lexical compositions. The

language factor is the most important factor that distinguishes the different situations under the child language context. In addition, the lexicons of Mandarin and Cantonese speaking people are more similar in lexical compositions, as the two clusters of the two languages are closer and attach on the same branch of the cluster tree. The lexical composition of English is different from that of the other two languages. This result is also consistent with my previous results based on ANOVA analysis of n/v ratio.

I further examined the similarity of situations occurring under the same language by doing cluster analysis on 16 situations of each language. The results show that adult lexicons often share a similar composition pattern, which differs from the lexical compositions of children. But as children grow up, their vocabularies' lexical composition becomes increasingly similar to those of their parents. For example, in the cluster tree of English situations shown in Figure 10, adults' lexicons and children's lexicons are grouped into two separate clusters. However, there are exceptions: the lexicons of children in age groups 7 and 8 (49-60 months old) share the same cluster with adults. This means that the lexical compositions of children under the two age groups are similar to the adults. Another interesting finding from Figure 10 is the situation of adult lexicon under age group 1 share the same cluster with many lexicons of children, which implies that the speech of adults to children in age stage 1 (13-18 months) has similar lexical compositions with children's speech. This result implies that adults might speak to their children in a type of "Motherese" style (Snow & Ferguson, 1977) when their children are young, while reverting to more adult speech as their children age.

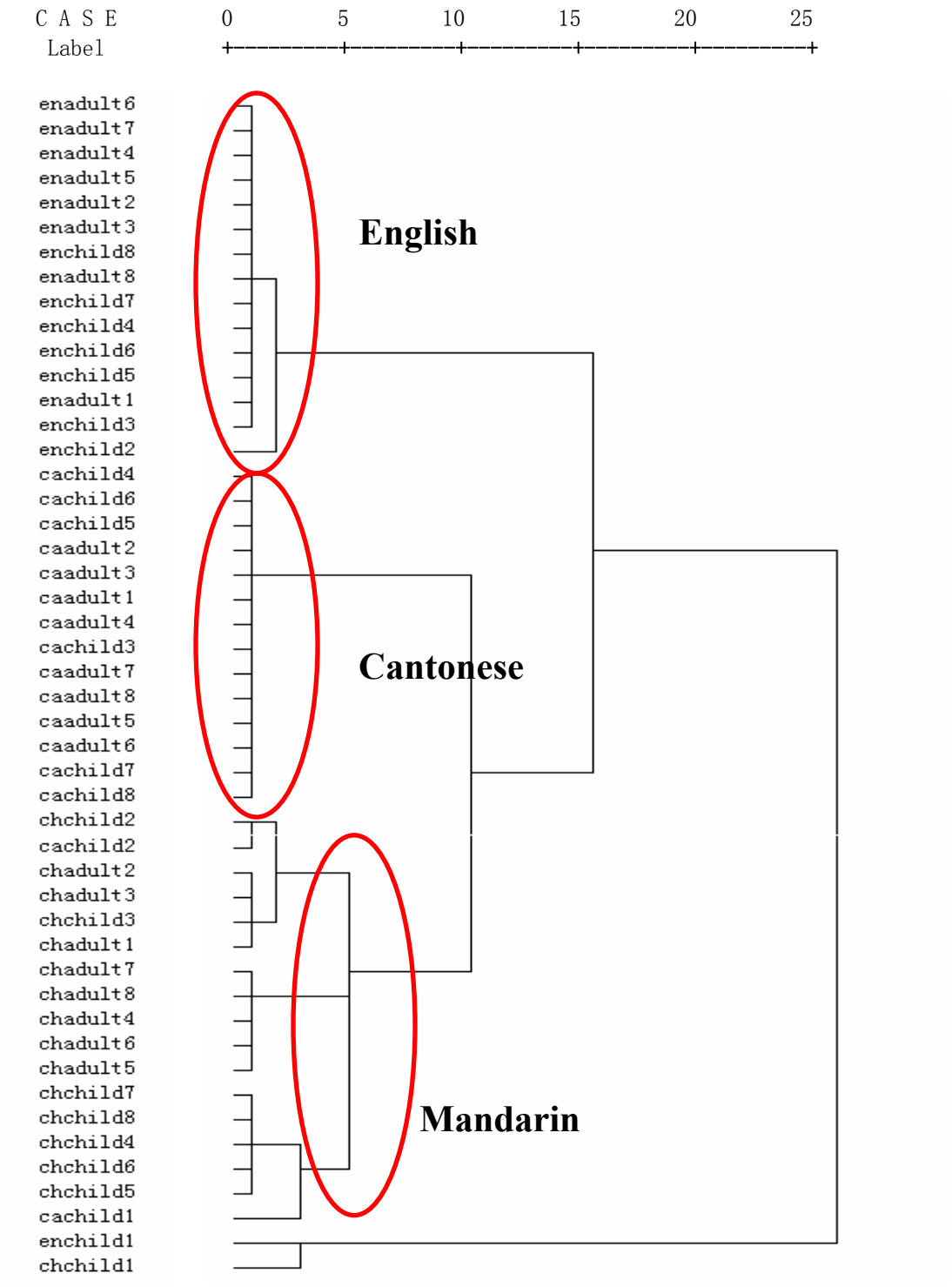

Figure 9. Cluster analysis of the similarity among the vocabularies of 48 situations across different language, age, and people. Ward method was used.

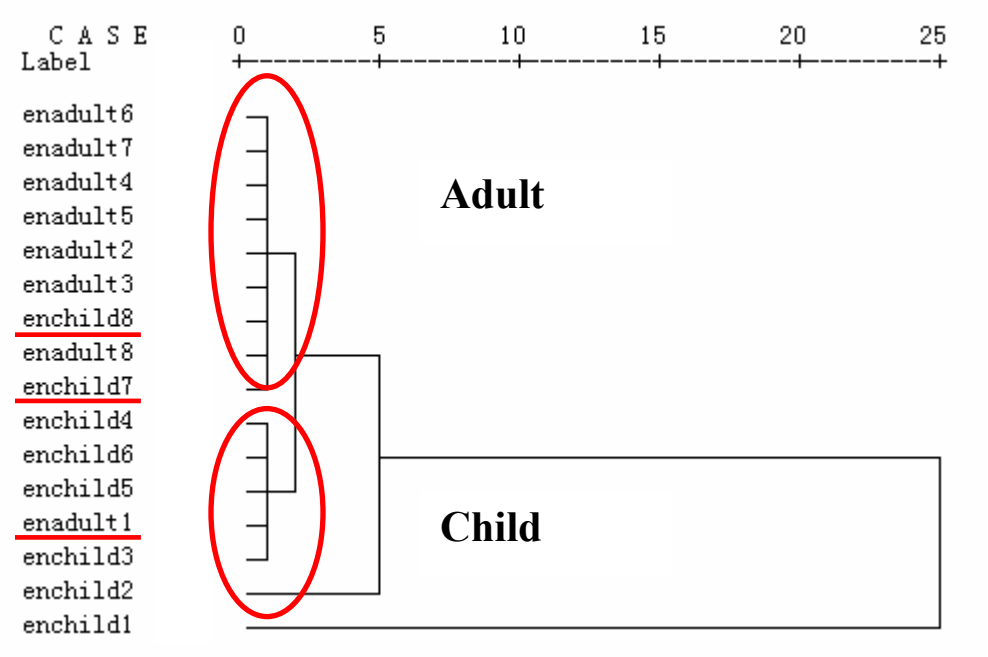

Figure 10. Cluster analysis for 16 situations in English. Ward method was used.

5. Discussions

The purpose of my thesis research has been to study the early lexical development patterns exhibited in children across different cultures and languages. In my thesis, I have hypothesised the existence of both commonalities and differences in children's early language acquisitions across different cultures. The results of my study to date clearly support my hypotheses.

5.1.Commonalities

Through the analyses of D-measure in Study-I and MLU of children's speech in Study-II, I can draw the conclusion that, no matter what culture is being analyzed, the language abilities of children undergo a gradual and incremental process of growth and

development that is characterized by an increasing lexical diversity, a greatly increased storehouse of words along with an increased competence and flexibility in the use of the lexicon, as well as increasing speech complexity, longer and longer sentence length. Our results are valid and reliable across all three cultures studied and the patterns that have been observed are fully consistent with our expectations and common sense.

Through the analyses of N/V ratio and lexical composition in the child language output across the three languages studied, I have discovered a somewhat mixed result when I tested for the existence of the highly debated, cross-linguistic phenomenon known as "Noun Bias."

The results of my study indicate that we must investigate this problem from a developmental point of view. Most importantly, my analysis showed that a universal "Noun Bias" does exist in each of the three languages studied. However, this preference of nouns over other word categories is only universal for children at their earliest stage of linguistic development, when they are capable of speaking only a few words (e.g. younger than 18 months old, as shown in Figure 4, Figures 6-8). This finding suggests that the so-called "*natural partitions hypothesis*" proposed by Gentner (1982) has a basis in empirically-quantifiable linguistic fact. No matter what culture is examined, my research shows that there are nouns in that culture that are conceptually salient and perceptually more basic and accessible than other types of words, thus making these nouns easier for the child to grasp, at an earlier stage in his or her linguistic development, than other categories of words, such as verbs. These more accessible and salient nouns are often those "referential style" words shown in Bates et al. (1995).

It is worth noted that my finding of early "noun bias" in Mandarin is not very consistent with Tardif's statement that "nouns are not always learned before verbs". This difference might be caused by the different age range in my research and in Tardif's studies (Tardif, 1996; Tardif et al., 1997; 1999). My study included the speech of children whose ages were younger than 18 months old, and it is in this age group that my studies showed the strongest noun bias in Mandarin. However, the mean ages of Tardif's studies were around 20-24 months old, and if we look at the N/V ratio of the same age group in my Study-II (Figure 4), we will find that my results actually are quite similar with Tardif's: there are almost the same amount of nouns and verbs. My results here suggest us that maybe in future empirical studies, we should consider more of the lexical compositions of those very young children.

Nevertheless, as children age and grow, their lexical composition patterns have a natural tendency to change over time. The "Noun Biases" that are observed in the child's early lexical compositions become weaker over time. This phenomenon is also common to all three languages. The reason for this weakening of the Noun Bias appears to be the fact that, as children develop linguistically, they acquire lexicons with more and more words that belong to other word categories, especially words that representing motions, ideas and events that take place in the child's daily life - complex concepts and relations that require verbs and other word categories to achieve expression. During this process, specific and unique features of the three languages begin to modulate the lexical development of the children more and more clearly. Many of the obvious differences

that are attributable to the three unique languages can be observed during the later age groups in my study.

5.2. Differences

From the ANOVA of N/V ratios in Study-I, and the cluster analysis of lexical composition in Study-II, my research clearly shows qualitative and quantitative differences that can be observed in the lexical development of children from the three different cultures and languages studied. The N/V ratios of children's vocabularies were significantly different across the three languages. Particularly, English children exhibit a significantly higher N/V ratio than children learning the other two languages, while the differences that were measured between Mandarin and Cantonese were not as significant. The same pattern can be observed on the dendrogram of the cluster analysis of the lexicons in Study-II. It is clear that the lexicons of English-speaking children at different ages are relatively close to each other, while greatly different from those of the two other languages. By contrast, Mandarin and Cantonese are not very different when represented on hierarchical cluster trees. This difference pattern is consistent with my hypothesis that the lexical development of children in languages that are similar in structure (e.g. Mandarin and Cantonese) will display similar developmental patterns.

From the lexical composition analyses in Study-II (Figures 4-5), I can find that the vocabulary distribution patterns of children evolve closer and closer to those represented in their language input. The initial "Noun Bias" decreases dramatically with time, to the point where differences in the lexicons of children speaking different

languages may or may not express a preference for nouns, depending on whether such patterns can be found in their language input. This result tells us that the final "end product" of a child's language is indeed more the product of learning and language input than a passive, purely genetic inheritance from his/her parents. The language input of children plays a very important role in sculpting and shaping language output, thus bringing about the different features in language development of children raised in different language environments.

5.3. Nature or Nurture

In recent years, nativists in language area have been extremely excited and energized, due to the discovery of a so-called "Gene of Language and Speech" – the FOXP2 gene (Enard, Przeworski, Fisher, Lai, Wiebe, Kitano, Monaco, & Pääbo, 2002). Through the investigation into the "KE" family, a family in London whose members exhibit unique difficulties with the marking of regular suffixes on verbs, investigators have determined that the mutation of a gene called FOXP2 is responsible for the family's severe speech disorders. For this reason, nativists have announced confidently that FOXP2 is the first language gene ever discovered, and that the FOXP2 gene discovery is strong evidence that our language ability is innate or "hard-wired" in our DNA.

However, as is the case with many controversies that involve language and human beings, things are not always so straightforward. In addition to the language disability identified, "KE" family members also display trouble with certain specific motor skills. For this reason, some researchers argue that FOXP2 is really not a gene that dictates

language abilities specifically, but is instead a gene that determines general motor control (MacWhinney, 2002; Elman et al., 1996). This argument finds further support in the fact that FOXP2 can also be found in many animals, rather than only humans. In addition, the language problem that the "KE" family are famous for is the difficulty in forming verb suffixes. Although verb suffixes are a specific characteristic of some western languages, they are rarely found in Chinese. This leads me to wonder what would be the result if a "KE" family member were to have Chinese as his/her native language. Will he or she still exhibit a severe language disorder? In my opinion, a language characteristic that is culture specific, such as verb suffixes, and that does not occur universally across all cultures, can not be reliably employed to answer questions such as the existence or non-existence of a language instinct. Certainly, if the language instinct exists, it would be expected to exist in all human beings, regardless of specific ethnic or genetic heritage.

My research has indicated that "Noun Bias" at the earliest developmental ages of children is a universal linguistic tendency for all three of the languages I studied. For this reason, I believe that Noun Bias may be used as a criterion for comparing and judging the nature of our human language faculty. If we are able to discover certain instances of an absolute or systematic absence of the occurence of "Noun Bias" in the early developmental speech of children from a specific family[6], and identify the specific gene responsible for causing this phenomenon, we could safely say that we have found

---

[6] For example, a case study where all children of a particular family can not produce nouns, just like the Anomic aphasia but without brain damage and only caused by genetic reasons

evidence of a language "instinct." However, in the investigations that have been carried out to date, there appears to be no such evidence of a genetically-induced Noun Bias deficiency.

Because the existence or non-existence of Noun Bias appears to be more culturally and linguistically determined than genetically determined, it would appear that the mystery that has arisen around the question of a "language instinct" may continue to confuse and confound investigators for a long time to come.

5.4. Limitations and future directions

From the previous discussions, it is clear that my research has a more complete picture of commonalities and differences of language development patterns found in children in each of these three unique cultures. This is a novel, cross-linguistic study that has yielded many interesting results. However, it is important to recognize that my research is still a preliminary study of child language development, with a few limitations. A great deal of additional research needs to be conducted to overcome and explain these limitations.

First, my research to date has included only three languages: English, Mandarin and Cantonese. Sometimes, Mandarin and Cantonese are treated as belonging to the same general "Chinese" language group. Inclusion of more languages into my research will certainly increase the validity of my results. Consequently, in the future I plan to investigate the lexical composition of additional languages in the CHILDES database, including, but not limited to, Spanish, French, and Japanese.

Second, although the corpus size of English is very large in my study, by comparison, the size of the available Mandarin and Cantonese corpuses were simply not as large as English. The smaller size of these two languages prevented me from splitting the corpus into additional age groups. As the research currently exists, I have only eight age groups with 6 months as a scaling unit. With a larger sample size, I will be able to obtain more detailed separations of age groups, and thereby obtain more precise patterns of lexical development in each of the three languages. For this reason, in the future I will try to include more samples of the speech of Mandarin- and Cantonese- speaking children into my analysis. However, this will require more collaboration and participation from my colleagues.

In conclusion, in this study I have combined a variety of statistical methods to investigate the various aspects and properties of the lexical and linguistic development of children. My research provides further insights into the area of lexical developmental patterns in children across different cultures and languages. My study is a novel attempt to identify and understand the basic mechanisms that underlie the language acquisition process in children. Finally, I hope that this exploratory study will help us to improve our understanding of the nature, origin and the very existence of the mysterious "language instinct" in human.

References

Aitchison, J. (1998). *The articulate mammal: An introduction to psycholinguistics*. London, UK: Taylor & Frances/Routledge.

Bates, E., Dale, P.S., & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher & B. MacWhinney (Eds.), *Handbook of child language.* Oxford: Basil Blackwell.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P.S., Reznick, J.S., Reilly, J., & Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language, 21*, 85-123.

Benedict, H. (1979). Early lexical development: Comprehension and production. *Journal of Child Language*, *6*(2), 183-200.

Blackwell, A. (2005). Acquiring the English adjective lexicon: Relationships with input properties and adjectival semantic typology. *Journal of Child Language*, *32*(3), 535-562.

Carey, S. (1978). The child as a word learner. In M. Halle, J. Bresnan & G.A. Miller (Eds.), *Linguistic theory and psychological Reality* (pp. 264–293). Cambridge, MA: MIT Press.

Caselli, M.C., Bates E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development, 10,* 159-199.

Chang, C. (1998). The development of autonomy in preschool Mandarin Chinese-speaking children's play narratives. *Narrative Inquiry, 8 (1),* 77-111.

Choi, S. (1997). Language-specific input and early semantic development: Evidence from children learning Korean. *The crosslinguistic study of language acquisition, Vol. 5: Expanding the contexts* (pp. 41-133). Lawrence Erlbaum Associates Publishers.

Choi, S. (2000). Caregiver input in English and Korean: Use of nouns and verbs in book-reading and toy-play contexts. *Journal of Child Language*, *27*(1), 69-96.

Chomsky, N. (1968). *Language and mind*. New York, NY: Harcourt, Brace & world.

Clancy, P. (1985). The acquisition of Japanese. *The crosslinguistic study of language acquisition, Vol. 1: The data; Vol. 2: Theoretical issues* (pp. 373-524). Lawrence Erlbaum Associates, Inc.

Dale, P.S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28*, 125-127.

Dromi, E. (1987). *Early lexical development*. Cambridge, UK: Cambridge University Press.

Enard W, Przeworski M, Fisher S, Lai C, Wiebe V, Kitano T, Monaco A, & Pääbo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature, 418,* 869-72

Elman, J., Bates, A., Johnson, A., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Fletcher, P., Leung, S. C-S., Stokes, S. F., & Weizman, Z. O. (2000). *Cantonese pre-school language development: A guide*. Hong Kong: Department of Speech and Hearing Sciences.

Foster-Cohen, S. H. (1999). *An introduction to child language development.* New York: Addison Wesley Longman.

Ganger, J., & Brent, M. (2004). Reexamining the vocabulary spurt. *Developmental Psychology, 40*, 621-632.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development: Vol. 2. Language, thought and culture* (pp. 301-334). Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldfield, B. (2000). Nouns before verbs in comprehension vs. production: The view from pragmatics. *Journal of Child Language*, *27*(3), 501-520.

Goldfield, B. A., & Reznick, J. S. (1990). Early lexical acquisition: rate, content, and the vocabulary spurt. *Journal of Child Language, 17,* 171–183.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W.C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H Brookes Publishing.

Lee, J., & Naigles, L. (2005). The Input to Verb Learning in Mandarin Chinese: A Role for Syntactic Bootstrapping. *Developmental Psychology*, *41*(3), 529-540.

Lee, T. H. T., Wong, C. H., Leung, S., Man. P., Cheung, A., Szeto, K., & Wong, C. S. P. (1991-1994). *The Development of Grammatical Competence in Cantonese-speaking Children*, Report of RGC earmarked grant 1991-94.

Li, P., Zhao, X., & MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science, 31,* 581-612.

Lǔ. S. (2001). *Eight hundred words in modern Chinese (现代汉语八百词).* Beijing, China: The Commercial Press.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (2002). Language emergence. In Burmeister, P., Piske, T., and Rohde, A. (Eds.) *An integrated view of language development - Papers in honor of Henning Wode*. pp. 17-42. Trier: Wissenshaftliche Verlag.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15*, 323-338.

Miller, G. A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3*, 235-312.

Institute of Linguistics of Chinese Academy of Social Sciences (CASS). (2002). *Modern Chinese dictionary (现代汉语词典)*. Beijing, China: The Commercial Press.

Pinker, S. (1994). *The language instinct: How the mind creates language.* New York, NY: Harper Collins Publishers Inc.

Reznick, J. S., & Goldfield, B. A. (1992). Rapid change in lexical development in comprehension and production. *Developmental Psychology, 28*, 406-413.

Sandhofer, C., Smith, L., & Luo, J. (2000). Counting nouns and verbs in the input: Differential frequencies, different kinds of learning?. *Journal of Child Language*, *27*(3), 561-585.

Snow, C. E., & Ferguson, C. A. (Eds.). (1977). *Talking to children: Language input and acquisition.* Cambridge: Cambridge University Press.

Tardif, T. (1993). *Adult-to-child speech and language acquisition in Mandarin Chinese*. Unpublished doctoral dissertaion, Yale University.

Tardif, T. (1996). Nouns are not always learned before verbs. Evidence from Mandarin speaker's early vocabularies. *Developmental Psychology, 32*, 492-504.

Tardif, T. (2006). The importance of verbs in Chinese. In P. Li, L.H. Tan, E. Bates, & O.J.L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. 1: *Chinese*). Cambridge, UK: Cambridge University Press.

Tardif, T., Gelman, S., & Xu, F. (1999). Putting the 'noun bias' in context: A comparison of English and Mandarin. *Child Development*, *70*(3), 620-635.

Tardif, T., Shatz, M., & Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of English, Italian, and Mandarin. *Journal of Child Language*, *24*(3), 535-565.

Thal, D., Bates, E., Goodman, J., & Jahn-Samilo, J. (1997). Continuity of language abilities in late-and early-talking toddlers. *Developmental Neuropsychology, 13,* 239-273.

Tomasello, M. & Slobin, D. (Eds.). (2005). *Beyond nature-nurture: Essays in honor of Elizabeth Bates.* Mahwah, NJ: Lawrence Erlbaum Associates.

Yin, B. & Felley, M. (1990). *Chinese Romanization*: *Pronunciation and orthography* (Hanyu pinyin he zhengcifa 汉语拼音和正词法). Beijing: Sinolingua.

Yip, V. (2005). "Early bilingual development in the Chinese context." In Li P., L-H .Tan, E. Bates & Tzeng, O. (eds.) *Handbook of East Asian Psycholinguistics* (Vol.1). Cambridge: Cambridge University Press.