L2 English Lexicon Acquisition in Bilingual Children: A Cantonese and Spanish Comparison

Aine Drelingyte and Axelle Gapin
Université de Lorraine, IDMC

Abstract

The aim of this study is to examine how native language structures impact the acquisition of English lexicon in children who are native speakers of Cantonese and Spanish. By providing an extensive review of relevant literature and exploring both existing and new hypotheses using the SLAtalkbank data, this research delves into the complexities of bilingual language development. The analysis focuses on the frequency of various parts of speech in the English vocabulary of these children at the ages of 1, 2, and 4. The findings highlight both similarities and distinct differences in language acquisition patterns, influenced significantly by each group's linguistic background. This study not only contributes to our understanding of bilingual lexicon development but also offers valuable insights for educational approaches in bilingual settings.

1 Introduction

The field of second language acquisition (SLA) has witnessed extensive research and inquiry into the intricate dynamics of linguistic transfer, language interference, and the impact of the first language (L1) on the second language (L2) acquisition process. While much attention has been directed towards understanding how the grammatical structures and phonological aspects of L1 may shape L2 acquisition, comparatively less spotlight has been placed on the role of L1 in influencing the lexicon, a crucial facet of language learning.
Some of the prominent theories on the L1's influence on L2 include Language Transfer (), Interlanguage (), and the Contrastive Analysis Hypothesis().

- **Language Transfer Hypothesis (LTH):** LTH focuses on the influence of L1 on L2 and posits that elements of the native language (L1) can affect L2 learning. LTH acknowledges that similar structures between L1 and L2 can lead to positive transfer, making certain aspects of L2 easier to learn. For example, if L1 and L2 share vocabulary or grammatical features, learners may find those aspects more accessible due to their similarities.

- **Contrastive Analysis Hypothesis (CAH):** CAH suggests that the similarities and differences between the structures of L1 and L2 can predict the areas where learners might encounter difficulties. It assumes that similar structures between L1 and L2 can lead to positive transfer (where L1 knowledge aids L2 learning) but also negative transfer (where L1 knowledge interferes with L2 learning), providing a different perspective into the first language transfer.

- **Interlanguage theory**, on the other hand, underscores the developmental stages that language learners progress through during their journey from L1 to L2 proficiency. As learners construct their evolving linguistic systems, the interlanguage concept highlights the continuous adaptation of language thereby providing possible insights into the dynamic nature of vocabulary acquisition.

Per the aforementioned fact, these theories have been mostly employed to explore the L1s influence on L2s grammatical structures and phonological aspects. This investigation, however, seeks to bridge this research gap by shifting the spotlight towards the intricate dynamics of lexicon acquisition during early language development. If we were to apply these theories to L2s lexicon acquisition, we hypothesize the following results:

- At an early age, children are expected to exhibit either positive or negative language transfer from their native language (Cantonese or Spanish) to English. This transfer may manifest as a higher or lower frequency of cognates,

which are words sharing similar meanings and forms, between their L1 and English lexicon. Regardless of whether the transfer is positive or negative, we anticipate significant differences in the content of vocabulary acquired by Spanish and Cantonese children due to their distinct linguistic backgrounds.

- At an early age, children's lexicon is expected to demonstrate the emergence of an interlanguage. This interlanguage is characterized by the incorporation of elements from both their L1 and English, resulting in a vocabulary that reflects a blend of both languages.

- In the cases of both positive and negative language transfer, as well as interlanguage development, we expect errors to decrease as children's proficiency in English grows. However, we anticipate that the types of errors will continue to differ significantly between Spanish and Cantonese speakers, reflecting the influence of their respective L1s.

2 Literature review

Investigations into the vocabulary expansion of children who speak only one language have been conducted for a long time. However, the focus on vocabulary development among children acquiring a second language or bilingual children has gained momentum only in the recent years. In comparison to children who speak only one language, there remains a gap of knowledge regarding how learners of a second language develop their vocabulary (Snow and Kim, 2007).

Scholars in the field of Second Language Acquisition (SLA), especially those focusing on Second Language Vocabulary Acquisition, agree that word knowledge encompasses multiple elements. However, there is a lack of consensus about the exact nature of these elements and their integration into the understanding of word knowledge. (Ma, 2009)

For instance, while research on early word comprehension among bilingual infants is limited, there has been investigation into their early vocabulary production. Numerous studies indicate that bilingual toddlers generally speak fewer words in each of their languages than monolingual toddlers do (Hoff et al., 2011). Furthermore, in situations where language exposure is uneven, these children often show a higher word production in their first language (L1) compared to their second language (L2) (HURTADO et al., 2014). However, several studies support the claim that when considering their total vocabulary across both languages, bilingual children typically match the word count of monolingual children.

The findings of (Byers-Heinlein et al.) indicated that the differences between monolinguals and bilinguals in terms of vocabulary varied based on the method of measurement for bilinguals' vocabulary. Bilinguals exhibited larger vocabularies in terms of both expressive and receptive words compared to monolinguals. However, while their receptive concept vocabularies were comparable in size to those of monolinguals, their expressive concept vocabularies were smaller. Additionally, when considering each language separately, bilinguals' vocabularies in a single language were smaller than those of monolingual speakers.

Several theories have been proposed to explain why language development differs between monolingual and bilingual children, particularly in terms of vocabulary size and ease of access. These differences could stem from varied experiences in acquiring and using languages. For example, bilinguals often learn words in environments dominated by one language, which could limit their vocabulary in the other language. When it comes to the difficulties bilinguals face in retrieving words, two main theories are prominent. The first, the weaker links hypothesis, posits that bilinguals may not access words as efficiently due to the reduced frequency of using word-concept associative networks, unlike monolinguals who have more consistent exposure to their single language (Gollan et al., 2008). The second hypothesis, the competition hypothesis, argues that bilinguals require more effort to access words in each language because they need to suppress interference from the other language, a challenge not faced by monolinguals (Dijkstra, 2005).

The differences in vocabulary size and ease of access observed between monolingual and bilingual individuals might partly stem from the unique cognitive challenges bilinguals face. For bilingual children, navigating two linguistic systems simultaneously requires cognitive flexibility but may also lead to a division in attention and exposure. This division can result in bilinguals having a smaller vocabulary in each language compared to monolinguals who focus on only one language (Bialystok, 2001). Moreover, the context of language use plays a crucial role; bilinguals often use each language

in specific domains or with certain interlocutors, which may limit the range of vocabulary encountered in each language. On the other hand, the competition hypothesis, which addresses the issue of lexical retrieval, proposes that bilinguals face unique challenges due to the parallel activation of two linguistic systems. This simultaneous activation necessitates a constant management of interference, where bilinguals must inhibit words from the non-target language to efficiently access the desired vocabulary in the target language. This added layer of cognitive processing might lead to slower or less accurate lexical retrieval in bilinguals (Dijkstra, 2005). The literature review reveals a wealth of research exploring the variance in vocabulary sizes between monolingual and bilingual individuals, yet studies focusing on the impact of a first language (L1) on second language (L2) vocabulary acquisition are notably less common. Therefore, our study shifts its focus from comparing vocabulary acquisition in monolinguals and bilinguals to examining the influence of L1 on L2 vocabulary development. We aim to conduct this investigation by analyzing bilingual speakers from two linguistically diverse language families. Previous similar research in this area, has been conducted by Uchikoshi,2014, and primarily concentrated on the overall size of the vocabulary. Our research intends to build upon this by delving into the differences in the acquisition of various parts of speech.

3 Linguistic Features of Spanish and Cantonese

This section will provide a brief overview of the features we will focus on in this paper.

Words are more than just standalone concepts; they are complex amalgamations of various concepts. A frequently mentioned illustration of this complexity is the difference in how verbs encapsulate semantic content in verb-framed languages compared to satellite-framed languages. In verb-framed languages, the verb typically bundles together the ideas of movement and path, whereas aspects like the manner of motion are often left unexpressed within the verb (Chenu and Jisa, 2009).

Chenu and Jisa research echoes the findings of Yu (1996), highlighting how similarities in language structure between English and Chinese facilitate Chinese speakers in mastering English motion verbs. Conversely, learners of Spanish, a language characterized as verb-framed, often use gestures to communicate how actions are performed, a tendency noted by Lantolf. This gesture-based communication is also prevalent among native Spanish speakers, as McNeill observed, due to the typical absence of manner descriptions in Spanish verbs. Therefore, both native and English-speaking learners of Spanish complement their verbal communication with gestures when the verbs in the language do not explicitly express the manner of action. For Spanish-speaking children learning English, this difference in language structure might present initial challenges in understanding and using English motion verbs, potentially influencing the development and composition of their English vocabulary, particularly with verbs denoting motion.

In addition, Both Cantonese-speaking and Spanish-speaking children may encounter challenges when learning English pronouns, but the nature of these challenges may differs due to the linguistic features of their native languages. Spanish, being a pro-drop language with gendered and number-inflected pronouns, means Spanish-speaking children are accustomed to the concept of pronouns, but might initially omit them in English due to verb conjugation indicating the subject. The primary adjustment for them lies in the frequent use of subject pronouns in English. In contrast, Cantonese-speaking children, coming from a language where pronouns are often dropped and neither gendered nor tied to verb conjugation, face a more significant challenge. They need to learn entirely new concepts for English, such as gender-specific pronouns and subject-verb agreement, which are absent in Cantonese. Consequently, while both groups might initially omit pronouns, we hypothesize that the learning curve may be steeper for Cantonese-speaking children due to these additional linguistic features they need to acquire such as gender specific pronouns.

Furthermore, the handling of singular and plural nouns in Spanish and Cantonese presents distinct challenges for speakers of these languages when learning English. Spanish, like English, marks nouns for plurality, typically by modifying the ending (e.g., "libro" for "book" and "libros" for "books"), a grammatical feature that Spanish-speaking children are already familiar with. This similarity suggests that Spanish-speaking children may more easily grasp the concept of pluralization in English, adapting to its specific rules and exceptions. In contrast, Cantonese does not inflect nouns

for number; plurality is usually inferred from context or indicated through the use of quantifiers and classifiers. Therefore, Cantonese-speaking children might find the concept of pluralizing nouns in English particularly challenging. They must learn not just the basic rule of adding -s or -es to create plural forms but also understand the numerous irregular plural forms in English. This difference makes the acquisition of plural noun forms in English a new and potentially more complex grammatical concept for Cantonese speakers compared to their Spanish-speaking counterparts.

These are just a few differences in linguistic structure between Cantonese and Spanish that we hypothesize could result in an L1 language transfer. Further cases will be discussed along with result presentation.

4 Study and Results

4.1 Datasets

The dataset used for Spanish-English is called "CHILDES Spanish-English FerFuLice Corpus" (Liceras et al., 2008) and the research was conducted by Raquel Fernández Fuertes and Juana Liceras. The dataset involves two identical twins who were given the pseudonyms Simon and Leo. They were born on December 28, 1998 in Spain. The twins were raised in a middle-class family. Their father is a native speaker of Spanish, and their mother is a native speaker of American English. The children are bilingual, with the father using Spanish and the mother using English with them. The parents usually communicate in Spanish together. What is dealt with is individual bilingualism: bilingual English/Spanish first language acquisition in a monolingual-Spanish social context. Throughout their study, the twins' language development was observed from 1 year old and 1 month to 6 years old and 11 months. Their family primarily communicated in Spanish, except during summer visits to the United States, or when someone exclusively speaking English was with them. The twins attended a Spanish-speaking daycare from the age of 1 year old and 10 months, with additional exposure to English during visits and summers. The recordings were made during natural play activities at home.

The dataset "CHILDES Cantonese-English Yip/Matthews Corpus"(Yip and Matthews, 2007) focuses on the language acquisition of five children exposed to both Cantonese and English from birth.The research was conducted by Virginia Yip, Stephen Matthews and Huang Yue-Yuan. The participants include Alicia, Charlotte, Darren, Janet, Kasen, Kathryn, Llywelyn, Sophie, and Timmy. Each child's language development is documented through regular audio and video recordings, transcripts, and additional data. Alicia, born in Hong Kong, attended both Chinese and English kindergartens from the age of 2 years old and 3 months. Despite the one parent-one language pattern, Alicia interacted in both Cantonese and English with her siblings.* Charlotte was born in Hong Kong in 1996 and exhibited a dominance in English during her bilingual development. Her parents are a teacher mother native in Cantonese and a UK professor father, who followed the one parent-two language strategy. Darren has native Cantonese-speaking parents and showed a relatively balanced language development from 1 year old and 7 months to 3 years old and 11 months. Some Code-mixing (where English words were incorporated into Cantonese sentences) was observed in parental speech. Janet, born to a Cantonese-speaking mother and British English-speaking father, exemplified childhood bilingualism with an extended silent period in English. Her Cantonese development outpaced English due to an input imbalance, with more production in English starting at 2 years old and 9 months. Kasen, who was exposed to both Cantonese and English from birth, exhibited distinct periods of language development. His preference for English between 2 years old and 10 months and 3 years old and 4 months was influenced by his enjoyment of learning in English at preschool. Kathryn, born in Hong Kong to a Cantonese-speaking father and English-speaking mother, attended an international kindergarten in Cantonese from the age of 2 years old and 7 months. Unlike her siblings Timmy and Sophie, Kathryn showed a balanced pattern of language development.* Llywelyn, who was born in 1993, demonstrated some phonological influence from English in his early Cantonese.His brother Kenny, advanced in language and cognition, played a significant role in Llywelyn's language ecology. Sophie, born in 1996, showed a lively and talkative personality during recordings from 1 year old and 6 months to 4 years old. Influenced by her grandmother's Chaozhou dialect, Sophie developed passive knowledge of the dialect. Timmy, the first-born in Hong Kong, had exposure to both Cantonese and English

from birth.Despite the one parent-one language principle, he had more Cantonese than English input in his first three years.

4.2 Methodology

To get all the information we needed from our two chosen datasets and analyze it, we had to develop a specific code. In this section, we'll outline the steps we took in coding to extract and process the data for our study. In the initial phase of data preprocessing, we extracted lines pertinent to children's speech, eliminating enclosed phrases through regular expressions. Following this, we tokenized the refined phrases into individual words using NLTK. To ensure linguistic accuracy, we filtered out non-English words with the assistance of language detection and a set of English words from NLTK's corpus. For a global overview of the dataset, we merged the tokenized data and employed word clouds to visually capture the most frequent words. This approach offered us an intuitive snapshot of the linguistic landscape. Then, we applied POS tagging to the tokenized words using NLTK's POS tagger, enabling a complete analysis of the syntactic structure in children's language. We also conducted frequency analysis on both tokens and POS tags, which revealed prevalent words and parts of speech across the datasets. Visualizing the most common tokens in Spanish involved the creation of word clouds, which facilitated the comparative analysis between both datasets. Additionally, we refined our approach for Cantonese/Spanish and English token separation, enabling a more nuanced exploration of language-specific patterns. The creation of word clouds and frequency analyses offered valuable insights into the predominant words, syntactic structures, and language patterns in both the Cantonese-English and Spanish-English datasets.

Furthermore, we have applied ANOVA (Analysis of Variance) statistical test to look into the significance of parts-of-speech frequency variations between different age groups and L1 groups. We have applied the test to explore the significance of variations between types of verbs, nouns, and pronouns because of the syntactical variation described in section 3.

The ANOVA test was selected based on its ability to compare means across multiple groups effectively and simultaneously. This is crucial in our study, as we are comparing more than two groups (different age groups and language backgrounds). ANOVA helps in determining whether any statistically significant differences exist in the frequency of usage of various parts of speech across these groups. Unlike t-tests, which are generally limited to comparing two means, ANOVA allows for the comparison of means across multiple groups, making it more suitable for studies like ours with multiple independent groups. Additionally, ANOVA can handle complex experimental designs and interactions between variables, which is pertinent given the multifaceted nature of our research that involves age, first language, and parts of speech. This makes ANOVA a robust and appropriate choice for analyzing the data in our study, providing a comprehensive understanding of how syntactical variations influence language acquisition in bilingual children.

4.3 Results

At the outset of examining our results, it's evident that the lexical development in bilingual children exhibits distinct patterns based on age and linguistic background. Cantonese-speaking children show a gradual increase in unique word usage, starting with 251 unique words at age 01, progressing to 434 words by age 02, advancing to 678 by age 04. Conversely, Spanish-speaking children present a different trajectory, with 138 unique words at age 01, escalating to 751 by age 02, and culminating at 1072 by age 04. These figures not only highlight a robust increase in vocabulary with age but also suggest substantial differences in the lexical acquisition rate and diversity between the two language groups. The higher count of unique words in older Spanish-speaking children could point to linguistic features shared between Spanish and English posing less difficulty to acquire.

4.3.1 Verb Usage

The ANOVA analysis conducted on the dataset reveals significant insights into the impact of age and language background on verb usage. The results indicate a pronounced effect of age on the count of verbs, as evidenced by a low p-value (0.000261) in the age group factor. This statistically significant finding suggests that the frequency of verb usage varies considerably across different age groups. However, when it comes to the language group (e.g., Cantonese vs. Spanish), the analysis does not show a significant effect on verb count, as denoted by the higher p-value (0.186869). This implies that the language background does not play a substan-

tial role in influencing verb usage patterns. Furthermore, the interaction between age and language groups also appears to be non-significant (p-value: 0.086412), indicating that the age-related differences in verb usage are consistent across different language backgrounds. The absence of a significant interaction effect suggests that the influence of age on verb usage is not contingent on the language spoken by the children. The warning about the covariance matrix not having full rank, however, points to potential complexities in the data or the model, which warrants further investigation for a more nuanced understanding of these relationships.

4.3.2 Noun Usage

On the other hand, the ANOVA results for nouns in the dataset provide insightful observations about the influence of age and language background on noun usage in children. The analysis reveals that age has a marginally significant impact on the number of nouns used (p-value: 0.051755). This suggests that as children grow older, there might be a trend towards an increase in the variety or frequency of nouns they use, although this trend is not strong enough to be conclusively deemed significant at the conventional 0.05 level.

Interestingly, the impact of the language group (distinguishing between Cantonese and Spanish) is more pronounced, with a p-value of 0.039587. This indicates a statistically significant difference in the number of nouns used between these language groups, implying that language background plays a notable role in shaping noun usage patterns in children.

However, the interaction between age and language group does not show statistical significance (p-value: 0.708594). This means that the difference in noun usage across age groups is relatively consistent regardless of the language background, indicating that the age-related development in noun usage is not specifically influenced by whether the child is a Cantonese or Spanish speaker.

4.3.3 Pronoun Usage

The findings demonstrate a significant effect of age on pronoun usage, as indicated by a notably low p-value of 0.000104. This significant result under the age factor suggests that as children grow, there are meaningful changes in how they use pronouns. This could be reflective of their developing language skills and cognitive abilities, where older children show a more sophisticated or varied use of pronouns compared to younger ones.

On the other hand, the language group (Cantonese vs. Spanish) does not exhibit a statistically significant impact on pronoun usage, with a p-value of 0.312263. This indicates that the differences in pronoun usage are not significantly influenced by whether the child is a native Cantonese or Spanish speaker. Such a finding implies that the developmental trajectory of pronoun usage is somewhat universal across these language groups, not significantly swayed by the specific linguistic environment.

The interaction effect between age and language group also does not reach statistical significance (p-value: 0.128941), suggesting that the age-related patterns in pronoun usage do not significantly differ between Cantonese and Spanish speakers. This lack of a significant interaction effect further supports the notion that the fundamental aspects of pronoun acquisition and usage in children are similar across different language backgrounds.

5 Discussion and Conclusion

The comprehensive analysis of verbs, nouns, and pronouns usage in the dataset provides valuable insights into the language acquisition patterns of Cantonese and Spanish children learning English. The results partially align with and offer nuanced perspectives on the proposed hypotheses regarding language transfer, interlanguage emergence, and error patterns in bilingual language development.

- **Language Transfer and Vocabulary Content:** The results indicate a significant influence of age on the usage of verbs and pronouns, suggesting developmental changes as children grow older. However, the impact of language background (Cantonese vs. Spanish) is more nuanced. While it significantly affects noun usage, it does not show a similar influence on verbs and pronouns. This finding partially supports the hypothesis that children exhibit language transfer from their native language to English, as it manifests in nouns but not uniformly across other parts of speech. The distinct linguistic backgrounds of Spanish and Cantonese seem to play a role in shaping noun usage, potentially reflecting differences in cognate frequency between these languages and English.

- **Interlanguage Development:** The absence of

significant interaction effects between age and language group in the usage of verbs, nouns, and pronouns may support a common developmental trajectory in bilingual language acquisition, regardless of the native language. This observation aligns with the concept of interlanguage, where bilingual children's vocabulary reflects elements from both their native language and English. The similar patterns of language development across both language groups indicate a blending of linguistic influences, which is a hallmark of interlanguage formation.The notion of this is also supported by the abundance of code switching by the children which is discussed in the dataset section.

- **Development with age:** The significant age-related changes in language usage imply that as children's proficiency in English grows, the nature and frequency of errors evolve. This is consistent with the expectation that errors decrease with increasing English proficiency. However, the distinct patterns observed in noun usage between the two language groups suggest that the types of errors or linguistic challenges faced by children might still reflect the influence of their respective native languages. The significant difference in noun usage between Cantonese and Spanish speakers, but not in verbs and pronouns, might indicate specific areas where language transfer and L1 influence are more pronounced.

In conclusion, the findings underscore the complex interplay between age, native language background, and language acquisition in bilingual children. While there is some support of language transfer hypothesis, this phenomena manifest differently across various lexical categories, reflecting the intricate nature of bilingual language development. The results highlight the importance of considering both age and native language influences in understanding how bilingual children acquire and use vocabulary in a second language.

References

Ellen Bialystok. 2001. Bilingualism in development: Language, literacy, and cognition.

Krista Byers-Heinlein, Anne M. Gonzalez-Barrero, Elizabeth Schott, and Helen Killam. Sometimes larger, sometimes smaller: Measuring vocabulary in monolingual and bilingual infants and toddlers. *First Language*, 0(0):1–18.

Florence Chenu and Harriet Jisa. 2009. Reviewing some similarities and differences in l1 and l2 lexical development. *Acquisition et interaction en langue étrangère*, 1(lia 1).

Ton Dijkstra. 2005. Bilingual visual word recognition and lexical access. In Judith F. Kroll and Annette M. B. de Groot, editors, *Handbook of Bilingualism: Psycholinguistic Approaches*, pages 179–201. Oxford University Press.

Tamar H Gollan, Rosa I Montoya, Cynthia Cera, and Tiffany C Sandoval. 2008. More use almost always a means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, 58(3):787–814.

Erika Hoff, Cynthia Core, Silvia Place, Rosario Rumiche, Melissa Señor, and Marisol Parra. 2011. Dual language exposure and early bilingual development. *Journal of child language*, 39:1–27.

NEREYDA HURTADO, THERES GRÜTER, VIRGINIA A. MARCHMAN, and ANNE FERNALD. 2014. Relative language exposure, processing efficiency and vocabulary in spanish–english bilingual toddlers. *Bilingualism: Language and Cognition*, 17(1):189–202.

Juana M. Liceras, Raquel Fernández Fuertes, Silvia Perales, Rocío Pérez-Tattam, and Kate T. Spradlin. 2008. Gender and gender agreement in bilingual native and non-native grammars: A view from child and adult functional–lexical mixings. *Lingua*, 118(6):827–851.

Qing Ma. 2009. *Second Language Vocabulary Acquisition*.

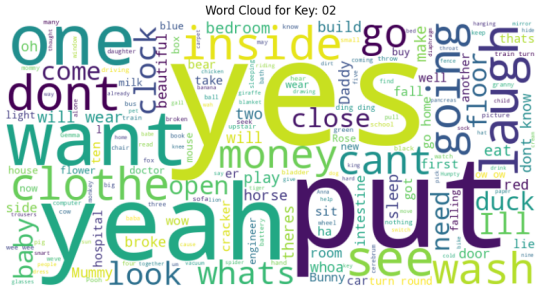
Catherine Snow and Young-Suk Kim. 2007. Large problem spaces: The challenge of vocabulary for english language learners.

Yuuko Uchikoshi. 2014. Uchikoshi, y. (2014). development of vocabulary in spanish-speaking and cantonese-speaking english language learners. applied psycholinguistics. *Applied Psycholinguistics*, 35:119–153.

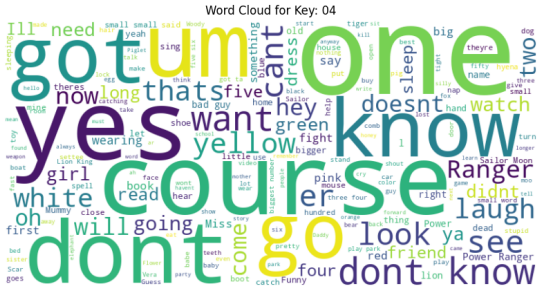
Virginia Yip and Stephen Matthews. 2007. The bilingual child: Early development and language contact.

6 Appendix

(a) Most frequent words produced by Cantonese children at the age of 01.



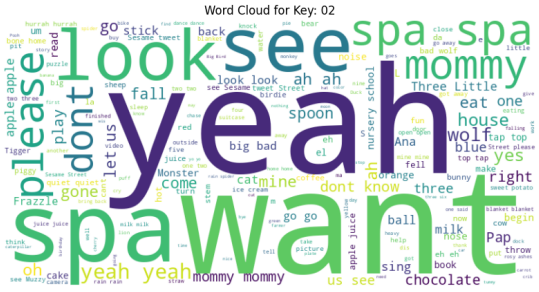
(b) Most frequent words produced by Cantonese children at the age of 02.



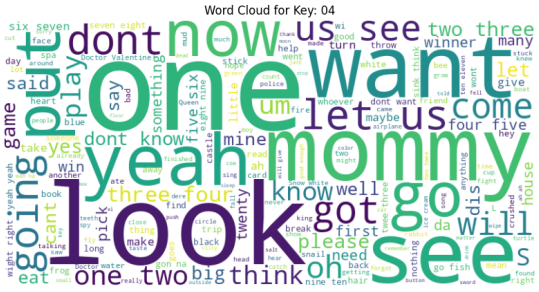
(c) Most frequent words produced by Cantonese children at the age of 04.

Figure 1: Frequent words produced by Cantonese children at various ages.

(a) Most frequent words produced by Spanish children at the age of 01.



(b) Most frequent words produced by Spanish children at the age of 02.



(c) Most frequent words produced by Spanish children at the age of 04.

Figure 2: Frequent words produced by Spanish children at various ages.