Spring 4-25-2022

Implications of Metalinguistic Awareness on Early Childhood Word-Referent Mapping

Zinan Yang
Macalester College, zyang1@macalester.edu

Implications of Metalinguistic Awareness on Early Childhood Word-Referent Mapping

Zinan Yang
Department of Linguistics, Macalester College

Honors Project
Thesis Advisor: Kevin Schaefer
Committee Members: Christina Esposito, Brooke Lea
Date Submitted: 04/25/2022

Abstract:

Children exhibit word-referent learning mechanisms like statistical learning (SL) proposed by Yu and Smith (2007) and propose-but-verify (PBV) mechanisms proposed by Medina (2011), but prior work has yet to investigate how children develop metalinguistic awareness within these two approaches. To evaluate the differences in corpus data predictions of the SL and PBV mechanisms, this study proposes a learning bias: an Extralinguistic Reference Bias. Statistical learning predicts a constrained trajectory of children's development of metalinguistic awareness. Children younger than approximately age 5 have limited access to metalinguistic language use while they are engaged in the initial mapping of forms to their primary, extralinguistic meaning. Children will acquire *metalinguistic* language use only after first understanding *extralinguistic* reference through word-referent learning mechanisms.

Using data from the Child Language Data Exchange System (CHILDES), this corpus-linguistic study coded every token of four metalinguistic verb lemmas (*say, ask, tell, talk*) across all corpora of mainstream North-American English's varieties with random forest classification. Prior to age 5, if children have limited access to metalinguistic reference as suggested by Piaget (1928) and Vygotsky (1962); their metalinguistic verb use is in fixed constructions that refer to speech acts, like *say cheese,* rather than reported speech. Furthermore, some of the isolated tokens that appear to be reported speech are instances of children's imitations of parents' modeled speech.

Additionally, the development of metalinguistic awareness is different within the SL and PBV approaches. For the SL mechanism, children would disregard any forms without an observable extralinguistic referent; whereas for the PBV mechanism, children would produce seemingly metalinguistic tokens without observable extralinguistic correlates.

Content

1. Introduction

Children's ability in pairing sounds with specific meanings is staggering. According to Bloom (2002), children learn at least ten words a day to reach the English-speaking adult vocabulary capacity, which is sixty thousand words. Two theories, statistical learning (SL) (Saffran et al., 1996) and "propose-but-verify" (PBV) (Medina et al., 2011), explain people's mechanisms of mapping words with meanings with a statistical approach - they differ in the degree of reliance on the statistical properties of the input.

SL suggests that when perceiving an unfamiliar linguistic signal, people will show gradual learning, making multiple hypotheses of the meanings, then gradually converge on a single meaning across learning instances as the word in various contexts provides more information (Trueswell et al., 2013). In contrast, the PBV mechanism claims that people retain one hypothesis of the meaning when hearing a word or seeing a sign using fast-mapping, i.e., the selection of a single candidate word-referent mapping from limited information in the input. If the meaning of that linguistic signal is confirmed across trials to be correct, people will hold on to that hypothesis; but if the hypothesis is refuted, people will make a new one. SL and PBV provide contrasting predictions on children's word-referent mapping mechanisms. The PBV mechanism allows children to make errors in word use and then to collect negative evidence from the response (of caregivers or, in the lab, the experiment computer interface), whereas the gradual learning of SL does not allow for active attempts with ambiguous statistical properties. *

Prior work in word learning focuses on mapping the words to the extralinguistic referents but not metalinguistic ones. The current thesis define extralinguistic referents as meanings that have correlates that are observable in the external world ("outside of language") for the subjects to map to, whereas the metalinguistic referents are those not visually accessible to subjects. For example, Yu and Smith (2007) uses visual stimuli like photos as extralinguistic references. The subjects receiving auditory linguistic input can observe the visual stimuli while conducting word-referent mapping. However, the visual stimuli for metalinguistic reference are missing because the referents made metalinguistically have to be interpreted literally. Take the sentence *turn off the light* as an example, the verb phrase is extralinguistic because the action of *turn off the light* can be represented visually. But the observable correlates of the action literally saying the sentence *turn off the light* is hard to be presented visually because *say* is referring to a metalinguistic reference which is saying the sentence.

Additionally, previous work has yet to investigate how SL or PBV word-mapping mechanisms work differently when encountering metalinguistic referents. The children who employ disparate word-mapping mechanisms will display different behaviors facing metalinguistic references. The predictions of two approaches are starkly different for domains that have no observable extralinguistic correlates. For SL, children will not make active attempts mapping what can be called the *zero informative* linguistic signal with non-observable extralinguistic referents, because they will disregard any forms with ambiguous statistical properties like the metalinguistic word *say* which (along with its direct object complement) cannot be mapped to any observable extralinguistic correlates. When encountering the sentence *say turn off the light*, the children who use SL will have trouble mapping the word *say* with metalinguistic referents because these referents are missing from their observable environment. Those children are no longer capable of calculating the co-occurrence rate of the word and the referent. So if these children employing SL can not measure the co-occurrence rate, they are unable to map the verb to metalinguistic referents.

The view that supports SL argues that one acquires a new word with full comprehension (Tunmer et al., 1984). Vygotsky (1962) also claimed that language awareness does not develop until the language has been acquired. From this standpoint, people produce language after they develop metalinguistic awareness of the language as suggested by SL. Children have to understand the word first and be able to calculate the co-occurrence rate of a word and the referent before using it. But for PBV, children will produce seemingly metalinguistic words because words are fast-mapped to the referents without comprehension. Circling back to the example of *say turn off the light,* the children who use PBV can not comprehend the word *say* either because the metalinguistic referents are still not visually accessible. However, PBV does not require children to calculate the co-occurence rate of the word and stimuli. Even though children are still potentially confused about the metalinguistic referents of the word *say*, they will propose a way of using it without understanding the nature or the function of the word *say*.

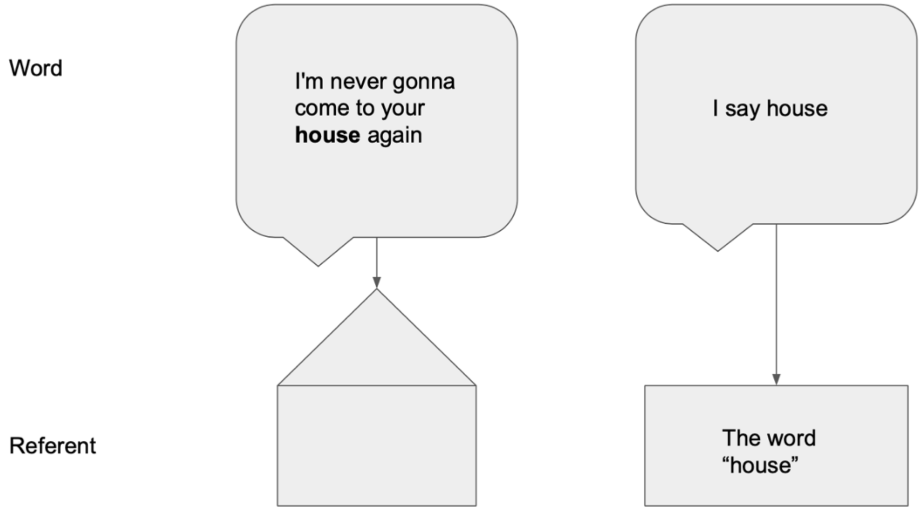
Some theoretical frameworks support the PBV mechanism because of the production-comprehension lag phenomenon which occurs in children or adults acquiring a first language or a second language (Benedict 1979; Schmitt 2008; Pilulski & Templeton 2004). Piaget (1928) suggested that children without sufficient metalinguistic awareness can still learn concepts that they are unable to use or evaluate usage. This is not a problem for PBV because children who use this approach will actively produce ostensibly metalinguistic words to check if their production is comprehended by others without having actual comprehension. The production-comprehension lag should be minimal if erroneous usage is factored into the development of children's production.

Based on the hypothesis that children who employ SL or PBV will exhibit disparate behaviors encountering metalinguistic usage, I propose a Extralinguistic Reference Bias (ERB) which suggests that children younger than approximately age 5 have limited access to metalinguistic language use. Children under 5 years old will only acquire extralinguistic referents but not metalinguistic language use. Figure 1 below shows that extralinguistic referent (e.g., the object house) and linguistic sign (e.g., the word *house*) are different entities.

Figure 1

Illustrations of extralinguistic and metalinguistic references.

.



Note. The left side represents the extralinguistic reference while the right side represents a metalinguistic reference.

We can assume an ERB in word-referent mapping akin to the Whole Object, Mutual Exclusivity, and Basic Taxonomic Level Biases (Markman, 1990). ERB is a motivated bias; in any situation where a linguistic sign (e.g., the word *house*) has an extralinguistic referent (the object “house"), the linguistic cues to this mapping overlap significantly with the cues to *metalinguistic usage* (i.e. *house* when it refers to ‘the word *house*’). So using a word when it refers to the word itself is mostly identical in using the word referentially.

Children's word-referent mapping process of metalinguistic and extralinguistic reference is in sequential order. The metalinguistic awareness can be achieved after children understand the extralinguistic referents. The complex cues to metalinguistic senses may not be acquired until after the child first maps *house* to its extralinguistic referent. Metalinguistic usage of a noun (e.g., the word *house*) will not be acquired until children understand *house*’s extralinguistic referent (e.g, the object “house”).

In many cases, the metalinguistic usage has a potential extralinguistic interpretation: *to say cheese* of course means, in context, literally to say *cheese* for a photo (because the anatomic effect of pronouncing the linguistic sign is essentially equivalent to producing a smile); however, there are substantial extralinguistic cues available to the child: the practice of smiling in the presence of a camera. In fact, it is possible that expressions like these directly contribute to children’s development of metalinguistic awareness, by generalizing across context-bound metalinguistic verbs. As suggested by Erickson and Thiessen (2015), word clusters are formed by similarity in semantics, and syntactic knowledge is word-specific; children learning English, for example, may learn some verbs with auxiliary-first question word order and others with verb-first command order. This perspective can also apply to early childhood word-referent mapping. Children first learn that each word can refer to various referents, then they group the referents together based on context-bound metalinguistic verbs; in other words, they see the pattern that the syntactic complements of *say* generally refer to the words themselves in one way or another.

To fill in the gap of the prior work, this thesis is an overview of the implications of metalinguistic awareness on early childhood word-referent mapping mechanisms through the corpus-linguistic study of how children use metalinguistic verbs up to age 5. Previous metalinguistic awareness studies paid more attention to children who are older than 5 years old because of younger children's deficit in metalinguistic awareness. For example, de Villiers and Villiers (1972), who had younger children make grammaticality judgments, reveals the problem that children under 5 years old cannot give grammatical feedback to the targeting sentences; rather, they focus on the content. Others also argue that “failure to elicit grammatical judgment” could be due to inefficient test techniques but not younger children’s limit in metalinguistic awareness (Tunmer et al., 1984). This study focuses on examining early children's production, especially their metalinguistic verb use like *say, tell,* and *ask.* I also coded ML awareness on a scale (see section 3.3 for detailed levels of metalinguistic awareness). The lowest ML awareness examples are metalinguistic word chunks (i.e., constructions) that have extralinguistic correlates while the most ML awareness examples are unambiguous reported speech. The thesis predicts that children progress through stages of metalinguistic awareness in their production, but early attempts to mimic reported speech constructions support the PBV mechanism.

The layout of the thesis is as follows: section 2 explains two approaches of word-referent mapping which are SL and PBV. Section 3 introduces the method of spaCy and the criteria table for each stage of metalinguistic awareness. Section 4 displays the data results, the relationship between children's metalinguistic use and age, the comparison between children and parents' speech, and the relationship between children's metalinguistic use and lexeme. Section 5 explains the results in full detail.

2. Two approaches of word-referent mapping

The following section 1.2.1 and 1.2.2 will explain two approaches of word-referent mapping: statistical learning and propose-but-verify. Statistical learning requires the subject to calculate the co-occurrence rate of the word and referent. Children employing statistical learning shows a gradual learning process because they need to map the word based on their prior knowledge. Children must understand a word first then use it. I argue that children would make less errors in statistical learning because children's production is based on a large amount of previous knowledge. Propose-but-verify does not demand subjects calculating any co-occurrence rate to perform word-referent mapping. Therefore, children who use PBV can fast-mapp one word to a referent. Children can produce a word without understanding it because their caregivers will correct erroneous usage. Then children will start making new proposals until they receive positive feedback and retain the correct word-referent mapping. The following Table 1 concludes the differences between statistical learning and propose-but-verify in word learning.

Table 1

Differences between statistical learning and propose-but-verify approaches in word learning

Statistical Learning	Propose-but-verify
Gradual learning	Fast mapping
Comprehension precedes Production	Production precedes Comprehension
Prohibits erroneous usage	Allows erroneous usage

2.1 Statistical learning in speech stream segmentation and the word-referent mapping

The theory of statistical learning originated from Saffran and Aslin (1996), which investigated hearing infants' auditory speech segmentation ability. Because the stream of speech does not represent the word boundaries, (this is known as *the segmentation problem)*, infants need to separate individual words from another. First, Saffran and Aslin (1996) played two minutes of audio which consisted of three-syllable pseudowords that align with English phonotactic rules but lack lexical meanings to eight-month-old infants. Then they presented subjects with part-words which were a combination of two syllables of one pseudoword and one syllable of another pseudoword to determine if subjects could differentiate pseudowords from part-words. The result showed that the eight-month-old infants took longer to listen to the part-words than pseudowords, indicating that the infants *could* identify the novel combinatorial properties of the additional stimuli. Saffran and Aslin (1996) suggested that infants were able to identify word boundaries through computing pairs of syllables that are less likely to exist in sequence because they are not found together within words. Infants could understand what syllables were often paired together: those that had high transitional probability. In part-words, the syllables in pseudowords were not paired together, which had low transitional probability, so infants took longer time to listen to them. Infants learned syllables' statistical relationship even with limited exposure: two minutes of listening to one language with no prosodic cues to word boundaries. Therefore, statistical learning is thought to be one way that children acquire the lexicon of a language.

Statistical learning also applies to other facets of language learning like syntactic acquisition (Saffran & Aslin, 1996) and word-referent learning (Yu & Smith, 2007). Yu and Smith (2007) tested statistical learning mechanisms in word-referent learning. Their findings supported the statistical learning procedure in which adult participants retain multiple references across the trials. They tested adult participants' word-referent ability by asking them to hear nonsense words then mapping each nonsense word with individual photos. The experiment tested 18 nonce words; and each nonce word had 6 trials. In each trial, subjects saw two, three, or four photos at a time while hearing one nonce word. Across 108 trials, Yu and Smith (2007) tested subjects' accuracy rate on pairing each nonce word with individual photos. Because the subjects' accuracy rate was above chance, Yu and Smith (2007) concluded that participants will put each novel word they heard with multiple photos, or referents, and then gradually converge on a single photo. This conclusion showed that subjects were measuring the co-occurrence rate of each nonsense word and photo throughout the experiment, then they merged on one pair that had the highest co-occurrence rate. So Yu and Smith (2007) argued that subjects employed statistical learning to measure the probability of which nonsense word and photo occur together.

2.2 Propose-but-verify in the word-referent mapping

Medina (2011) criticized Yu and Smith's (2007) test stimuli for being oversimplified from real-world input. Medina (2011) expanded on Yu and Smith (2007) methodology by using a videotape which embedded nonsense words into conversations as test stimuli for adult participants to make word-referent pairings in more complex circumstances. Subjects viewed a 40-second parent-children interaction videotape which was muted everywhere except the nonsense word part. Additionally, the subjects' correctness was tested trial by trial. But Trueswell and Medina's (2011) method was also questioned for being unable to assess how participants would perform differently if they were the observers rather than the utterance addressees. So Trueswell and Medina (2013) used artificial stimuli, similar to Yu and Smith's (2007), to test "propose but verify" learning strategies in adults.

"Propose but verify" learning strategy claims that word learning is a "fast mapping procedure" rather than a gradual statistical one. According to Trueswell and Medina (2013), people only retained one referent when they heard a nonsense word, if the referent was correct, they would confirm it; but if the referent was incorrect, they would search for a new one. In Trueswell and Medina (2013)'s experiment, adult participants were asked to pair artificial nonsense words with visually presented objects over different referential ambiguities. Each participant heard a sentence which included a nonsense word, then decided which object was the correct referent of that sound. The trials have either low informativity with five objects or high informativity with two objects. For example, the subjects would hear a sentence “Oh look, a zud!” while looking at two or five objects. The findings showed that participants can seize upon the correct mapping quickly when informativity is high. Moreover, participants made single-meaning hypotheses in word-referent pairing rather than multiple-meaning hypotheses. Subjects did not show memory of alternative incorrect referent if they did not select the correct referent in the first place. They only retained the memory of previously correct referents. Adults use the “propose-but-verify” mechanism in word-referent learning, but it remains to be seen whether this is at work in infants’ language acquisition.

2.3 Two parallel theories for the thesis

The current study draws from three theories: the simultaneous development of metalinguistic awareness development and language development, the Verb Island theory, and mutual exclusivity constraints in word learning. The simultaneous development of metalinguistic awareness and language acquisition suggests that language and metalinguistic awareness are related. The verb-island theory narrows the target word of evaluating metalinguistic awareness to verbs because each verb forms isolated grammatical organization. The mutual exclusivity constraints explain children’s difficulty in comprehending both extralinguistic and metalinguistic awareness because children can not map one word to both extralinguistic and metalinguistic referents.

The concomitant of language development and metalinguistic awareness development is first proposed by Clark (1978) who uses a repair test to demonstrate that children’s metalinguistic awareness is a parcel of children's language acquisition development. From this theory, the current study predicts that metalinguistic learning is related to language acquisition.

The Verb Island theory uses item-specific schema to investigate verbal argument structure construction. This theory treats the children's grammatical development of single verbs isolated from others. The complex pattern is the collection of every simple verb-island. From this theory, because the schema is defined by each predicate involved (Tomasello, 1992), verbs are an important component for understanding children's grammatical development. The current study will treat the various comprehension levels of verbs as different levels of metalinguistic awareness development. For example, *say meow* and *I said no* shows different levels of verb comprehension. *Say meow* presents a low comprehension level of verbs because it is mimicking animal sound, the comprehension of metalinguistic verbs used is not shown in the predicates. As for *I said no,* however, it shows a high level of verbs comprehension because the predicates convey a more metalinguistic message like denial (see Section 2.2).

Markman (1990) suggests that the whole object assumption is the initial step for children to identify a single object, then children acquire the properties of the object by following mutual exclusivity (ME). In this thesis, ERB assumes that children younger than approximately age 5 have limited access to metalinguistic language use due to ME. ME suspends overlapping referential objects, but in some cases of ERB, the word that refers to itself is mostly identical with using the word referentially (See introduction). Children who apply ME can not differentiate the differences between the addressing word itself and using the word referentially. The level of ME is controlled in prior work and current thesis. Goldstone (2010) tests bilingual children who have low ME while the current thesis uses age to control the children's access level of ME. Bilingual children are supposed to have low ME because two terms are given to the same objects. In ERB, when children grow older, or closer to 5 years old, their ME will be lower because of their PBV approach which allows them to receive corrections upon erroneous metalinguistic language use.

3. Method

I use corpus-linguistic study to identify the metalinguistic awareness of children prior to age 5. The random forest classification is used to identify the metalinguistic awareness level of each sentence. The corpus-linguistic analysis is used for compensating the naturalness lacking in Yu and Smith (2007) research which uses self-created visual stimuli. As Medina et al (2011) points out, Yu and Smith's study are less naturalistic than real-environment because the visual stimuli and pseudowords are customized by researchers. Corpora, on the other hand, are transcripts from parent-child interactions in real life. Furthermore, the results of corpus-linguistic analysis are reliable as Gries (2005) suggests: the results from corpus-linguistics analysis are aligned with experimental results.

Section 3 proceeds as follows: Section 3.1 presents the data being used. Section 3.2 shows the overview of the participants. Section 3.3 demonstrates the criteria table for categorizing each level of metalinguistic awareness and how spaCy is used to code the data.

3.1 Data

All data were retrieved from The Child Language Data Exchange System (CHILDES) - North America collection and syntactically parsed en masse using spaCy, a software library for

Natural Language Processing. CHILDES is a corpus composed of written, audio, or video transcripts from parent-child interactions collected by developmental psychologists and language acquisitionists. With modifications, spaCy's dependency parser was used to extract any verb dependents like Subject (Subj), Direct Object (DObj), and Indirect Object (IObj) and their semantic domains. These values were used in turn to code the data (75102 raw tokens, 41,766 cleaned tokens of metalinguistic verbs by children or their caregivers) using a random forest model to classify tokens based on a large training set (18,613 tokens) (see Section 3.3 for information on the coding process). The tokens are sentences that contain metalinguistic verbs: *say, says, said, saying, tell, tells, told, telling, ask, asks, asked, asking.* Other metalinguistic verbs that can not have clausal complements are not included in the list. For example, the metalinguistic verb *talk* is not included in the list because the utterance **she already talked that she went to work* is not grammatically acceptable. The asterisk marks the grammatically incorrect speech.

3.2 Participants

Imported data from the CHILDES - North America collection consists of $N = 167$ participants who are 5 years old or younger. The mean age is 4.05 years old. Participants interact with adults in various conditions like play sessions, home visit, and daily dialogs.

3.3 Procedure

Though metalinguistic awareness does not have a clear definition or and is not definitively measurable, I propose the hypothesis that metalinguistic awareness level is quantifiable if it is operationalized as the ability to actively speak metalinguistically. Recall the ERB: children prior to 5 years old have limited access to metalinguistic usage. They will progress through different stages of metalinguistic awareness until they fully develop their comprehension and production metalinguistic reference ability. To identify various developmental stages of metalinguistic awareness, Table 2 shows gradually increased levels of metalinguistic awareness with examples. Value 0 indicates the examples of indeterminate because children are either being unconscious about the metalinguistic verb like *say* or being highly metalinguistic, and the immediate context does not shed light on the intended usage - these cases were all coded as Category 4 for the initial model. Value 1 demonstrates non-linguistic awareness because the direct object is observable and *anatomical:* acts denoted by these constructions are the result of anatomy like a *meow,* or refer to the changes in anatomy like mouth configuration when a person says a word like *cheese*. Value 2 demonstrates neutral metalinguistic awareness because it contains a single lexeme as a speech event. Using the single lexeme already indicates a basic level of metalinguistic awareness but the verb complements are missing or omitted so no more advanced level of metalinguistic information is given. Value 3 indicates the intermediate metalinguistic awareness. Children in this category have speech-act utterances that show an extralinguistic effect like denial. Value 4 is the most advanced metalinguistic awareness because children demonstrate reported speech unless the children are imitating the parents' utterances which are indeterminates showing no metalinguistic awareness. These are indeterminate in the data as collected, which does not contain any data from the interactional context.

In value 1, children do not display the comprehension of verbs because the animal sound do not convey high animacy information like in value 2 and above. In value 2 and 3, children make less mistakes in verb functions, so they show more advanced metalinguistic awareness, but not as much awareness as reported speech. At these stages, the caregivers' correction rate will be lower compared to value 0 if parents give negative response to children's seemingly metalinguistic verb use at value 0. At value 4, children present the most advanced stage of metalinguistic awareness because the information is conveyed through clausal complements. The children who use reported speech correctly already comprehend the nature and function of the metalinguistic verb. Their cognition is prepared for comprehending their own usage of language.

Table 2

Criteria for Each Stage of Metalinguistic Awareness

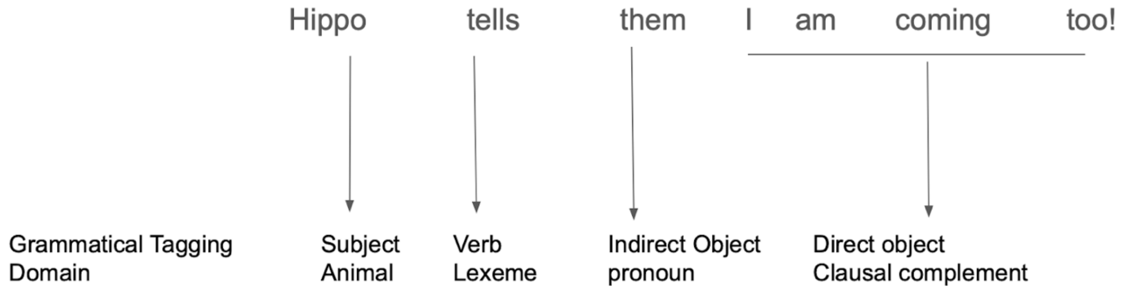
Value	Category	Type	Example	Corpus	Age and Name	Note
0	indeterminate	Say X	"**say cheese**"	Demetras2	2:5,8, Jimmy	High information
1	nonlinguistic	Anatomical phenomena	"saying **meow**"	Clark	3;9.6, Shem	Call or make a sound
2	metalinguistic	Speech events	"**ask**"	Braunwald	3;6, Laura	Single lexeme
3	metalinguistic	Near speech acts	"when you come to my house **I'm say no**"	Hall	4;9, Kip	Not speech acts but involving extralinguistic effect, like denial
4	metalinguistic	Reported speech	"I said I might sleep in it"	Gelman	4;11, Ronald	Clausal complements

Note. Table is designed by the author based on observations of instances.

The random forest classification algorithm, a predictive learning model, categorizes the tokens into different values based on the table above. The variables are words, or metalinguistic verbs, subject words, direct object words, indirect object words, prepositions, possessive object words, subject domains, direct object domains, indirect object domains, possessive object domains.An example of spaCy coding for a token is shown in the figure 2 below.

Figure 2

Example of spaCy coding for a sentence



Note. The first line represents the grammatical category of each word. The second line represents the semantic and grammatical domain that each word belongs to.

Firstly, spaCy identifies the target metalinguistic verb which is *tells,* then identifies the dependency like subject, indirect and direct object. After extracting the verbs and dependents, their semantic/grammatical domain is determined.

The coded grammatical variables contribute to categorizing most of the data because words, semantic and grammatical domain could indicate the levels of metalinguistic awareness The subject domain could indicate ML because if the subject is an animal, it is likely that the token is not metalinguistic. The actual word suggests the semantic and grammatical properties of the subjects and objects which are diagnostic of ML usage. If the object is a clausal complement, then that is automatically a Value 4.

Subject domains have the following categories: animals, work of art, generic, date, linguistic, GPE (see Table 3 for the explanation of these abbreviations), LOC, NORP, ORG, NNP, PRODUCT, and PRP. Object domains have all of the above and speech acts. See the following table for detailed descriptions of each domain. However, spaCy fails to correctly identify appropriate tokens for each domain. In NORP, LOC, and ORG domains, person names are included incorrectly. The entire list of domain is listed in Table 3. Some of the domains are automatically identified by spaCy and rest are identified using customized code.

Table 3

Domains for each word

Domain	Description	Identification Software
NORP	Nationalities, religious, or political groups	spaCy
LOC	Non GPE locations	spaCy
GPE	Countries, cities, states	spaCy
PRODUCT	Objects, vehicles, food, etc.	spaCy
DATE	Absolute or relative dates or periods	spaCy
WORD_OF_ART	Titles of books or songs,etc	spaCy
ORG	Companies, agencies, etc	spaCy
ANIMALS	Animals and their children forms like dog, doggie, cat, kitty,etc	custom
NNP	Noun phrases	custom

Domain	Description	Identification Software
PRP	Pronouns	custom
LINGUISTIC	Lexical noun objects like *question*, *grace*, *no* (denial), or *sorry* (apology).	custom
SPEECH ACT	Dependency of the verb was a clausal complement, suggesting reported speech	custom

Note. The domains are either identified by spaCy or manually customized by the author.

4. Limitations of current research

Section 4 discusses the limitations of this thesis in two ways: the limitations of corpus-linguistic study and the limitations of theoretical framework.

4.1 Problems with corpus-linguistic study

Before the method section, a few limitations of corpus-linguistic study need to be laid out first. The current corpus-linguistic method exhibits some problems. First, because some corpora do not provide video or audio transcription, the environment of the subjects are unknown. It is impossible to determine what subjects can see or hear. Furthermore, the corpora only reflect a small amount of children's passive and active language usage. Corpora may fail to holistically display children’s metalinguistic development.

Though the corpus-based approach has limitations, Gries (2005) claims that the data from corpus-linguistic analysis are aligned with experimental data. This thesis uses production as a test of comprehension by disregarding the production-comprehension lag presented in children. Given the results reported in Section 5, it is clear that children select metalinguistic verbs from the earliest utterances, so it seems that comprehension would lag production. This is exactly what is predicted by PBV, and another framework might not be as conducive to corpus-linguistic study for the reasons I note above.

4.2 Problems with theoretical frameworks

The approach taken in this thesis has a parallel in other veins of research: the concomitant of metalinguistic awareness development and language development, the verb-island theory, and the existence of production-comprehension lag in children. The concomitant of metalinguistic awareness development and language development leads to the connection between metalinguistic awareness development and word-referent mapping. The verb-island theory narrows the target word of evaluating metalinguistic awareness to verbs. However, these three two theories are both debatable.

Furthermore, the relationship between metalinguistic awareness and language development are debatable. This study assumes that children's metalinguistic awareness is a part of the language acquisition development process (Clark & Anderson 1979; Marshall & Morton 1978). But other theories of metalinguistic awareness development exclude it from the language development process, considering it instead as a general change in information processing during the middle childhood (Tunmer et al., 1984), or as an effect of learning to read (Donaldson, 1978). Tunmer et al (1984) claims that metalinguistic development is related to the changes in general information processing ability rather than language development. Donaldson (1978) argues that the development of metalinguistic awareness is due to the beginning of formal schooling.

The current research is based on insularity of verbal argument structure constructions, the verb-island theory, which treats each predicate as having its own syntactic structure. But other theories refute the insularity hypothesis, claiming that children's verbal argument development is not item-specific: Ninto (2003) claims that children's grammar development is not a collection of verb-islands but a system. She states that a complex system is not based on "combining simple constructions" but related to language input.

5. Results

Section 5 proceeds as follows: Section 5.1.1 to Section 5.1.5 explain each step of model implementation in detail. Section 5.2 evaluates the model performance using gini index. Section 5.3 displays the variable importance in determining the level of metalinguistic awareness. Section 5.4 presents the relationship between children's age and metalinguistic usage. Section 5.5 demonstrates the comparison of children and parents' metalinguistic usage. Section 5.6 presents the relationship between metalinguistic awareness and lexeme.

5.1.1 Model Implementation

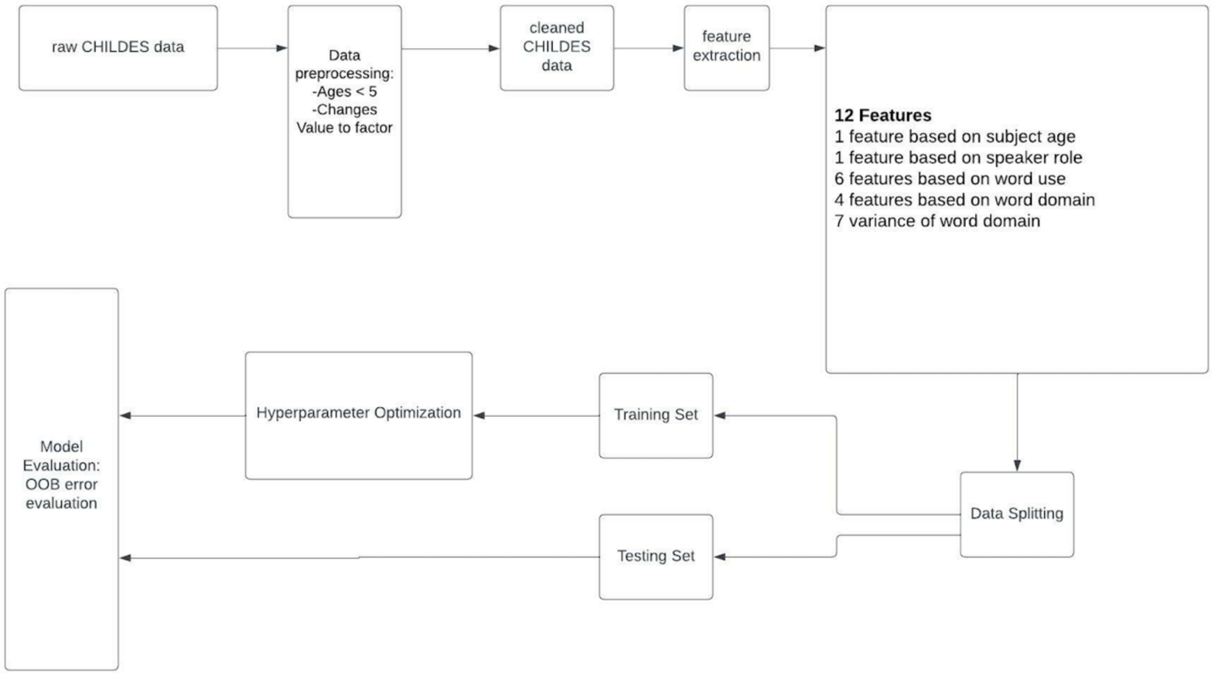
The current research uses a non-binary random forest classification approach to systematically analyze tokens' stages of metalinguistic awareness. The flowchart in Figure 3 below shows the random forest model implementation steps.

Decision tree learning is predictive modeling that divides data into branches by their values on different parameters (coded variables). These trees differ from each other in terms of the number of branches or splits that occur, the order of splits, and the variable levels or values that the splits occur at. When a sequence of splits yields a majority value in the dependent variable, that set of variables is used to conclude the target value from observations represented by each branch.

Random forests operate by building a multitude of decision trees. The classification operation will give the output that most trees selected. Random forests avoid the overfitting of the model to the training set in decision trees by feature randomization.

Figure 3

Workflow to optimize the random forest classification model


raw CHILDES data
Data preprocessing: -Ages < 5 -Changes Value to factor
cleaned CHILDES data
feature extraction
12 Features
1 feature based on subject age
1 feature based on speaker role
6 features based on word use
4 features based on word domain
7 variance of word domain
Data Splitting
Training Set
Testing Set
Hyperparameter Optimization
Model Evaluation: OOB error evaluation

5.1.2 Data preprocessing

Though the age of onset of metalinguistic awareness is undetermined, according to Tunmer (1984) and Vygotsky (1962) suggest that children do not develop cognition of language until school age. Additionally, as mentioned in the previous introduction section, the limitation of conducting grammaticality judgments for younger children also precludes the participants who are older than age 5. To diversify the participants' age range and fill in the gap of the previous study, this research will pay more attention to children who are under 5 years of age. In the data preprocessing step, the tokens of children who are older than 5 years of age are filtered out. The tokens of adults are filtered out in this model as well and considered separately in the characterization of children's input over time.

Based on the Table above, the value of each token ranges from 1 to 4. Value is included in the data analysis as a feature (see Section 3.3). The raw CHILDES data has 75,102 tokens of the three metalinguistic verbs targeted here. After data preprocessing, the number of tokens is 41,766. Child-produced tokens are 40,680; child-directed tokens made by parents are 1,086.

5.1.3 Feature extraction

This study only concerns the dependency parse regarding metalinguistic verbs, therefore the following features are included: *subject word, direct object word, indirect object word, possessive word, possessive object word, subject domain, direct object domain, indirect object domain, possessive object domain*. Irrelevant features, *events* and *corpus* are included for qualitative analysis in the follow up of early metalinguistic usage.

5.1.4 Data splitting

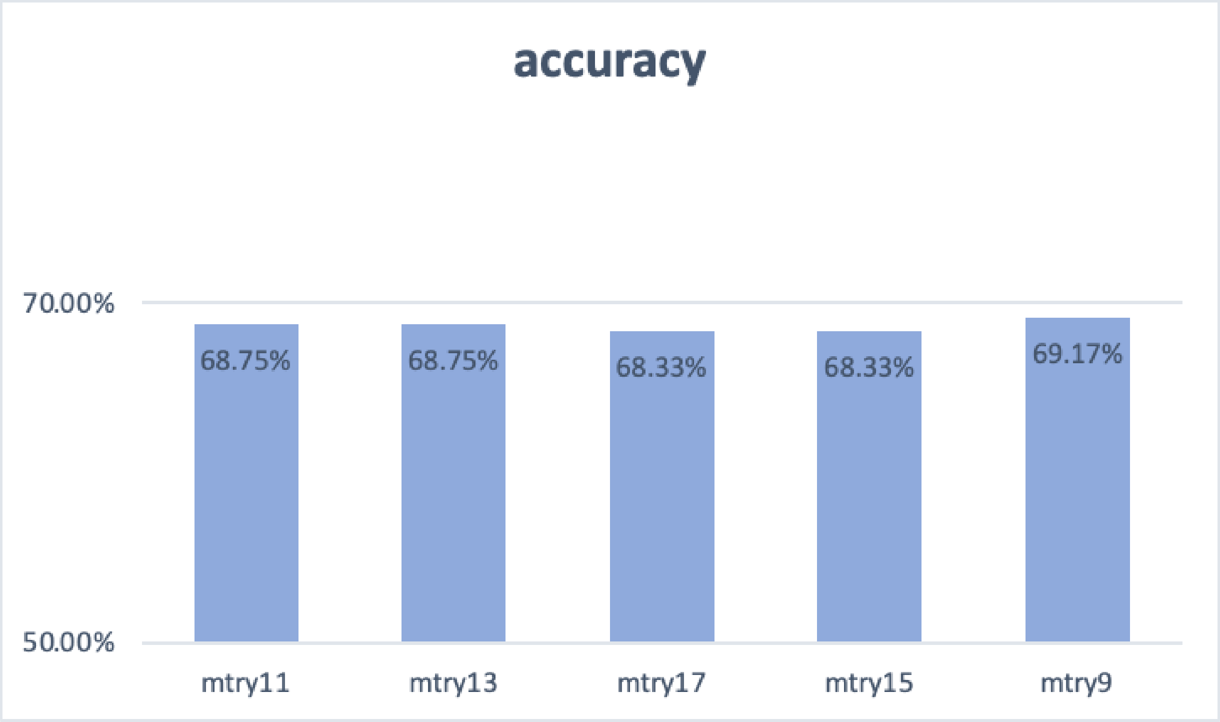
344 instances are split into training and testing sets. The current study assigns the training and testing ratio to 7:3. The number of tokens in the training set is 18,613; the number of tokens in the testing set is 7,977. This is done to quantify the accuracy rate of the model on known values and use that rate to estimate the accuracy of the rest of the data classified by it.

5.1.5 Optimization

The hyperparameter, represented by *mtry*, in random forest classification is the number of variables allowed to be used in each split of node in the decision trees that make up the random forest. The features are selected randomly in the node to ensure the outputs are in low correlation therefore boosting accuracy. Because the performance of a random forest classification model depends on the hyperparameter, *mtry* is tuned in the classification model. Five models are used in various *mtry*. The Figure 4 shows the various accuracy regarding different numbers of *mtry*. Four hyperparameters are used to optimize random forest classification model accuracy. According to the figure below, when *mtry* equals to 9, the true accuracy is 69.71% which is the highest. The final model only considers 9 features out of 17 at each split.

Figure 4

Accuracy across different hyperparameters



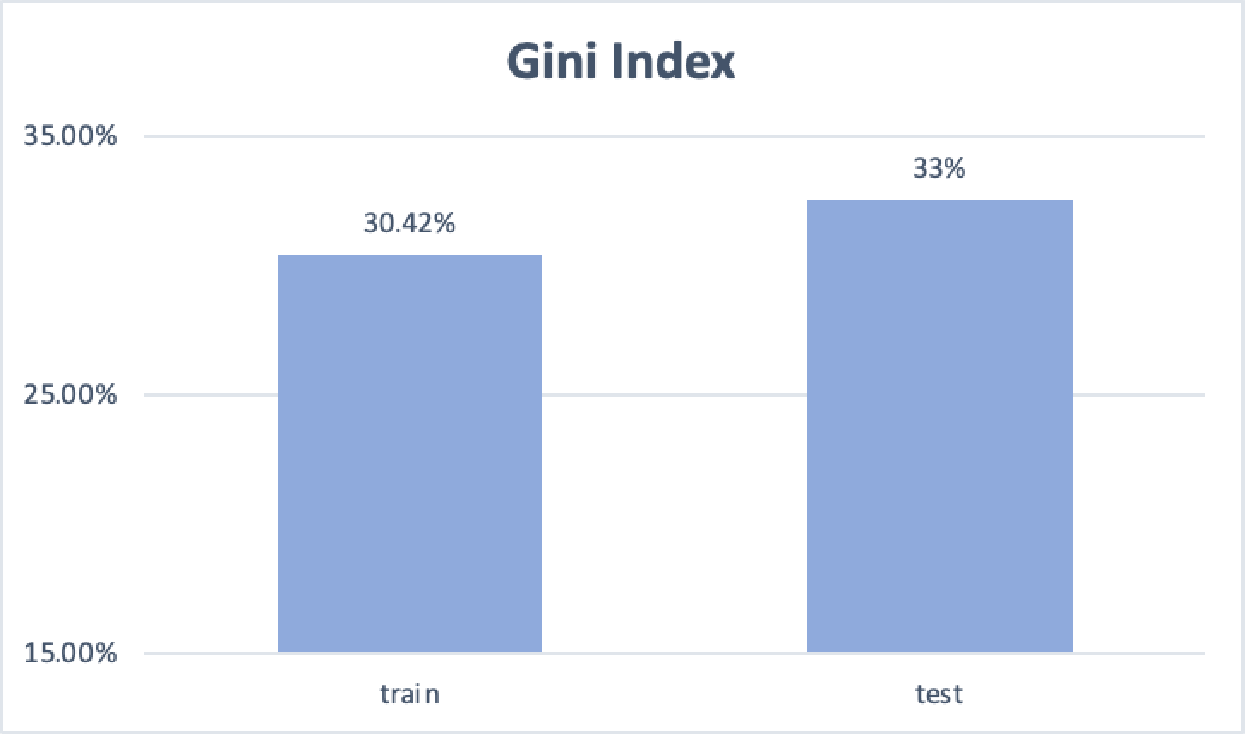
Note. mtry refers to the hyperparameter that is tuned during the optimization step.

5.2 Model Performance

The gini index suggests the probability of the wrongly classified features when randomly selected. The gini index of training and testing set is shown in the Figure 5 below. For the training set, the gini index is 30.42 % which means that the training set has 69.52% of true accuracy rate. For the testing set, the gini index is 33% which means the testing set has 67% of true accuracy rate.

Figure 5

The gini index for testing and training set



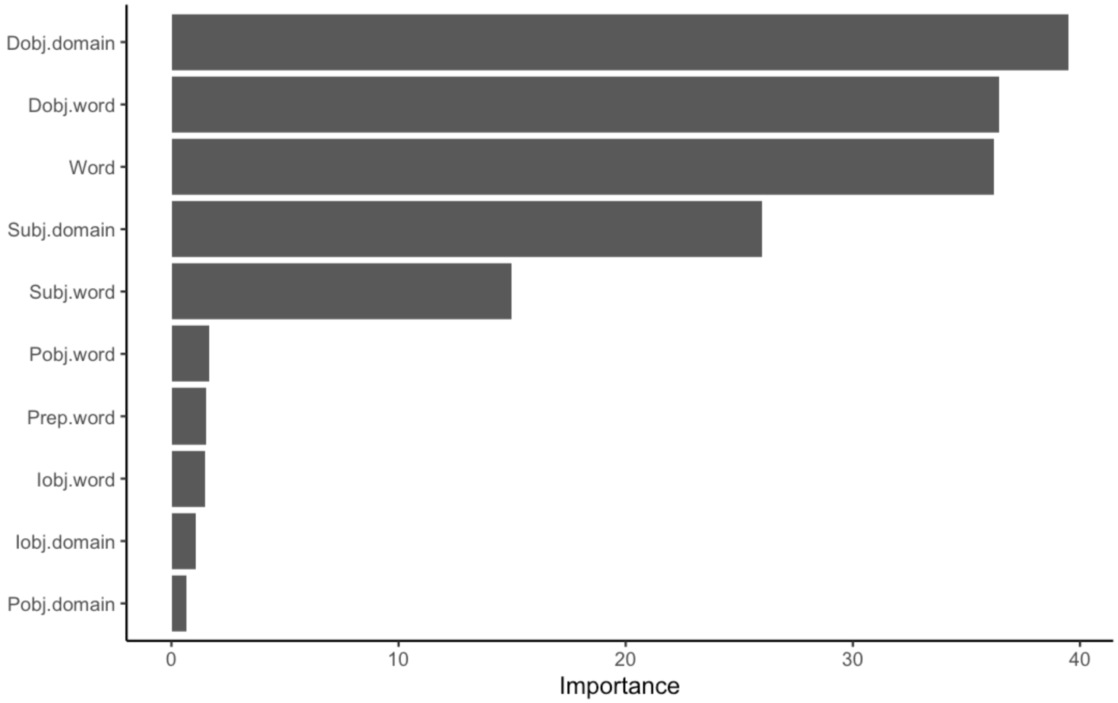
Note. This figure demonstrates the gini index of training and testing set.

5.3 Variable Importance

The variable importance is listed in Figure 6 below. The *direct object domains, direct object words,* and *words* are the three most important features in determining the metalinguistic stages of the tokens. The importance of direct object domain could be explained by an example of inquiry. If the direct object domain belongs to a question, then the metalinguistic awareness stage is likely to be 3 because it presents extralinguistic effects like inquiry. Direct object words are important because if the word is an animal sound, then the value is probably 1. So both direct object words and domains contribute to the stage of metalinguistic awareness.

Figure 6

Variable importance in determining the stage of metalinguistic awareness



Note. The figure demonstrates the variable importance in determining the stages of metalinguistic awareness.

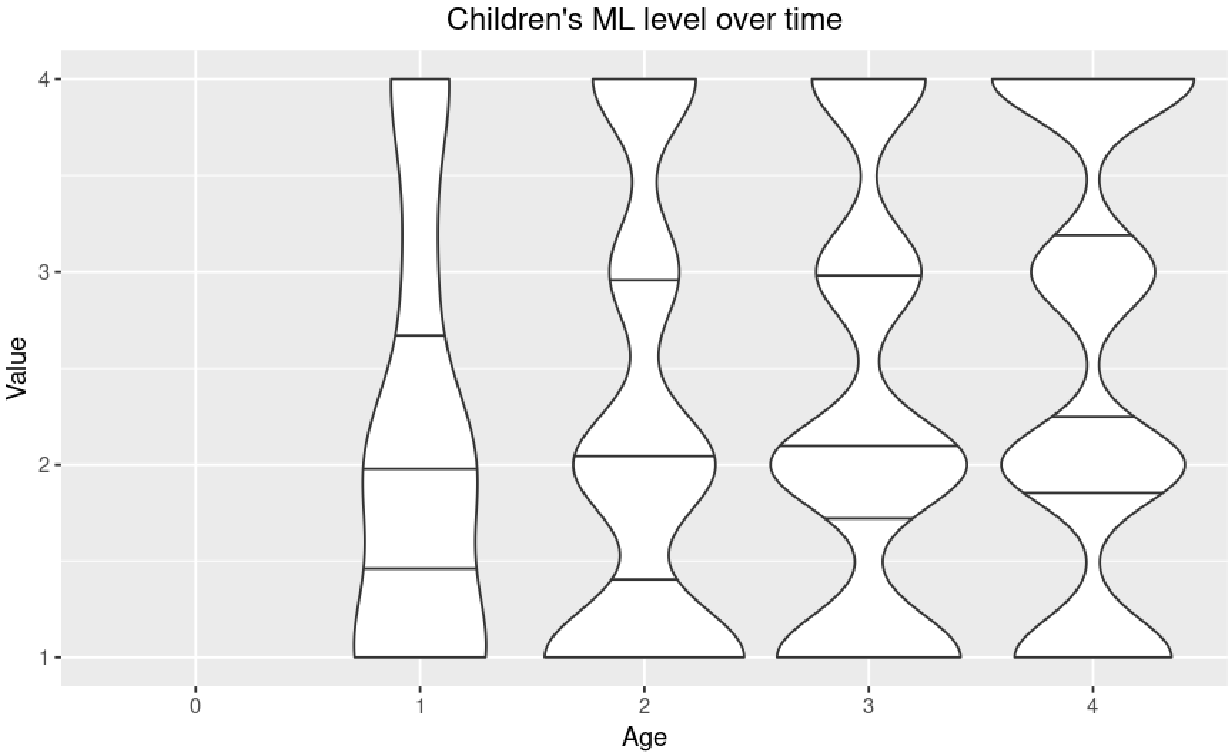
5.4 The relationship between children's metalinguistic awareness stage and age

The results support the hypothesis that children prior to age 5 do have limited access to metalinguistic awareness. The data were fit to a multinomial regression, with age as an independent variable and metalinguistic level as the dependent variable. However, the P value of the relationship between the value and age, 0.4377, is higher than the significance threshold 0.05. Based on P value, the null hypothesis that regardless of age, children's ML stages are in equal distribution can not be rejected. Therefore, I conclude that age is not a relevant feature in predicting children's metalinguistic awareness. Because age is irrelevant in determining the ML levels, the null hypothesis - that the means are the same across ages - cannot be rejected. The results do not support SL, but they cannot conclusively reject it either. However, some erroneous usage of children at early childhood is associated with PBV. The future model should split erroneous usage using context so that PBV is tested directly.

Figure 7 uses a violin plot which shows the median value (the middle line in each "violin") and quantiles of age of each level of ML. For Figure 7, the median value slightly increases across ages. The results of Figure 7 show gradual development when observing only the mean.

Figure 7

Children stages of metalinguistic awareness across ages



Note. The figure demonstrates the median and interquartile range of the stages of metalinguistic awareness that children at each age have.

However, some instances of children who exhibit metalinguistic awareness as early as age 2 do support PBV. The isolated tokens that appear to be reported speech are instances of children's imitations of parents' modeled speech like example 1:

1. *Peter 2,05.3* Bloom:020503 (1349-1474) (Bloom 1975)
   MOT: can you say [inaudible] off?
   CHI: light off.
   MOT: Yeah.
   MOT: [say] don't turn the light off.
   CHI: say don't turn that light on.

Peter, who is 2 years old, produces the ostensibly metalinguistic sentence *say don't turn the light off.* Peter's utterance contains clausal complements *don't turn off the light* so random forest classification will identify Peter's utterance as stage 4. But Peter does not understand the use of *say* in his utterance because he only mimics the mother's speech. So, Peter's behavior is aligned with PBV for the fact that his comprehension lags the production.

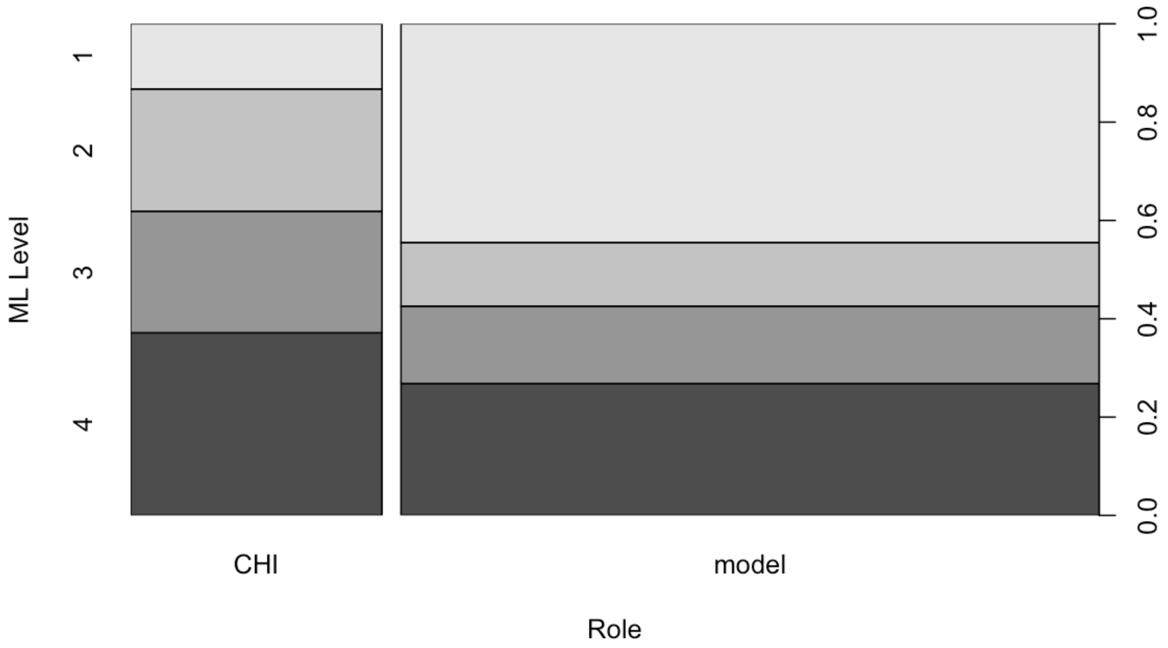
5.5 The comparison of metalinguistic usage between parents and children

Figure 8 shows the comparison between parents' and children's metalinguistic awareness stage. Surprisingly, children produce overall a higher number of values 2, 3 and 4 than parents have. Parents tend to produce a higher number of value 1 than children do. So parents produce a smaller portion of metalinguistic tokens than their children do. The conditions of children and parents' interactions can be assumed as child-directed speech (CDS) where parents engage in a slower, repeated, and more simplified strategy when talking to their children. According to

Harley (2010), CDS is used by adults who repeatedly produce shortened grammatically simplified sentences when talking to their children.

Figure 8

The comparison of parents and children's metalinguistic usage



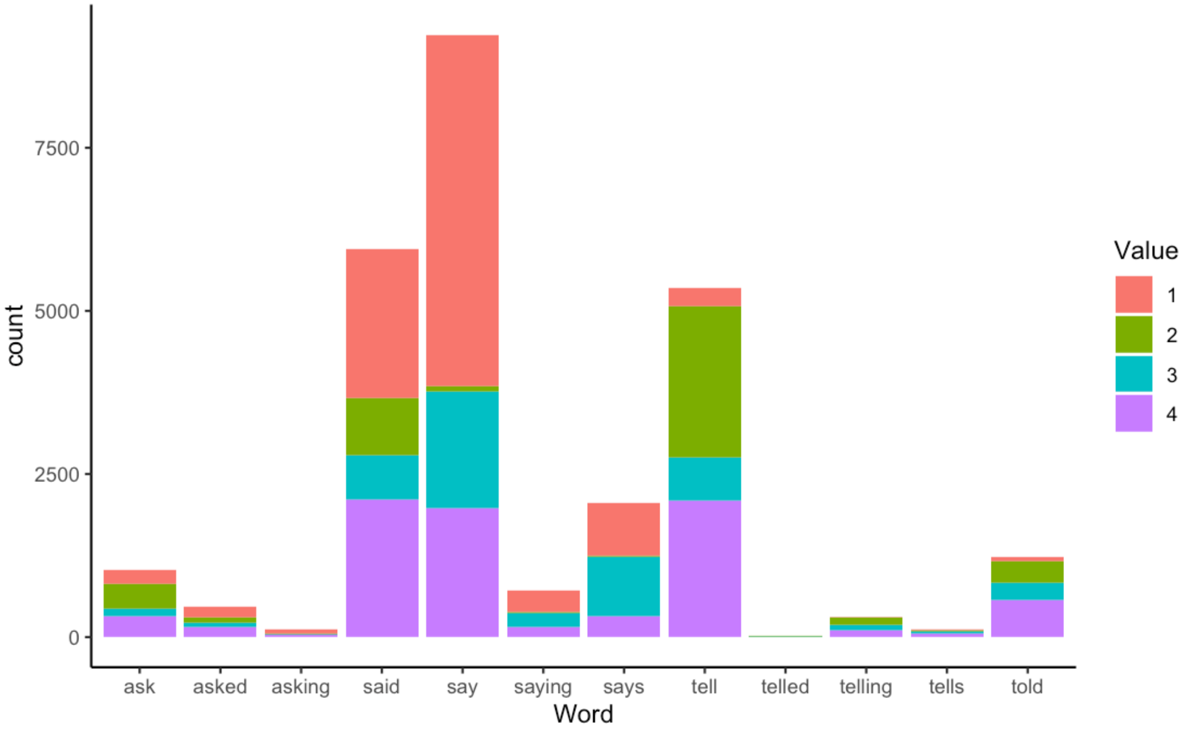
Note. The label 'model' refers to the parents' metalinguistic usage.

The tendency that children produce more instances that are in advanced levels of metalinguistic awareness supports fast mapping which allows error-ridden usage. Children won't produce the advanced level of metalinguistic usage in SL which prohibits erroneous usage.

5.6 The relationship between metalinguistic verbs and metalinguistic awareness level

Because the metalinguistic verb is one of the top three important variables in determining the metalinguistic awareness, the relationship between lexeme and metalinguistic awareness level is presented in Figure 9. Figure 9 shows the number of instances and the number of each metalinguistic value in every lexeme. Firstly, the number of instances of each lexeme is calculated. Lexeme *say* has the highest number of instances in the dataset. Lexeme *say* also has the highest portion of value 1 across its own distribution of metalinguistic awareness level. This skewed distribution is due to the biased Table 2 in Section 3.3. Table 2 tends to classify the anatomical configuration after lexeme *say* but no other lexemes like *ask* or *tell.* Because lexeme indicates a more general speech event while ask or tell suggests a specific communicative event. It is impossible to produce **tell meow* or **ask meow*.

Figure 9
The relationship between metalinguistic verbs and metalinguistic awareness level



Note. The figure demonstrates the number of instances for each lexeme at each stage of metalinguistic awareness.

Figure 10 below shows that each lexeme has a similar median age across value 1 except lexeme *say* and *says*. Lexeme *say*, *saying* and *says* in value 1 has a median age of 2 to 3 years old while the median age of other lexemes in value 1 is around 3.5 to 4.5 years old. This phenomenon suggests that in the preliminarily metalinguistic stage, children tend to learn the metalinguistic verb *say* first, then pick up other metalinguistic verbs like *ask* and *tell*. But the median age across each lexeme does not differ much when moving towards more advanced stages of metalinguistic awareness. Therefore, *say* indicates more generic communicative event while lexemes *tell* and *ask* indicate more specific communicative event, denoting information exchange. I argue that the lexeme *say* is more generic because it can incorporate anatomical sounds like *meow*. The phrase say meow is grammatically appropriate while **ask meow* or **tell meow* are not. When moving towards higher stage of metalinguistic usage, all lexeme ask, tell, or say can incorporate clausal complements. Table 4 below shows the comparison between acceptable metalinguistic verb complements. The asterisk marks the ungrammatical phrases.

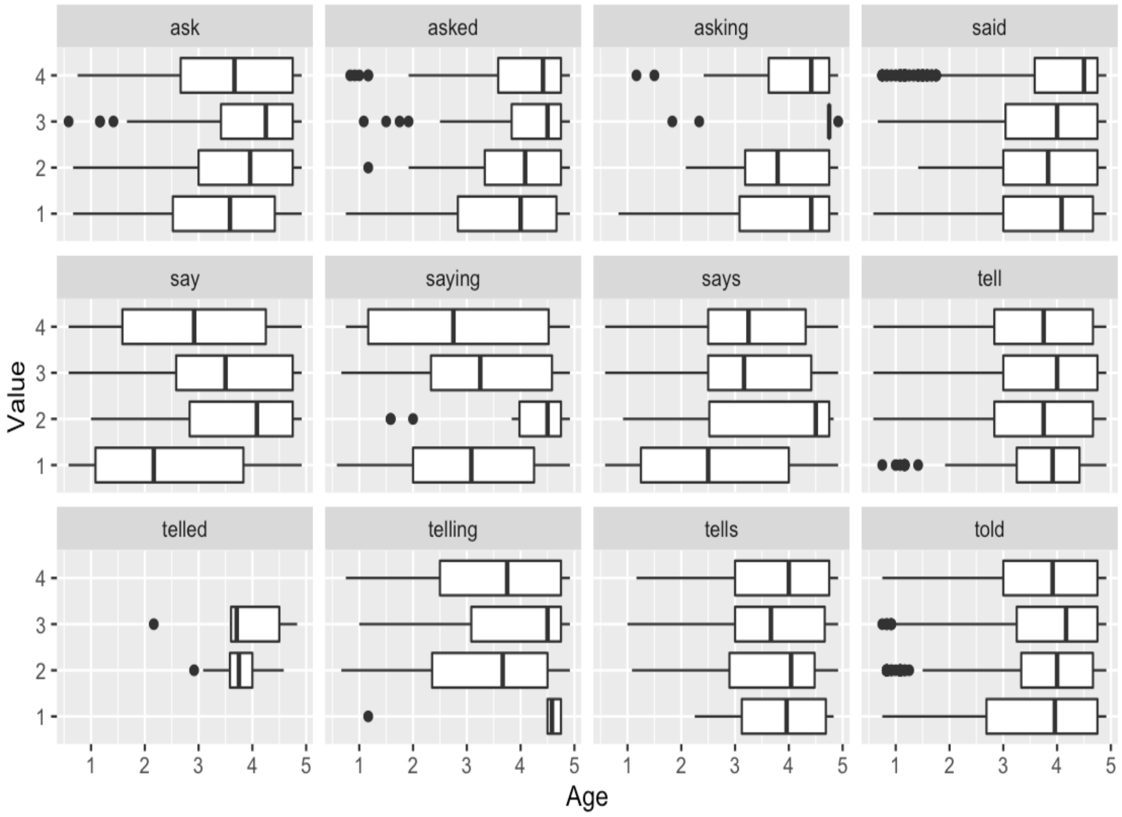
Table 4

Comparison between Acceptable Metalinguistic Verb Complements

Anatomic sounds	Clausal complement
Say meow	Say “ I might sleep in it”
* Tell meow	Tell someone “I might sleep in it”
* Ask meow	Ask someone if “I might sleep in it”

Figure 10

The median age of metalinguistic awareness for each value across lexemes



Note. The figure demonstrates the median age for each stage of metalinguistic, from 1 to 4, awareness across all lexemes.

6. Discussion

The thesis predicts that before age 5, children will have little access towards metalinguistic usage. Children progress through stages of metalinguistic awareness in their production, but children who exhibit the most advanced level of metalinguistic usage early support PBV mechanism that children will produce seemingly metalinguistic utterance without observable utterances.

My findings do not support the ERB hypothesis that children have limited metalinguistic awareness prior to age 5 because children's age can not predict the ML level. The high value of significance threshold 0.4377 indicates that feature age does not predict the increasing development of ML level.

However, the erroneous use presented in young children demonstrates the PBV mechanism. Some children even present the advanced level of metalinguistic awareness at 2 years of age. These children's behavior is aligned with PBV mechanism that children will produce metalinguistic referents without observable correlates. In this case, children will make erroneous usage of a verb without knowing how the verb works. Furthermore, some of the isolated tokens that appear to be reported speech are instances of children's imitations of parents' modeled speech like example 1 (repeated here for clarity):

1. *Peter 2,05.3* Bloom:020503 (1349-1474) (Bloom 1975)
   MOT: can you say [inaudible] off?
   CHI: light off.
   MOT: Yeah.
   MOT: don't turn the light off.
   CHI: say don't turn that light on.

Example 1 shows the process of a child ostensibly producing a metalinguistic token, *say,* as the proposing part of the PBV mechanism. The child has not acquired the metalinguistic sense, but randomly selects this word in the face of negative grammatical evidence of their utterance.

The differences in quantile between Age 1 and others in Figure 7 can be explained by the Two-Word Stage (TWS) and Telegraphic Stage (TS). Children at TWS will produce mainly nouns while omitting grammatical markers (Devilliers & Devilliers, 2013). At TS, children's utterances will include more content words featuring nouns and verbs like example 2:

2. *Laura 2;08.19* Braunwald: 020819 (33-35) (Braunwald, 1973)
   CHI: sit
   MOT: What?
   CHI: Mommy sit.

Children drop the verb inflections like *Mommy sit* which is supposed to be *Mommy is sitting* while maintaining the content words that carry out the highest level of information like the verb *sit.*

Parents on the other hand, present fewer metalinguistic awareness compared to their children. Parents in many instances are engaged in CDS in which they produce slower, repeated and simplified grammatical speech when talking to children like example 3:

3. *Erin 1;01.26* New England:14/04 (935-939) (Snow & Pan, 1994)
   MOT: There is the wow_wows
   MOT: and kitty kitty.
   MOT: meow.
   CHI: [inaudible] (smiles)
   MOT: you like the bow+wow huh?

Example 3 shows that when talking to a 1-year-old, parents mimic the animal sounds *wow, meow, bow,* though children do not give any response. The words *kitty* and *wow* are repeated twice; and the grammatical structure of the polar question omits the auxiliary verbs. Shore (1997) suggests that CDS enables children to understand the basic structure and function of the sentences. Therefore parents tend to use CDS to help children develop cognitively.

The relationship between lexeme and metalinguistic awareness suggests that at the basic level of metalinguistic awareness, children may pick up the word *say* first then acquire other metalinguistic verbs. When moving towards more advanced stages of metalinguistic awareness, children do not show any priority in picking up any lexeme.

Though the data collected from corpus-linguistic study do not differ much from experimental data according to Gries (2005), corpus-linguistic study still has unresolved limitations. For example, the value 0 in table 2 is indeterminate because spaCy cannot identify which contexts are relevant or which context contributes to the false identification of the metalinguistic usage. Without any relevant context information, the interpretation of sentences that have multiple meanings like *say cheese* is hard to determine. Additionally, it is difficult to detect tokens of the target verbs repeated from parents. Future study can investigate the ERB cross linguistically using an experimental approach. Furthermore, the mean length of utterance (MLU) of children could also lead to more clear differentiation between ML levels. More search should be done through examining the relationship between MLU and ML levels.

References

Berk, S., & Lillo-Martin, D. (2012). The two-word stage: motivated by linguistic or cognitive constraints?. *Cognitive psychology*, *65*(1), 118–140. https://doi.org/10.1016/j.cogpsych.2012.02.002

Berthoud-Papandropoulou, I. (1978). An experimental study of children's ideas about language.

Benedict H. Early lexical development: comprehension and production. J Child Lang. 1979;6:183–200.

Brown, R. (2013). *A First Language: The Early Stages*. Cambridge, MA and London, England: Harvard University Press. https://doi.org/10.4159/harvard.9780674732469

Callanan, M. A., & Sabbagh, M. A. (2004). Multiple labels for objects in conversations with young children: parents' language and children's developing expectations about word meanings. *Developmental Psychology*, *40*(5), 746–763. https://doi.org/10.1037/0012-1649.40.5.746

Clark, Eve. (1978). Awareness of Language: Some Evidence from what Children Say and Do. 10.1007/978-3-642-67155-5_2.

Clark, Eve & Andersen, E.S.. (1979). Spontaneous repairs: Awareness in the process of acquiring language. Papers and Reports on Child Language Development. 16. 1-12.

de Villiers, Peter & Villiers, Jill. (1972). Early judgments of semantic and syntactic acceptability by children. Journal of psycholinguistic research. 1. 299-310. 10.1007/BF01067785.

De Villiers, P. A., & De Villiers, J. G. (2013). Early language. In *Early Language*. Harvard University Press.

Doherty, M., & Perner, J. (1998). Metalinguistic awareness and theory of mind: Just two words for the same thing? *Cognitive Development*, *13*(3), 279–305. https://doi.org/10.1016/s0885-2014(98)90012-0

Donaldson, M. (1978). Children's Minds. Glasgow, 1978.The Aboriginal Child at School. 8. 55-56. 10.1017/S0310582200010968.

Erickson, Lucy & Thiessen, Erik. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. Developmental Review. 37. 10.1016/j.dr.2015.05.002.

Goldstone, Robert L., and David Landy. "Domain‐creating constraints." Cognitive Science 34.7 (2010): 1357-1377.

Gries, Stefan. (2005). Syntactic Priming: A Corpus-based Approach. Journal of psycholinguistic research. 34. 365-99. 10.1007/s10936-005-6139-3.

Harley, T. A. (2010). *Talking the talk: Language, psychology and science*. Hove, East Sussex: Psychology Press.

Karmiloff-Smith, A., Grant, J., Sims, K., Jones, M.-C., & Cuckle, P. (1996). Rethinking metalinguistic awareness: Representing and accessing knowledge about what counts as a word. *Cognition*, *58*(2), 197–219. https://doi.org/10.1016/0010-0277(95)00680-x

Markman, Ellen M (1990). "Constraints children place on word meanings." Cognitive science 14.1: 57-77.

Ninio, Anat. (2003). NO VERB IS AN ISLAND: NEGATIVE EVIDENCE ON THE VERB ISLAND HYPOTHESIS.

Ninio, A., Snow, C., Pan, B., & Rollins, P. (1994). Classifying communicative acts in children's interactions. *Journal of Communications Disorders, 27*, 157-188.

Marshall, John & Morton, John. (1978). On the Mechanics of Emma. 10.1007/978-3-642-67155-5_12.

Medina, T. N., Trueswell, J. C., Snedeker, J., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(22), 9014–9019

Piaget, J. (1928). *Judgment and reasoning in the child.* Harcourt, Brace.

Piccin, T. B., & Waxman, S. R. (2007). Why nouns trump verbs in word learning: new evidence from children and adults in the human simulation paradigm. *Language Learning and Development*, *3*(4), 295–323. https://doi.org/10.1080/15475440701377535

Pilulski JJ, Templeton S. *Teaching and Developing Vocabulary: Key to Long-Term Reading Success.* Boston, MA: Houghton Mifflin; 2004.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Schmitt N. Instructed second language vocabulary learning. Lang Teach Res. 2008;12:329–63.

Shore, R. (1997). *Rethinking the brain: New insights into early development*. New York: Families and Work Institute.

Tomasello, M. (1992). *First verbs : a case study of early grammatical development*. Cambridge University Press.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156.

https://doi.org/10.1016/j.cogpsych.2012.10.001

Tunmer, W. E., Pratt, C., Herriman, M. L., Bowey, J. (1984). *Metalinguistic awareness in children : theory, research, and implications* (Ser. Springer series in language and communication, 15). Springer-Verlag.

Vygotsky, L. (1962). *Thought and language.* (E. Hanfmann & G. Vakar, Eds.).

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

Acknowledgements

I am extremely grateful to my thesis advisor, Kevin Schaefer, for guiding me through the project. His brilliant advice on anxiety management, telling me to stress about one thing at a time, really helped me to soldier through the ups and downs of the research. I sincerely appreciate his kindness, encouragement, and one important suggestion: don't regret the decision I made. If I become an educator one day, I hope I can provide others with the same amount of professional and emotional support I received from Kevin.

I also want to thank Christina Esposito and Brooke Lea for being on my honors committee and providing me with valuable feedback that sharpens my thoughts and brings my work to a higher level.

Lastly, I would like to thank my friends, Albert Liu, Zhaoheng Li, Esther Han, Yutong Wu for listening to me talk about children and language. I also like to express my greatest gratitude to my parents. I could not have completed this thesis without their support and sympathetic ear.