Verb Islands in Child and Adult Language

Alix Kowalski and Charles Yang
University of Maryland and University of Pennsylvania

1. Introduction

How do we know what the child knows about her language? Regardless of the investigative methodology one chooses to study child language, the comparison between child and adult language seems inevitable. To wit, the universal adoption of Roger Brown's 90% criterion is exactly a metric that pits child language against adult forms.

While the literature does occasionally exaggerate children's achievements (see Yang 2002 for extensive criticism), there is no denying that many aspects of children's language use appear remarkably adult like. The recent,[1] usage-based, approach to language and learning turns this notion on its head. Perhaps we have been given the child too much credit.

Consider the acquisition of determiners and their usage in grammar, which has generated quite a bit of interest in recent acquisition research. It has been known since Valian's classic work (1986) that English learning children's use of determiners is virtually error free from very early on, which has been taken as evidence for early and abstract syntactic knowledge. The usage based approach not dispute these facts but suggests that children's use of determiners is not as productive as adults. Productivity is a quantitative measure of usage diversity, directly following the influential Verb Island Hypothesis (Tomasello 1992) to which we return momentarily. For instance, one could measure the percentage of (singular) nouns that appear with two determiners (e.g., *the* and *a*) out of the number that appeared with either of them. Pine & Lieven (1997) found the percentage to be low, which was taken to support the usage based perspective. But the failure to compare these measures against adult language apparently led to premature conclusions. In a rejoinder, Valian, Stewart & Solt (2008) applied the same metric to child directed speech and found that, paradoxically, mothers' determiner usage diversity is comparable to children's and both figures are fairly low (e.g., considerably below 50%, or less than half of the nouns appeared with both determiners.

---

[1] It should be mentioned that the usage based approach is quite similar to the pivot grammar hypothesis (Braine 1963).

The present paper addresses another empirical case that has played a central role in the usage based approach, the acquisition of verbal morphosyntax. One of the earliest proposals in the usage learning tradition is the Verb Island Hypothesis (Tomasello 1992): that a great majority of verbs are used in one or two constructions, that some verbs are used in more or different constructions than others, that most verbs are morphologically bare while relatively few appear in distinct inflectional forms (see Tomasello 2000a for a summary). These findings have been influential and been extended in later studies. For instance, Diessel & Tomasello (2001) observed that out of the possible combination between a matrix verb and the types of embedded clauses it can take, relatively few are actually attested in child speech.

We note, however, that these claims have not subject to a proper comparison against adult language, much like the determiner studies. In the original study of the Verb Island Hypothesis (Tomasello 1992), one finds numerous observations of child language but no statistical test to show the distributional patterns of children's verbal morphosyntax are *different* from adults'. (For that matter, nor was there any statistical test that children's language is in fact *consistent* with the predictions of the Verb Island Hypothesis.) In this paper, we provide a preliminary analysis of the distributional patterns in children's verbal syntax, specifically situating them in a comparative setting with adult language. Our analysis of corpora from the Harvard study shows that the diversity of verb usage for children and their mothers are remarkably similar. We conclude the study with a general discussion of the proper assessment of child grammar.

2. Data and Methods

We analyze and compare verbal morphosyntax of children and their mothers from the Harvard study (Adam: 2;3-4;10, Eve, 1;6-2;3, Sarah 2;3-5;1). We focus on *think, see, be, have, go* and *put*, verbs which feature prominently in previous work and are sufficiently frequent to allow meaningful comparisons. Each subject's transcripts were scanned for instances of the verbs of interest, and labeled using the tags provided by the CHILDES database.

The verbs *think* and *see* are notable for their wide range of arguments including clausal complements. Diessel and Tomasello (2001) conclude that *think* was a fixed phrase, used in the same sense as *maybe*, to indicate uncertainty. They claim that *think* was not used like a verb because they did not find evidence of non-first-person usage or instances of *think* used with negation. Verbs such as *see* (and *look*) are treated as

merely serving a linking purpose to full sentences in order to direct attention. Diessel and Tomasello interprete their results as children using utterance schemas and concrete phrases that are limited to specific constructions.

Our analysis is entirely distributionally based. This is to be contrasted with Diessel and Tomasello (2001), which classifies complement clauses by their illocutionary functions. We have found it difficult to provide consistent analysis of this type due to its subjective nature. For us, syntactic distribution was determined by comparing the co-occurrence of syntactic categories with the verb phrase. The syntactic categories are PRO(noun), N(oun), D(eter)miner, A(djective), (adve)RB, Q(ua)N(tifier), P(reposition), and CP (clausal phrase, in the case of *think*). A clause was considered to contain one of the lexical items only if the item functioned in a grammatically correct way, and if the phrase was not a repetition of their mother's speech.[2] TWe evaluate the claims of the Verb Island Hypothesis in light of our findings.

In addition to the verb environments, we assessed the inflectional flexibility of each verb, by tagging each item for its tense, person, and number. Verbs were identified as being in the first, second, or third person or having no subject (noS). Tenses included were present, progressive, infinitive, modal, past, and perfective. Use of the same verb in different tenses and different person and number agreements indicates the child's ability to manipulate the verb in reference to different events. As the mothers' speech was also transcribed in each session, these items were coded for comparison.

3. Results

Think Though *think* is suggested to be "fixed" in the first person present tense, we see it used with varied person and number agreements as well tenses. The frequency distribution is close between child and mother. Adam in particular used *think* in non-first person 27% of time and his mother did 12% of the time. He used it in non-present tense 12% of time; his mother 7.3%. Diessel and Tosmasello (2001) pointed to a high frequency of complement clauses as evidence of formulaic use, but we see just a high of occurrence in the speech of adults. Adam and his mother, for example, used complement clauses very frequently, Adam

[2] Because a verb may appear in multiple syntactic contexts, the percentages of its occurrences over all contexts may sum to a value exceeding 100%.

84% of the time and his mother 92%. Both Adam and his mother used exactly one overt complementizer for *think*, but that low figure is consistent with corpus findings of newspapers (Kearns 2007).

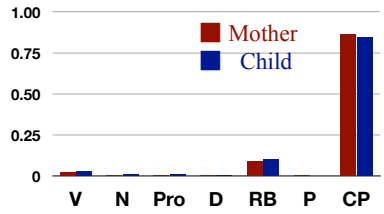


Figure 1. Syntactic distribution of *think*.

See The word *see* provided sufficient data to divide the children's usage into 3 stages based on mean length of utterance (MLU): less than 2.5 words, between 2.5 and 3.5 words, and over 3.5 words. These early, middle, and late stages give a depiction of language development that can be more accurately compared between children than chronological age.

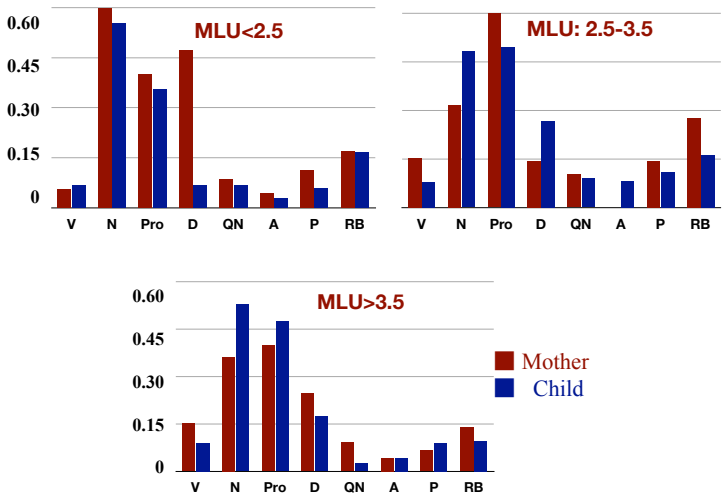


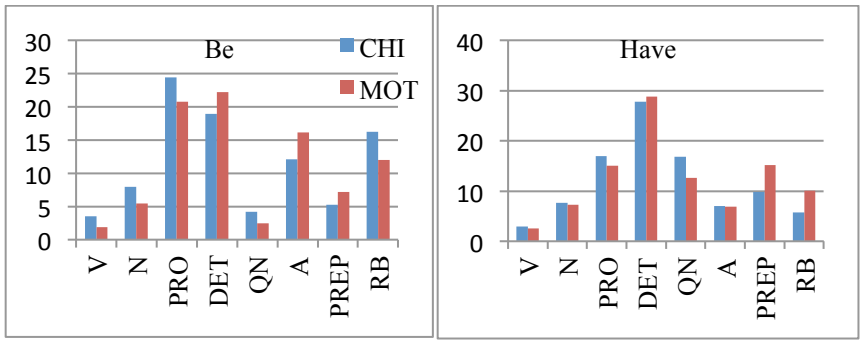
Figure 2. Distribution of verb environments of *see* across three MLU stages

Again, we see similar results for all three children and little difference across the three MLU stages. Distribution of verb environment does not significantly differ between the child and mother's utterances. The only exception is the use of determiners. The number of times *see* is used with a determiner appears to follow a developmental trajectory that the other syntactic contexts speech do not. This is more likely due to an independent issue of phonological development, wherein unstressed syllables like *the* and *a* are dropped in earlier stages of language acquisition.

Though Diessel and Tomasello (2001) claim that *see* occurs almost exclusively in the imperative form, only 36.1% of the utterances of *see* across all MLU stages are in the infinitive form. The children use the infinitive less frequently than the mothers in the MLU <2.5 stage (children: 28.1%; mothers; 40.1%) and the MLU >3.5 stage (children: 31.7%; mothers: 37%). In the middle stage of MLU 2.5-3.5, the children use the imperative 46.6% of the time, and the mothers 40.2%. Both children and mothers use the progressive and past tenses less than 20% of the time, and the perfective less than 10% in all MLU stages. Negation constitutes a low 4%, but that's similar to the mothers' usage (5%).

Though children and mothers use *think* and *see* with comparable morphosyntactic distributions, more frequent verbs are more likely to be treated as fixed expressions and thus may offer support for the usage-based learning approach. We thus turn to the highly frequent matrix verbs *be, have*, *go*, and *put*.

Be, Have, Go and Put



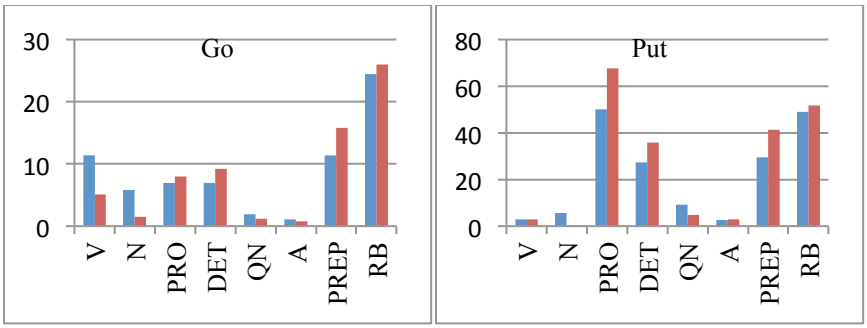


Figure 3. Distribution of verb environments of *be, have, go,* and *put*.

For the most part, the percentage of each child's utterances in various verb frames follows the same distribution as the mother's utterances. The percentages are not exact, which suggest that children are not imitating child directed input. The fact that the language learners are using all relevant syntactic contexts contradicts the claims of the usage-based approach.

In addition to flexible verb frames, all three children showed variability in person, number, and tense agreements.

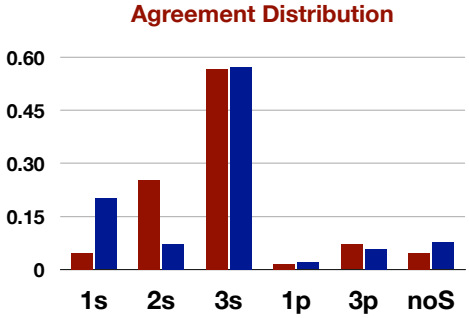


Figure 4. Distribution of person/number of *be, have, go*, *put* (noS: null subject)

Figure 4 shows the percentage of the subjects' utterances that are in the first, second, and third person singular, and first and third person plural for each target verb, compared to their mother's person/number agreements. Unlike Tomasello's assertion that most early utterances are in reference to the self and not others, the data shows that more of the children's speech is in the third person singular (Adam 51.15%, Eve 37.4%, Sarah 37.2%). All three children use the first person plural and

third person plural with similar frequency as their mothers. Adam and Sarah used first, second, and third singular agreements with the same frequency as their mothers. However, the percentage of utterances in the first, second, and third person singular are significantly different between Eve and Eve's mother. This discrepancy seems to stem from a difference in the frequency of third person singular agreement. Eve's mother tends to use verbs in the third person (55%), whereas Eve's agreement is more evenly distributed between the first and third person (25.18%, 37.44%). Adam and Sarah's utterances shows the same phenomenon, but to a lesser (non-significant) extent. The difference could be due to the nature of their conversational roles as mothers and children, but it should not be taken as evidence against any of the children's flexibility of use, because these children are using more varied agreements than their mothers.

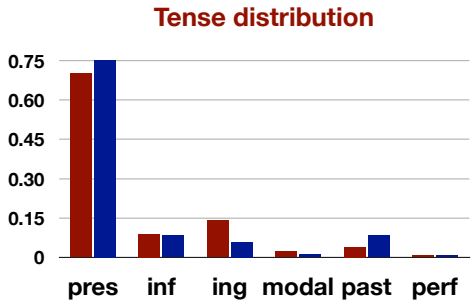


Figure 5. Distribution of tense of *be, have, go, put* (Adam, Eve, Sarah).

Figure 5 represents the percentage of the subjects' utterances in the present, infinitive, progressive, past, and perfective tenses, as well as data from their mothers. Verbs in clauses were necessarily in their infinitive form, and are noted as such (inf). Any unconjugated verbs, for example "I be back", or "She have one," were not included in the graphs because those mistakes were rare (0.29%) and are even seen in the mothers' speech (0.24%), when they imitate children's grammatical errors. One immediately sees the high percentage of utterances in the present tense, for all subjects and their mothers. The children used the present tense 71.7% of the time, the mothers 72.8%. The past and perfect tenses show similarly close percentages. With the use of the present tense high for both child and mother, the use of progressive, past, and perfective tenses is necessarily quite low for everyone.

In sum, it seems quite clear that the diversity of verb usage is remarkably similar for children and their mothers. Before we discuss

these findings in the general setting of language learning, it is instructive to understand the statistical distribution of verbal syntax.

4. Islands Everywhere

No one who has studied the quantitative properties of natural language corpora would be surprised by the uneven distribution of morphosyntactic combinations. (The only surprise is that children and their mothers are so closely matched.) One of the most robust statistical properties of language is Zipf's law (1949), that relatively few words are highly frequent while many more occur rarely or only once in a linguistic sample, occupying a characteristic long tail. Similar observations have been made for *n*-grams (Ha et al. 2002) and phrase structure rules (Buttery & Korhonen 2005). In fact, as the combinatorial possibilities of multiple word expressions grow exponentially, even fewer types will appear frequently with the vast majority of perfectly grammatical forms never attested. This has been referred to as the *sparse data problem*, an inherent challenge in computational linguistics.

Given the sparse data problem, it is impossible to expect anything other than verb islands in a sample of language use. We examine constructions that involve a transitive verb and its nominal objects, including pronouns and noun phrases. Following the definition of "sentence frame" in Tomasello's original Verb Island study (1992, p242), each unique lexical item in the object position counts as a unique construction for the verb.

We extracted 1.1 million adult sentences from the CHILDES database. After applying a state of the art Part-of-Speech tagger (Brill 1995), we extracted the top 15 most frequent verbs immediately followed by a nominal. For each verb, we count the frequencies of its top 10 most frequent constructions, which are defined as the verb followed a unique lexical item in the object position (e.g., "ask him" and "ask John" are different constructions, following Tomasello 1992). The results are given in Table 1.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
put	401	164	124	15	12	12	11	10	8	5
tell	245	64	49	49	45	36	22	16	14	13
see	152	100	38	32	28	21	14	14	12	11
want	158	83	36	24	19	15	13	9	5	4
let	238	38	32	23	22	17	8	6	3	3

give	115	92	59	32	31	7	5	5	5	5
take	130	57	30	21	18	15	14	9	8	7
show	100	34	27	21	19	17	12	8	7	7
got	58	37	14	12	11	9	7	7	7	4
ask	45	41	27	24	12	10	8	8	4	2
make	67	20	12	10	9	7	7	4	3	2
eat	67	42	14	8	6	5	5	3	3	3
like	39	13	9	6	4	4	4	4	3	3
bring	43	30	17	15	10	10	3	3	3	3
hear	46	22	13	9	6	4	4	3	3	3
total	1904	838	501	301	252	189	137	109	88	75

Table 1. Frequencies of the most frequent nominal object frames for the most frequent transitive verbs in 1.1 million child-directed utterances.

As one can see in Table 1, the frequencies of constructions decline rapidly as their frequency rank increases—and these are among the most frequent verbs in a sample of 1.1 million utterances In light of these distributional reality of language, it is impossible to expect anything other than Verb Islands, especially when we deal with much smaller sample sizes as is usually the case with child language.

5. Discussion

This preliminary analysis of children's morphosyntactic usage reveals much similarity and continuity with adult forms. The findings are consistent with the interpretation that children's language closely matches adults', and are inconsistent with claims of limited, item-bound, productivity from the usage based literature.

Additional tests need to be carried out to assess children's grammar with greater reliability. After all, an advocate of usage based learning may wish to claim that we find the similarities between child and adult language because we have been *overestimating* adult language. If the child just repeats back what the adult says, or only just the most frequently used expressions, without a systematic grammar like the adult's, she would also be error free, leading to the impression of linguistic mastery (see Tomasello 2000b for exactly such a proposal).

To settle the grammar vs. storage and retrieval disrupt would require precise quantitative predictions of what each approach predicts, which

has been lacking from both sides. We direct the reader to our treatment of these issues elsewhere (Yang 2011), where the usage based approach is again found wanting.

References

Braine, M. (1963). The ontogeny of English phrase structure: The first phase. *Language,* 39, 3-13.

Buttery, P. & Korhonen, A. (2005). Large-scale analysis of verb subcategorization differences between child directed speech and adult speech. Workshop on the Identification and Representation of Verb Features and Verb Classes, Saarland University.

Diessel, H. & M. Tomasello. (2001). The acquisition of finite complement clauses inEnglish: A corpus-based analysis. *Cognitive Linguistics*. 12(2). 97-142.

Ha, Le Quan, Sicilia-Garcia, E. I., Ming, Ji. & Smith, F. J. (2002). Extension of Zipf's law to words and phrases. *Proceedings of the 19th International Conference on Computational Linguistics*. 315-320.

Kearns, K. (2007). Epistemic verbs and zero complementizer. *English Language and Linguistics*. 11. 475-505.

Pine, J. & E. Lieven, Elena. (1997). Slot and frame patterns in the development of the determiner category. *Applied Psycholinguistics*, 18, 123-138.

Tomasello, Michael. (1992). *First verbs*. Harvard University Press.

Tomasello, M. (2000a). Do young children have adult syntactic competence. *Cognition,* 74, 209-253.

Tomasello, M. (2000b). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics,* 11, 61-82.

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology,* 22, 562-579.

Valian, V., Solt, S, & J. Stewart. (2008). Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language*. 36. 743-778.

Yang, C. (2002). *Knowledge and Learning in Natural Language.* New York: Oxford University Press.

Yang, C. (2011). A statistical test for grammar. *ACL* 2011.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Addison-Wesley.