Hi, Thanks, and Goodbye: From the Wug Test to AI

Jean Berko Gleason

1. Introduction

The creation of the Wug Test, which showed that even young children have an internalized representation of their language, took place near the beginning of the cognitive revolution. In this paper, I will first describe some of that background and my own personal and academic part in that history, along with a discussion of what endures and what does not, including relevant examples from the Wug Test itself. I will then note some of the much broader questions and psycholinguistic domains, like formulaic speech, that emerged and were studied over the ensuing years in our own laboratory, or in naturalistic settings, such as on the street on Halloween. A final discussion considers some contemporary issues and points about future research, including the ability of AI to pass the Wug Test and answer other questions we might pose.

2. Background—personal and linguistic

Through a remarkable combination of luck and happenstance my article that describes the Wug Test was published in 1958, just a couple of months after I submitted my doctoral dissertation, on which the article was based, to Radcliffe-Harvard. It was called "The Child's Learning of English Morphology" (Berko, 1958) but over the years the image of the little Wug I drew as part of my study has appealed to people so much that the whole set of questions I asked is just called The Wug Test, and people in various parts of the world have put wug tattoos on their chests and wug couples on their wedding cakes. Since that ancient time, my colleagues and I have conducted a great variety of psycholinguistic studies, not just with American children and families, but in other areas as well that vary widely, from work with Roma in little towns in Hungary to studies with aphasic patients in our local hospitals. Most recently, we have turned our thoughts to Artificial Intelligence and asked some questions based on the Wug Test to ChatGPT4 rather than to a 4-year-old, in an effort to shed light on the kind of language processing LLMs (Large Language Models) engage in. Do they develop productive rules for

*Jean Berko Gleason, Boston University: gleason@bu.edu. With many thanks to the organizers of the 50th anniversary meeting of the Boston University Conference on Language Development and to Nan Bernstein Ratner, who provided an in person introduction to the earlier spoken version of this paper, while I appeared on the Zoom screen.

 Proceedings of the 50th annual Boston University Conference on Language Development, ed. Romi Hill et al., 1-14. Somerville, MA: Cascadilla Press.

dealing with new material, or is their output based purely on pattern recognition?

The Wug Test revealed that even very young children have sophisticated knowledge about language that goes beyond what they have heard or could imitate. The world has changed since 1957 when it was designed, and I need to point out some of the things we would have to change in a contemporary Wug Test and what that says about the nature of research: we need to remember that we are creatures of our times and that the questions we ask and the way we ask them are constrained by the contemporary world we live in. So there were things we did then that we would not do today. And 50 years from now we or others may look back on what we do today and find some things quite outrageous. But first: how did we get here? I was one of the people who was present at the beginning of the cognitive revolution that included a new and intense interest in thinking and on the inner representation of language, so I will begin with a personal explanation of how I got there.

My own interest in language has been lifelong. I was born in Cleveland, Ohio to immigrant parents from Hungary, and lived there until I left for Radcliffe College, near Boston, at the age of 17. My parents were from the part of Hungary, now Romania, known as Transylvania, and though my mother was a native speaker of Hungarian my father had been brought to the U.S. as a 10-year-old and he was an English speaker. But I heard lot of Hungarian growing up: Uncles who pinched my cheek and said, "*Szép kislány!*" which means "*pretty little girl.*" And my mother used Hungarian for words she didn't know in English, like body parts, terms of endearment, dog poop, and, so forth. My mother also talked in her sleep in Hungarian, so in the night I would hear her deep Hungarian voice calling out "*Jaj! Istenem!*" Which means *"Oh my god!"* She also spoke in Hungarian on the phone to her sister quite often and I would try to listen in, but what I heard most often was her saying *"Nem tudom! Nem tudom!"* And it took me years to find out that *nem tudom* means *"I don't know."*

I had an older brother Marty, too. He was 6 years older than me and he was very sensitive and intelligent. He ultimately got a PhD in experimental psychology from Cornell University. But he had cerebral palsy as the result of a long difficult birth as a large baby being born to a very little woman. People had trouble understanding what he said. I always understood him, so early on I served as his translator and learned to pay particular attention to language that way. The combination of the Hungarian background, the translating for my brother, and the lifelong interest in language had a powerful influence on my choices in life.

Radcliffe College was actually a way for women to go to Harvard. Harvard did not admit women, but women could apply to Radcliffe, which had its own administration, and on acceptance they would be housed in dormitories a mile from Harvard Yard, where they had to have all their meals and follow various strict rules of conduct. And to the surprise of the men of Harvard, Radcliffe women were their fellow students in their classes. There

was never a separate Radcliffe faculty or curriculum. We were not, however, allowed to play in the band or write for the Harvard Crimson, or go into Harvard's Lamont Library, or even to wear pants on our walks to class in the middle of the gelid New England winter. But we were grateful to be there because of the extraordinary education we got.

As an undergraduate, I majored in history and literature, specializing in the nineteenth century, which entailed wonderful courses like Proust, Joyce, and Mann, taught by the incomparable Harry Levin, and The Symbolist Poets, taught by Renato Poggioli. I couldn't resist taking elective languages: French, German, Latin, Norwegian, Sanskrit, for instance. I read *The Wild Duck* by Ibsen in Norwegian. And in Sanskrit, we read what everyone reads, the passage from the Mahābhārata in *Lanman's Sanskrit Reader* (1906), all about a prince named Nala and the girl he wants to marry. I loved Sanskrit, which was taught by Daniel Henry Holmes Ingalls in a little room on an upper floor of Widener Library. But it wasn't until my senior year when I took a course by a young assistant professor named Roger Brown called the Psychology of Language that I realized that that was what I needed to do. How do we acquire language, lose it, store it, understand it, produce it? This was psycholinguistics, but it did not yet exist as a field – it was 1952 and just at the dawn of the cognitive revolution, when researchers really began to study mental processes.

3. The beginnings of the cognitive revolution

The cognitive revolution became an important part of the intellectual and psychological history of our times. It was made possible by the confluence of a number of things, both new ideas and new technology, such as the new technology that produced the tape recorder. In those days, tape recorders were huge contraptions, the size of a large suitcase, and they could store much less than any mobile phone today. But for the first time we could capture language and analyze it later. They were called tape recorders because recordings were made on a reel of magnetic tape. We had to hire someone to carry our tape recorder for us, but the tapes made it possible to listen repeatedly, and to write down what we heard. Now we had real data, for the first time. New ideas helped to put linguistics and psychology together. Descriptive linguistics had long been a part of the anthropologist's arsenal, but now the broader academic world began to appreciate the componential nature of language and the subsystems like phonology and morphology that could be described. We began to ask: Is there a psychological reality to the units the linguist describes? We talked about *isomorphism*: the linguist may describe how the plural is formed in English, for instance, but does that mean that speakers have those same units and rules in their brains? Roger Brown's course raised questions of this kind, and there was clearly a whole new world to explore.

In order to pursue these questions, I switched fields to linguistics for a master's degree and then to both linguistics and psychology for the doctorate

and this involved having a committee oversee what my course of study would be. This included taking all the qualifying courses in both linguistics and psychology, including both Old French and clinical psychology. Roger Brown was my adviser, and after getting my degree I spent a postdoctoral year with him while he was at MIT, prior to returning to Harvard. We published a number of papers and chapters together during that time. Roger was a wonderfully creative, intelligent, and generous mentor. Roger's students still love him—I know this from a recent correspondence with Steven Pinker, Howard Gardner, Ellen Winner, Helen-Tager Flusberg, Dan Slobin, Peter and Jill DeVilliers, Kenji Hakuta, and a number of other eminent people in our field who were Roger's students in the cohorts that followed mine. My cohort was just three people: myself, Eric Lenneberg, and Don Hildum. There is no room here to talk about the incredible body of research produced by all of these people and others, but if you are interested in Roger Brown's legacy, there is a very comprehensive book by Frank Kessel (1988) called *The Development of Language and Language Researchers: Essays in Honor of Roger Brown.*

Noam Chomsky was of course just getting established at this time, but we were not involved with him. In designing our studies we were able to rely on descriptive linguistics, which had long been used by anthropologists and also missionaries who wanted to be able to analyze and use any language they encountered at home or in the field. So I got to take field methods courses with Dell Hymes and Charles Ferguson. Our textbook was *An Introduction to Descriptive Linguistics* by Henry Allan Gleason, Jr. (1955). Gleason was himself both a linguist and a missionary who taught at the Hartford Seminary Foundation, and later at the University of Toronto. By an odd coincidence, I met his younger brother Andrew Gleason when I was a graduate student and he was a young mathematics professor at Harvard, and we were married in 1959. Andy and I became the parents of three daughters, Katherine, Pam, and Cynthia, born in the three years just after I got my PhD and we were married for 49 years until his death.

Roger Brown and I set out to see if even young children between the ages of 4 and 7 have internalized rules for dealing with language: do they just imitate what they hear, or are they able to construct new utterances and words they have never heard before? All of this was before Roger embarked on his famous 1973 study, *A First Language*, of the three children he called Adam and Eve and Sarah, whose transcripts provided the basis for so many dissertations. And there is an update on one of those early child participants. We were in touch with Adam not long ago. He grew up to be a CEO and philanthropist. He is currently finishing his term as a college president because he has been elected lieutenant governor of one of our most important states!

I settled on the basic parts of the English morphological system for my study. This included the inflectional and derivational morphology found in young children's vocabularies, such as plurals, past and progressive tenses,

possessives, diminutives, and adjectives. Could children make plurals or past tenses of words they had never heard before? Could they turn an unknown noun into an adjective, since they knew that something covered with *dirt* could be called *dirty*?

As part of the my study I also asked the children about some compound words in their vocabularies—such as why did they think that a birthday is called a *birthday*. This did not require pictures, but we found that children of different ages gave typical metalinguistic answers. The youngest children simply insisted that it was called a *birthday* because that is what it is. Somewhat older children named salient features of the event, like you get presents or eat cake. Etymological-appearing answers were produced when a salient feature was also part of the name: a fireplace is called a *fireplace* because has fire in it. Only the oldest children sometimes, but rarely, explained both parts of the word: a birthday is the day of your birth, for instance. It is useful that I asked those questions, because recently I had a chance to ask the same questions of present day AI to see how ChatGPT4 compares with a 4-year-old in explaining why a birthday is called a *birthday*.

I made the morphology test using nonce words that children had never heard before to find out if they knew the rules: Do they just imitate what they hear from adults or can they actually create a plural of a word they have never heard before? I couldn't use a real word like *dog*, because of course they no doubt heard people say "dogs." But no one had ever talked about *wugs*. I made the test myself, drawing pictures to represent the imaginary things like *wugs* and *biks* and *tors* and *nizzes*, and then gave the test to 80 young children between the ages of 4 and 7. The results were clear and compelling. Even a 4-year-old knew that if you had one *wug* and then another you now had two *wugs*. I wrote up the results and got my degree in June of 1958. A month later I presented the research at a meeting on aphasia at the Boston VA Hospital and Uriel Weinreich, who was the editor of the journal *Word* was one of the other participants. Uriel Weinreich was an expert on Yiddish, and after he heard my talk he went home and asked his young daughter a series of Wug Test questions using his own made up rather Yiddish-like words: "This is a *mish*. Now there is another one." He was stunned when she answered just the way I said young children would. The next day he asked for a copy of my work and published it right away. So that is how I managed to get my degree in June of 1958 and have the article appear in print two months later in the August issue of *Word*. Life is full of happenstance and luck.

4. Basic and enduring findings of the Wug Test and some good and bad news

The Wug Test investigates some basic features of English inflectional and derivational morphology and word compounding that all speakers of the language must acquire. It is helpful to remember that morphology refers to basic units of meaning in a language. In English, we have free and bound

morphemes. For instance, the word *cats* consists if two morphemes, the free morpheme *cat*, which can stand alone, and the bound morpheme *–s* that marks the plural when appended to a free morpheme.

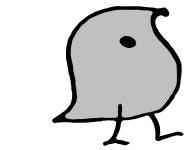
Inflectional morphology does not alter the meaning or part of speech of the basic word. Rather, it has a grammatical function that signals such things as tense, person, number, and case. Inflectional morphology is required in order to make meaningful utterances, and we found that all of the English inflections were present in the language of the children we were studying.

Derivational morphology by contrast derives a new word with a new meaning, and can change the part of speech. We can derive the agent *baker*, for instance, by adding the bound morpheme *–er* to the verb *bake*. Derivational morphemes can be seen in words like *unhappy, happiness, doggie,* and *thankful.* Among the derivational morphemes, diminutives and agentives are very common in the speech both of and to young children. That led us to hypothesize that the children might tell us that a man whose job is *to zib* is a *zibber* and that a very tiny *wug* is a *wuggie*.

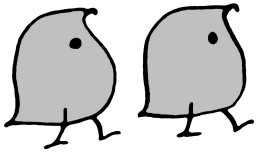
Compounds typically contain two free morphemes joined together with altered stress. Words like *hotdog* and *hot dog* are easily recognized as referring to either sausages or warm canines by adult speakers, even when pronounced in isolation, but this ability is acquired fairly late in childhood, not until children are around 9 years old. However, compound words are very frequent in the speech of young children.

The Wug Test pictures were in color, and they were on 27 cards with illustrations and simple questions of the general format: This is a Wug. Now there is another one. There are two of them. There are two ____? For the past tense elicitation the question was similarly: This is a man who knows how to rick. He is ricking. He did the same thing yesterday. What did he do yesterday? Yesterday he ____?

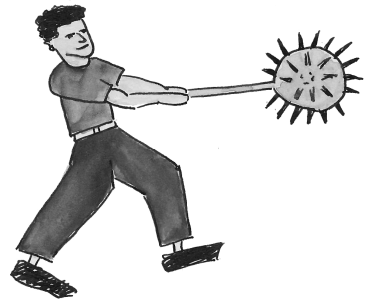
An example of these iconic questions can be seen in Figure 1.



This is a WUG.



Now there is another one.
There are two of them.
There are two _____.



This is a man who knows how to RICK.
He is RICKING. He did the same thing
yesterday. What did he do yesterday?

Yesterday he _____.

Figure 1. Two example items from the Wug Test. As described in Berko, Jean. (1958).

The English morphology we set out to investigate with these young children covered the following:

- The plural, possessive, and third person of the verb. These have similar forms: The sound of the affix is –z –s or –iz depending on the last sound of the root of the word, as in *wads, wits,* and *witches.*
- The progressive of the verb. This is invariant –ing as in *eating.*
- Past tense, which takes the sound –d –t or –id, as in *played, hopped,* and *planted.*
- The diminutive formed with –y as in *doggy*
- The agentive formed with –er as in *baker.*
- The adjective in –y, as in *dirty,* and the comparative and superlative, as in *dirtier,* and *dirtiest*

Our findings were fairly dramatic. In the first place, unlike Large Language Models (LLMs), children understood the questions. Moreover, even 4-year-olds could make plurals and pasts of words they had never heard before. Children produced regular and simple forms, clearly following rules, but simpler ones than those used by adult speakers: to make a plural, they added the appropriate –s or –z but not –iz, basically treating words that ended in –s and other sibilants as if they were already plural. They used a similar strategy in making past tenses. To make a past, they added –t or –d but not –id, treating nonsense words that ended in –t or –d as if they were already in the past. Children did not use irregular patterns, and the ability to add the required –iz or –id endings increased with age, although even 7-year-olds had

difficulty with these. They found the progressive of the verb, which is invariant and thus does not require a choice of forms, particularly easy.

Children's control of derivational morphology took a back seat to their production of the inflectional affixes. Even though their vocabulary and speech to them contain many diminutives, they did not use common diminutive –y. More commonly, they produced a compound word when asked what they would call a very tiny wug. Rather than a *wuggy*, they were liable to produce *baby wug*, and there was no evidence of the many diminutives adults offer when asked the same question: *wuglet, wugling, wuggette, wuggie.* The agentive –er was provided only rarely. Where adults all agree that a man whose job is *to zib* is a *zibber*, children typically said he was a z*ibman*, again using the compound word stress. They also used compounds rather than adjectives formed with –y (a dog covered with *quirks* was a *quirkdog*) and they were fairly unfamiliar with the comparative and superlative forms of the adjective.

We were thus able to show with the Wug Test that children have mental representation of their language, and that they are doing much more than repeating what they have heard from the adults around them. Their language is obviously not simply rote memory or the result of direct teaching. But we are not sure how they do what they do. Are they forming analogies? Or are they abstracting rules? We have been asking these questions about their cognitive processes since 1958, and similar results in many languages indicate that universal processes are involved. With the advent of Artificial Intelligence, it is no surprise that researchers are now giving forms of the Wug Test to Large Language Models, asking some similar questions about pattern recognition and abstract rules in the machine world. How does ChatGPT4 compare with to a 4-year-old in processing language?

Before turning from the Wug Test, it is necessary to sum up a bit both the good news and the bad news about this study. To begin with, the bad news is that Wug Test is incredibly sexist. Every human in it is a man. Man was the unmarked human in my era. If you read the article I wrote, "The Child's Learning of English Morphology," you will see that the child is always a *he* and never a *she*. That was the writing rule. But my own drawing of only men is rather shocking. If we did the test again I'm sure we would think to add some females. But who knows what we are doing today that will in 50 or 60 years be seen as outrageous. We are creatures of our times.

The good news is that after all these years, the test replicates. The test is robust. It was the first experimental demonstration that children really operate with rules, not just memorized words. And if you give it to a 4-year-old today, you will get similar answers. This is not true of many studies in psychology and we are in a replication failure crisis according to some. So the Wug Test endures.

But why does the wug have a special place in our hearts? Yes, we know the test showed something that had never been shown before. And it used made up or nonce words. Other research used made up words as well, like

blickets, which are well known, but not well loved. There are no blicket t-shirts or coffee mugs, that I know of. I think it's because all our nonce words refer to living things, mostly animals. Wugs are little animals, and we love animals. This love is shared with children. When you take a walk with a little child, you point out living things: "Look at the birdie!" Not: "Look at the parking meter!" Parents pass on a world view that children are already disposed to share, and that is that we love and pay attention to living things, especially animals, but also plants rather than rocks. Children's early vocabulary is full of animal terms, including quite rare words like *elephant* and *zebra*. My colleague Brenda Phillips at the University of Derby and I have looked, in work in progress, at what parents call children's attention to. In a group of over 60 children, we looked at hundreds of phrases where the parent used the phrases *look at* or *see the*. Eleven of the 30 most frequent target words used by parents referred to animals. The most common animal term used was *kitty*. Adults may believe that living things are of particular interest to children, and adults themselves may be drawn to animals as part of our genetic inheritance—we are all attuned to the natural world. So that is one explanation for why we like our wugs so much.

5. Hi, thanks, and goodbye: formulaic language

I turn now to one other kind of research we have conducted that is very different from the very cognitive based inquiry built into the work on the acquisition of morphology, which is not explicitly trained. We don't teach our children how to make possessives, for instance. And that is our work on the acquisition of formulaic language or what we called *routines* at the time. As research on child language progressed, many of us turned to broader questions such as social aspects of language and to more diverse populations of children to study. I missed a lot of the intense work on syntax conducted in the 1960s, much of it inspired by Chomsky's work *Syntactic Structures* that came out in 1957, that had researchers thinking about transformations, phrase structure rules, deep structure, kernel sentences, etc. Much of that has now been, as they say, updated, as syntactic theory has changed. My own interest turned to questions often related to differences and variation among speakers rather than the universals and innate characteristics emphasized by the research on syntax. These included, among other things, speech to children, gender differences, work in the Roma community, fathers' speech, and a number of other topics that shed light on how language makes us the people we are. So we increasingly studied interaction with other people rather than just the emergence of language in children.

In our own lab at BU we turned our attention to these broader aspects of language, for instance to the use of what we called *routines*, and what is now called *formulaic speech*. And that is where the Hi, thanks, and goodbye part of my title comes from. Early on, I had noticed that although the general consensus was that babies' first words were something referential like *mama*

and *dada*, even before this parents were actively teaching their infants to say *bye bye*. Even long before the babies could say anything, parents were teaching their babies to wave goodbye. Perhaps the best known of our studies in that period was one that Sandra Weintraub and I conducted over several Halloweens, published in 1976. We collected data on the American custom where children go door to door on Halloween hoping for gifts of candy, and are expected to say *"Trick or treat"* when the door is opened, "*Thank you"* when they get the candy, and "*Goodbye"* on leaving. This is a very constrained routine, but, for instance, saying *"Good evening"* would be totally unexpected unless the child were dressed as Dracula, and then it would have to be *"Good evening"* in a Dracula accent.

We found that parents were actively teaching children to perform this routine. For instance, one mother stood on the curb and called out to her son, "Don't forget to say trick or treat and thank you!" The children themselves also behaved in a very regular way that resembled developmental stages: the youngest said nothing, somewhat older kids said only *"Trick or treat."* The next oldest said *"Trick or treat"* and *"Thank you,"* and the oldest performed the whole routine, usually without knowing the meaning of the words *trick* and *treat*. Routines are performative in nature and explicitly taught. As social conventions, they also change over time. On the most recent Halloween when I answered the door I did not hear many children say *"Trick or treat."* There were a number of parents who came right to the door and said *"Thank you*" for their children.

Routines are a special kind of language children learn as members of their generation in a particular society, and they are important markers of membership. They differ from referential language in a number of ways. As noted earlier, they are taught explicitly and may be some of children's earliest expressions. In English, parents mark this teaching with the word *say*. Routines are performative and taught without semantic explanations. We teach children to say *"Thank you"* when they get a birthday present, but we don't spend much time on what it means to be thankful. Rather, we emphasize the importance of uttering the formula appropriately rather than knowing what it means. As reflections of society, which changes over time, routines change as well. But they are also very important, as is evidenced, for instance, by the consequences of failing to greet a friend you meet on the street.

There is no space here to recount everything we did over the course of many years, but to sum up where I think our own work has gone as well as that of many others in the child language world, I would say that we have progressed from a purely cognitive approach to one that includes social and emotional components. We now include speech to the child as well as the child's language, and the speech of not just mothers, but fathers and all others. The range of children included in research is now immense, with many different languages, and children both typical and atypical from different kinds of societies.

6. LLMs and the Wug Test

And now we are on the brink, supposedly, of a new technological revolution in artificial intelligence that is impacting all our lives. We don't really know what kind of intelligence LLMs have, for instance to what extent they can employ logic and inference and produce language that is more than a reflection of their ability to pull information from vast amounts of text that they have access to. Scholars in that field like Gary Marcus (2024) have distinguished 3 types of artificial intelligence that LLMs aspire to: the first is Artificial Narrow Intelligence, and this means that they are able to perform fairly well in narrow domains, such as providing driving directions. This is something that they are good at. A second kind of Artificial intelligence is Artificial General Intelligence, which would be a much broader kind of ability that goes beyond finding examples or the likely next words in the database. It would involve inference and reasoning, knowing when they do not know something, and remembering what they said earlier. Basically it means that they would be like humans. There is disagreement as to whether LLMs are about to have general intelligence, or when they will have it, or if they ever will. The third kind of artificial intelligence is ASI, artificial super intelligence, and it is the stuff of science fiction: machines that become smarter than we are, that can replicate themselves and do more than we can, and will take over and rule the world.

The Wug Test is an interesting way to see if LLMs have any of that general intelligence, since we can give them tests with words they do not know and cannot therefore look up in a database. My colleagues Patrick Juola, George Mikros, Nan Bernstein Ratner, Lise Menn, and I (2024) set out to do that. Not surprisingly, the idea of presenting LLMs with versions of the Wug Test also occurred to a number of other researchers. One group in Spain, for instance, (Pantelidou, Leivada, & Morosi, 2025) tested a version of the Wug Test they created with nonce words in 4 different languages, English, Spanish, Catalan, and Greek. They wanted to know if success in producing the forms called for was better predicted by their grammatical complexity or by the size of the database available in that language. Their results indicated that the LLMs did a good job of producing the correct morphology, but the degree of accuracy was not related to the complexity of the form so much as to the size of the database available and the number of speakers of the language. Results were better with English and Spanish than with Catalan or Greek, leading these researchers to conclude that the machines were exhibiting excellent pattern learning but very little in the way of structural grasp of the sort that humans have.

In our own study, we asked ChatGPT4 a variety of Wug Test items plus some additional ones using a bank of new nonce words that we made up. ChatGPT4 did very well with many inflectional items, providing appropriate plural forms, past tense verbs, possessives, and agentive nouns but could not process the prompt or succeed with progressives and adjectives. It also had

trouble understanding the instructions. For instance, it had problems understanding that the nonce word *blick* was not a typo for *black*. It then warned us to be careful when discussing matters of race. Young children, of course, do not do this. Some of the AI derivational answers were inexplicable, for instance where the LLM told us that a "very tiny *skring*" is "a very tiny." In comparing children and LLMs it is also helpful to consider that LLMs do not do as well as 4 year olds in some things even though they have vastly more experience. Researchers have pointed out that to date LLMs have been exposed to more than 2 trillion words, whereas one human's database for an entire lifetime is estimated at 3 billion words.

I was also curious to know if the LLM could think like a child, so I conducted a mini study using the part of the Wug Test that isn't pictured but simply asked children why they think some compound words in their vocabulary are called what they are. We found, as noted earlier, that very young children simply insisted that, e.g., a birthday is called a *birthday* because that is what it is. Somewhat older children focused on a salient feature: Thanksgiving is called that because you eat turkey. Only a very few first graders paid attention to both parts of any compound word. And the highest percentage of etymological responses was for Thanksgiving, but even there 67% of children said it is called *Thanksgiving* because you eat turkey, and said nothing about giving thanks, something that is often included in lessons.

I asked the LLM to answer the questions about compound words in just the way that a 4-year-old would, to say what a child would say.

General Instruction: I want you to answer the following questions the way a 4-year-old would answer.
Question: Why do you think a birthday is called a birthday?
LLM answer: Well, cause it's like a party day for when you got born! It's like a big Yay, you showed up!

Question: Why do you think Thanksgiving is called Thanksgiving?
LLM: You're just giving thanks! It's like a big feast to say "Hey, we're thankful for all the good stuff."

Question: Why do you think Friday is called Friday?
LLM: It's actually named after the Norse goddess Frigg, who was associated with love and beauty. So, yeah, it's like Frigg's day.
Question: Is that really what a 4-year-old would say?
LLM: Absolutely. It's all about making those concepts into tiny little pieces; 4-year-olds love that stuff.

These answers are obviously not possibly those that a 4-year-old might give. They are all etymological and simply use exclamation marks and words like *Yay!* but are combined with an explanation of both parts of the compound

word, which 4-year-olds do not do. The LLM at that point in time apparently did not have access to text in a database about how young children would answer these questions, and had no way to create anything other than standard definitions for the words.

The Wug Tests of various types with Large Language Models show that the LLM does well on regular inflections, especially in languages that have many speakers and a large body of data available. That does not imply that they are doing anything like what humans do in terms of processing. They are less successful with derivational morphology. And when it comes to explaining the words in a 4-year-old's vocabulary, they can talk like a stereotypical child, but they do not seem to be able to think like a child. There is no evidence of the organization of language that we see in people. At least for now, LLMs do not appear know what a preschooler knows about language and their performance only superficially resembles a child's. That does not mean that Artificial General Intelligence won't eventually be perfected but we are not there now.

7. Conclusion

Many things have happened in the child language world since 1958. We no longer see children as little Martians trying to decipher a strange code. Rather, they are little humans trying to join our society, and we understand that language development is a shared human enterprise that involves extensive interaction between young children and older speakers. It is not just a cognitive, but a social and emotional achievement as well. It is also a uniquely human achievement, and we do not know if machines will ever reach the level of General Intelligence that we see in children who can pass the Wug Test. Meanwhile, our job for the next 50 years is to carry on with our cooperative and human research.

References

Berko, Jean. (1958). The child's learning of English morphology. *Word*, *14,* 150-177.

Brown, Roger. (1973). *A first language: the early stages*. Cambridge, Mass. Harvard Univ. Press.

Chomsky, Noam. (1957). *Syntactic structures*. Mouton.

Gleason, Henry A. (1955). *An introduction to descriptive linguistics*. Henry Holt.

Gleason, Jean Berko & Weintraub, Sandra. (1976). The acquisition of routines in child language. *Language in Society, 5,* 129-136.

Juola, Patrick, Mikros, George, Menn, Lise, Berko Gleason, Jean, & Bernstein Ratner, Nan. (2024). Pseudo-psycholinguistic behavior of large language models. In A. Dister & D. Longrée (Eds.), *JADT 2024: Mots comptés, textes déchiffrés* (Tome 1, pp. 449–458). Presses Universitaires de Louvain.

Kessel, Frank. (1988) *The development of language and language researchers.* Psychology Press.

Lanman, Charles R. (1906). *A Sanskrit reader: with vocabulary and notes*. Ginn.

Marcus, Gary F. (2024). *Taming Silicon Valley*. MIT Press.
Pantelidou, Nikoleta, Leivada, Evelina, Morosi, Paolo. (2025). Resource-sensitive but language-blind: Community size and not grammatical complexity better predicts the accuracy of Large Language Models in a novel Wug Test. *arXiv preprint arXiv:2510.12463*.

Proceedings of the 50th annual Boston University Conference on Language Development

edited by Romi Hill, Xinhan Jiang, Danutham Worapipat, and Aditya Yedetore

Cascadilla Press Somerville, MA 2026

ISSN 1080-692X
ISBN 978-1-57473-047-0 (2 volume set, paperback)

Ordering information

To order a copy of the proceedings or to place a standing order, contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, sales@cascadilla.com, www.cascadilla.com