BUCLD 45 Proceedings
To be published in 2021 by Cascadilla Press
Rights forms signed by all authors

The Distributional Learning of Recursive Structures

Daoxin Li, Lydia Grohe, Petra Schulz, and Charles Yang*

1. Introduction

The ability to form recursive structures is almost surely innate (Berwick & Chomsky 2017, Yang 2013) but languages do differ with regard to the domain of recursion, which must be learned on the basis of language specific experience. For instance, nouns can be infinitely stacked in the English prenominal *s*-possessive (1a) but cannot do so in the postnominal *of*-possessive (1b-c). Similarly, the German postnominal possessive *von* 'of' can embed freely (2a), while the Saxon prenominal *-s* is only used with a narrow set such as proper names and some kinship terms (2b-c) and cannot be stacked (Weiß 2008, Pérez-Leroux et al. 2021). In Mandarin Chinese, noun embedding takes place freely when the possessive marker *de* is present (3a) but is generally restricted to kinship terms without *de* (3b) (Li & Thompson 1981).

(1) a. the man's neighbor's book
b. ?*the book of the neighbor
c. *the book of the neighbor of the man

(2) a. das Buch von dem Nachbarn von dem Mann
the book of the neighbor of the man
'the book of the neighbor of the man'
b. Marias/Vaters/*Manns Buch
Maria/father/*man's book
c. *Peters Nachbars Buch
Peter's neighbor's book

(3) a. na ren de linju de shu
the man GEN neighbor GEN book
'the man's neighbor's book
b. *na linju shu
the neighbor book
'the neighbor's book'

*Li & Yang, University of Pennsylvania, {daoxinli, ycharles}@upenn.edu. Grohe & Schulz, Goethe University Frankfurt, {grohe@lingua, P.Schulz@em}.uni-frankfurt.de.
We thank Lila Gleitman, Norbert Hornstein, Riny Huijbregts, Julie Legate, Joan Maling, Ana Pérez-Leroux, Yves Roberge, Tom Roeper, Rushen Shi, Jill de Villiers, Helmut Weiß, and the participants of a research seminar at the University of Pennsylvania for discussion.

c. *na ren shu
the man book
'the man's book'

In this paper, we present a proposal for how recursive structures should be formulated, which leads to a theory of how they are acquired by children. We stress from the outset that our study is not about the *capacity* for recursion, but the realization of recursive structures in specific languages and constructions, with the English, German, and Chinese possessives as a test case. Note also that we use the term *possessive* here only to refer to the formal syntactic structures; it is well known that the meanings expressed by these structures are quite varied and not limited to canonical notion of possession such as ownership (Quirk et al. 1985). Our study focuses on the formal and semantic properties of these structures available in the early stages of language acquisition: as we will see, they provide the core element for the extension of the possessives in broader usage.

Our proposal consists of two logically independent components. We first put forward a conceptualization of recursion, which reduces embedding of infinite depth to a structural property that holds in strictly one-level data. We then provide an account of how this distributional property is acquired, as a special instance of productivity in language (Yang 2005, 2016). Distributional analysis of child-directed English, German, and Mandarin Chinese shows that the major formal and semantic properties of the possessive constructions, and the condition under which they undergo recursion, are learnable from very simple linguistic data.

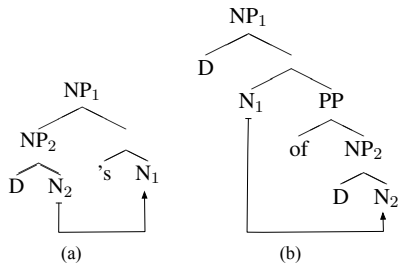
2. Proposal: Recursion as Structural Substitutability

Recursion in language has traditionally been construed as the self-embedding of a linguistic object (see Huijbregts 2019 for review). For example, *the man's neighbor* can be expressed by a rewrite rule NP→NP's NP, where the first NP self-embeds another NP to derive *the man's neighbor's book*. Similarly, the complementizer phrase (CP) recursion is expressed as CP→NP VP and VP→V CP, thereby embedding one clause inside another one. Once a rule such as NP→NP's NP is available, infinite embedding immediately follows. For the learner, the problem is how recursive rules become part of their grammar.

Clearly, the mere attestation of multilevel embedding, however deep, cannot be sufficient to support recursion. First, multi-level embedding, like all complex linguistic structures, is vanishingly rare and may not reliably appear in the input. Second, and more important, there is no principled reason why the presence of *N*-level embedding would ensure even (*N*+1)-level embedding, never mind infinite embedding. To paraphrase Sherlock Holmes, recursion of infinite depth must be learnable—however improbably—from level-one evidence.

To do so requires an alternative conceptualization of recursion. We propose that recursion does not derive from the self-embedding of a linguistic object (e.g., NP or CP) but from a property, dubbed *substitutability*, that concerns *two* positions in the formal linguistic structure.

(4)



(4) gives the syntactic representation of the English possessives. We use N_1 and N_2 to denote the two head nouns. A structure is recursive if N_1 and N_2 are *substitutable*: nouns used in one position can also be used in the other, indicated by a directional arrow ($\mapsto$). For example, the *s*-possessive (4a) is recursive if $N_2 \mapsto N_1$: nouns in N_2 can also be used in N_1, i.e., the possessor can be possessed. The *of*-possessive (4b) is recursive if $N_1 \mapsto N_2$: nouns used in N_1 can also be used in N_2, i.e., the possessum can possess. Note that the directionality of substitutability is different for the two structures, which is a distributional property that English-learning children must acquire.

Importantly, under the current conception of recursion, there cannot be embedding of finite depth (say, up to level 2 or 7): a structure is either not recursive or infinitely recursive (see also Huijbregts 2019). In principle, a structure may be recursive for only one lexical item. Imagine a hypothetical variety of English where $N_2 \mapsto N_1$ holds in the *s*-possessive only for the noun *mother*. That is, the language contains expressions such as *mother's car*, where *mother* is in N_2, and *the student's mother*, where *mother* is in N_1; no other noun appears in both positions. Nevertheless, the structure is recursive albeit only for a single item: *the mother's mother*, *the 40-year-old mother's 80-year-old mother's 120-year-old mother*, etc. are immediately available and recursion ensues. If multiple nouns enable $N_2 \mapsto N_1$, then learner will seek to form generalizations about what makes these nouns eligible for N_2 (and thus recursion-triggering). If the generalization turns out to be valid—in a sense to be made precise in Section 3—novel, and potentially infinite many, items that follow the generalization can also be used recursively. The learning of recursion, then, becomes the problem of learning the lexicon for which structural substitutability holds.

In light of this view of recursion, consider again the two English possessives. Despite the contrast in (1), we have been careful in not calling the *of*-possessive non-recursive. While it is ill-formed with respect to nouns such as *neighbor* and *book*, the *of*-possessive does allow embedding as shown in (5):

(5) The top of the tip of his hat (attested in CHILDES input)
The end of the story of the kitty (attested in CHILDES input)
The color of the cover of the book

The middle of the third inning of the deciding game
The son of the President of the Union of Retired Professors

Again, the question is to identify the properties of the nouns that enable recursion ($N_1 \mapsto N_2$) in the *of*-possessive, which are evidently different from those in the *s*-possessive.

While our proposal reduces the problem of infinite recursion to level-one structural substitutability, the learnability problem may still seem intractable. To know that $N_2 \mapsto N_1$ holds for the English *s*-possessive, for example, is to know that *all* nouns that can appear in N_2 can also appear in N_1. Of course, just the fact that an N_2 noun *can* be used in N_1 does not mean it will be in fact used as such, not least in a modest-sized child-directed input corpus. To form a recursive generalization, then, requires a leap of faith.

3. Proposal: Productivity and Generalization

The generalization problem is not unique to the acquisition of recursion but encompasses every aspect of language acquisition (and learning more generally): How does a grammar of infinite capacities arise from a finite sample of data that embodies the grammar?

The Tolerance/Sufficiency Principle (TSP; Yang 2005, 2016) is a theory of learning that enables such generalizations. The TSP states:

(6) Let a rule *R* be defined over a set of *N* items in the input. *R* is productive if and only if *e*, the number of items not supporting *R,* does not exceed $\theta_N = N/lnN$.

We use the term *rule* in the most general sense to denote any property associated with the items. It may be a familiar input-output process (e.g., add *-ed* to verbs) or a pattern that holds of the items (e.g., nouns stress the initial syllable). The TSP states that in order for *R* to generalize productively, the proportion of supporting items in the input data must exceed a precisely specified threshold.

We will not review the evidence of the TSP, which has been applied to a wide range of empirical studies. Additional support can be found with artificial language studies where the generalization threshold can be precisely manipulated. A striking example is found in Emond & Shi (2021; this volume). These authors designed two sets of stimuli, each of which consisted of 16 distinct items. In the first set, 11 out of the 16 items followed a word order pattern; in the second, 10 out of the 16 items followed the pattern. The design reflects the prediction of the TSP. For N=16 items, the critical threshold is θ_{16}=5: 10 is insufficient for generalization despite being the majority but 11 is. Indeed, 14-month-old infants generalized the pattern from the 11/16 set, but not the 10/16 set.

A critical property of the TSP is that the threshold for generalization is lower as a proportion of *N* when *N* is smaller. For instance, if a child's vocabulary has 10 lexical items, a rule defined over them can generalize if supported by at least 6 items (θ_{10}=4). However, when the vocabulary rises to 100, a rule would require

at least 79 supporting items (θ_{100}=21) for generalization. Small is better under the TSP, which may provide an account for why young children, who know relatively few words, can nevertheless accurately extract the productive rules in their language. These matters are discussed extensively elsewhere (Yang 2016, 2018): learning recursion via structural substitutability benefits from a small vocabulary in a similar way.

4. Recursive Structures in Child-directed Language

4.1. Methods

We extracted child-directed English, German, and Mandarin data from the CHILDES database (MacWhinney 2000). We start with the English possessives (Section 4.2); the results from Mandarin Chinese and German are summarized and compared with English in Section 4.3. The English input corpus contains 12.6 million words, which constitute approximately two years of input for typical American English-learning children.

We used the %mor tier annotation provided by CHILDES. Due to their structural simplicity, the English possessives were extracted via a regular expression pattern matcher, followed by manual inspection that eliminated a small number of annotation errors. To approximate a young language learner's vocabulary, we confined our analysis to the possessives where the N_1 and N_2 positions are occupied by the 50 most frequent nouns in early child English (Carlson et al. 2014; appendix). All in all, there are 1,070 '*s*-possessives and 1,158 *of*-possessives. Our search procedure returned many expressions that are typically analyzed as measure or classifier phrases (e.g., *two cups of water*, *a piece of paper*). These were not eliminated a priori: Measure phrases also need to be learned by children, and it is useful to investigate how they are distinguished from the other uses of the *of*-possessive, which have the same formal structure.

4.2. Quantitative and Semantic Analysis

As discussed in Section 2, we assume that the child attends to the nouns that enable structural substitutability in order to identify the condition on recursion. Thus, we assume that the syntactic representation in (4) is available to the learner: statements such as $N_1 \mapsto N_2$, which make reference to the structural configuration, can be formulated and evaluated. Table 1 provides a summary of the main results. It contains the counts of the N_1 and N_2 nouns in the two possessives. The 20 most frequent nouns in each position are also listed: given the Zipf-like distribution of linguistic items, these nouns constitute the bulk of the input. The Venn diagrams in Figure 1 displays the quantitative relation between the N_1 and N_2 sets, which plays a critical role in the determination of structural substitutability.

Table 1. The distributional properties of the English possessive structures (measure words marked in italics)

Construction	N (Count)	20 Most Frequent Nouns
s-possessive (N_2's N_1)	N_1 (42)	name, head, hair, nose, mouth, room, hat, house, car, bed, hand, chair, food, cup, mommy, juice, water, truck, daddy, school
	N_2 (22)	baby, daddy, boy, mommy, dog, girl, man, cat, bear, fish, truck, train, cup, name, door, day, way, hat, color, car
of-possessive (N_1 of N_2)	N_1 (24)	*piece*, top, *bit*, picture, name, *cup*, time, color, day, head, door, *box*, way, hair, thing, mouth, book, school, room, man
	N_2 (45)	cheese, cake, head, book, train, house, water, milk, box, baby, hair, car, juice, food, school, fish, hat, day, dog, man

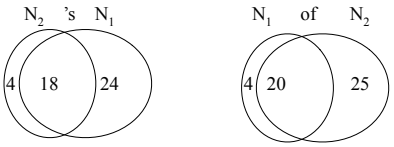


Figure 1. Set relations between the N_1 and N_2 nouns in the English possessives. The numbers indicate the cardinality of N_1, N_2, and the cardinality of their intersections

We analyze the two structures separately by mimicking the computational process that the child learner may follow. The results clearly point to the conditions under which the two possessive structures? can, and cannot, be used interchangeably, a topic of much previous research.

4.2.1 The *s*-possessive

As Figure 1 (left side) illustrates, there are 22 N_2 nouns in the *s*-possessive, of which 18 also appear in the N_1 position. The four that fail are *girl, bear, fish*, and *color*, which clearly can appear in N_1 but simply did not have the opportunity to do so in the child-direct speech corpus. Nevertheless, the TSP threshold for generalization is met (θ_{22}=7): $N_2 \mapsto N_1$ thus holds. In other words, if a noun is used in N_2, it can be used in N_1 as well, thereby enabling recursion. Note that substitutability in the other direction, i.e., $N_1 \mapsto N_2$, is untenable: 18 out of 42 isn't anywhere near sufficiency under TSP (31 is required).

Having identified the condition for recursion, the learner can now discover what makes nouns eligible for N_2 thereby triggering recursion. At least in the child-directed corpus, almost every *s*-possessive expresses the meaning of possession, which can be divided into two kinds. The first kind can be called *internal possession*: N_1 is a kin, body part, attribute, characteristic, and other inherent property of N_2, perhaps as an extension of inalienability and part-whole relations, which are often grammatically marked in the world's languages. Here we find *dog's hair, man's son, baby's name,* etc. The second kind can be referred to as *external possession*, where N_2 expresses ownership and other contingent relations with N_1. Here we find *baby's cat, daddy's bed, mommy's lunch,* etc.

Previous corpus analyses of adult speech and written corpora have found that the *s*-possessive strongly favors animate nouns in the N_2 position (e.g., O'Connor et al. 2013). Indeed, the statistical tendency is overwhelming in terms of *token* frequency: 99% of the *s*-possessives have an animate noun in the N_2 position. However, as Table 1 shows, inanimate nouns are not uncommon in terms of *types*, which is the quantity over which productivity is calculated under the TSP. In fact, 12 out of 22 N_2 nouns are inanimate, even though most appear only once. The animate nouns, despite their abundance, cannot be regarded as a sufficient condition for N_2 while relegating the inanimate nouns as lexicalized exceptions (θ_{22}=7). The semantic property of internal and external possession, however, does appear sufficient. When an inanimate noun is used in the N_2 position, it is always a case of internal possession (e.g., *car's name*) or anthropomorphic extension (e.g., *bus's house*, *train's way, flower's face, next door's cat*).

4.2.2 The *of*-possessive

In comparison to the *s*-possessive ($N_2 \mapsto N_1$), the structural substitutability in the *of*-possessive goes in the direction of $N_1 \mapsto N_2$. As illustrated in Figure 1 (right side), 24 nouns are used in N_1 of which 20 appear in N_2 as well, easily clearing the TSP threshold (θ_{24}=7). Thus, the *of*-possessive is recursive.

Again, let us consider the eligibility of nouns in N_1, the crucial condition for recursion. As shown in Table 1, 4 of the N_1 nouns are measure words as in *a piece of fish*, *a bit of cheese, two cups of juice,* and *a box of food*. The remaining Ns all express internal possession in the sense defined earlier: N_1 is a part, component, attribute, characteristic, and other inherent property of N_2. Examples include *picture of the boy, middle of the night, time of the day, day of the week, color of the flower, head of the man,* etc. Importantly 20 out of 24 guarantees productivity by the TSP (again, θ_{24}=7): internal possession can be regarded as the productive condition for N_2 nouns, with the 4 measure words lexicalized as exceptions.

Unlike the *s*-possessive, noun animacy does provide a categorical criterion for the N_2 position in *of*-possessives: 22 out of 24 nouns are inanimate, and the two animate exceptions are both nouns referring to humans, in a fixed expression: *daddy/man of the house*. This usage is also consistent with the semantic characterization of internal possession—if the child understood it as intended: the *man* or *daddy* is understood as the (senior) male member of a family as opposed to some male or father residing in the house.[1] It is possible that some kinship terms, always animate, may be used in the *of*-possessive. Although none is found

[1] The inanimacy condition is likely to remain productive even if we expand the learner's vocabulary. The entire 12.6-million-word corpus contains 630 N_1 nouns in the *of*-possessive, only 17 are animate: *man, daddy, mother, baby, friend, boss, prince, fan, director, boy, girl, principal, president, father, family, cousin,* and *author*. These are numerically well below the TSP threshold and can be lexicalized as exceptions, or learned as a group (i.e., nouns referring to humans). Notably, nouns referring to animals, of which there are many, never make an appearance in N_1.

in the input corpus, both *the son of the mother* and *the mother of the child* can be found in adult corpora. One reason may be that *mother* and *child* inherently possess each other by definition; see also footnote 1. Another, simpler, reason may be that learners eventually encounter expressions such as *mother of Dragons* and (*in*) *the name of the mother*: the attestation of *mother* in both N_1 and N_2 guarantees recursion by fiat according to our formulation.

From the child-directed input analysis, we conclude that in addition to measure phrases,[2] the relation of internal possession between N_1 and N_2 enables $N_1 \mapsto N_2$ and therefore recursion for the *of*-possessive. Moreover, N_1 must be inanimate as a rule. Note that external possession (e.g., ownership), freely available in the *s*-possessive, cannot be expressed with the *of*-possessive. This can be seen in the contrast between *the screen of the laptop* and **the monitor of the laptop* (meaning the external display): the screen is a built-in and thus inherent component of the laptop but the monitor is a detachable and independent device.

We can now account for *of*-possessive embedding in (5), repeated below:

(7) The top of the tip of his hat (attested in CHILDES input)
The end of the story of the kitty (attested in CHILDES input)
The color of the cover of the book
The middle of the third inning of the deciding game
The son of the President of the Union of Retired Professors

These examples all obey the criterion that N_1 is an internal possession of N_2. Their rarity is likely due to the fact that it is unusual for multiple nouns to form an appropriate Russian-dolls-like chain of internal possession, but it is not difficult to construct examples that obey such relations thereby extending recursion: *the hue of the color of the cover of the book, the middle of the third inning of the deciding game of the World Series*, etc.

4.2.3 Summary

The distributional analysis of the possessive structures in child-directed English produces the following conditions for recursive embedding. Only valid generalizations are listed: exceptions can nevertheless trigger recursion if attested in both N_1 and N_2 sufficiently often to be lexicalized as such.

(8) a. *s*-possessive (N_2's N_1) is recursive if N_2 and N_1 are related by internal or external possession (i.e., the possessor can embed).
b. *of*-possessive (N_1 of N_2) is recursive if inanimate N_1 and N_2 are related by internal possession (i.e., the possessum can embed).

[2] Measure words that appear in both N_1 and N_2 positions appear to allow recursion as our formulation predicts. From *the price of a pint* (N_2) and *two pints of beer* (N_1), recursion follows for *pint*, as *an order of three pints of beer* is possible.

Although the conditions in (8) are established on a very small set of child-directed data, their applicability is considerably broader. For example, (8) accounts for a well-known semantic difference between the two possessives (Chomsky 1970). The *s*-possessive *the man's picture* has two readings: a depiction of the man or an item in the man's collection, i.e., internal and external possession (8a). For the *of*-possessive *the picture of the man*, by contrast, only the internal possession reading (8b) is available.

We must note that these conditions are, and probably can only be, approximate. On the one hand, they are extracted from child-directed input data, which does not contain many of the rare and non-canonical uses of the possessives noted in the literature. On the other, and more fundamentally, these conditions depend on the language user's understanding of noun concepts and their relations, which can be flexible and fluid as the anthropomorphized examples in the input corpus illustrate (e.g., *the flower's face, the truck's home*).

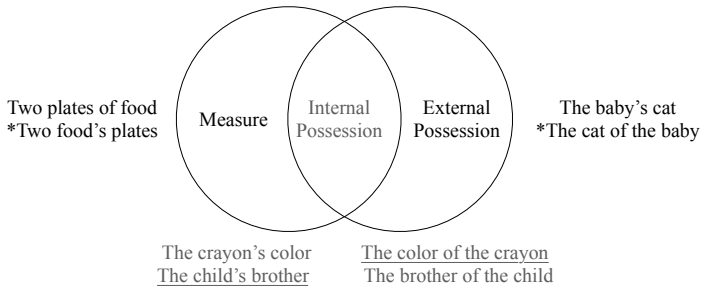
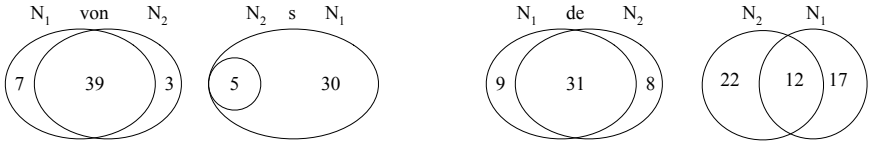


Figure 2. The semantic conditions for the English possessives. The underlined expressions are statistically preferable

Figure 2 illustrates the semantic scope of the two possessives: for internal possession, both possessive structures are possible, as illustrated by the greyscale examples. Again, there are strong statistical tendencies that favor one (underlined) over the other. Recall that N_1 position in the *of*-possessive is overwhelming inanimate and the N_2 position in the *s*-possessive is overwhelming animate. However, as discussed earlier in light of type frequencies and the TSP, the inanimacy condition in the former is a categorical one whereas the animacy condition in the latter is merely a matter of preference. Other factors contribute to the language user's preference when both options are grammatically available; see O'Connor et al. (2013) for a detailed quantitative analysis. The current study can be viewed as a prerequisite: it identifies the conditions on the possessives—and thus when they do (and do not) overlap.

4.3. Analysis of German and Mandarin Chinese

We now turn to German and Mandarin: tkey results are given in Figure 3.



(a) The German possessives (b) The Mandarin possessives

Figure 3. Set relations between N_1 and N_2 nouns in the German (a) and Mandarin (b) possessives. The numbers indicate the cardinalities of N_1 and N_2 and the cardinality of their intersection

For German we analyzed 5 CHILDES corpora (yielding a dataset of 3.5 million words of input). To determine the nouns in children's early vocabulary, we searched for the 50 most frequent nouns in the input and extracted all possessive structures containing these in N_1 or in N_2 position.[3] A total of 368 *s*-possessives and 888 *von*/*vom*-possessives were found.

Table 2. The distributional properties of the German possessive structures

Construction	N (Count)	20 Most Frequent Nouns
s-possessive (N_2's N_1)	N_1 (35)	Auto (car), Nase (nose), Bett (bed), Mama (mommy), Baby (baby), Kopf (head), Stuhl (chair), Hand (hand), Mund (mouth), Fuß (foot), Papa (daddy), Haus (house), Maus (mouse), Schuh (shoe), Eis (ice) Hund (dog), Milch (milk), Ball (ball), Geld (money), Karte (card)
	N_2 (5)	Mama (mommy), Papa (daddy), Oma (grandma), Baby (baby), Katze (cat)
von-possessive (N_1 von N_2)	N_1 (46)	Stück (piece), Bild (picture), Buch (book) Papa (daddy), Mama (mommy), Teil (part), Haus (house) Fuß (foot), Kopf (head), Kind (child), Nase (nose), Bett (bed), Auge (eye), Tür (door), Baby (baby), Auto (car), Sache (thing), Farbe (color) , Hose (pants)
	N_2 (42)	Baby (baby), Mama (mommy), Maus (mouse), Puppe (doll), Katze (cat), Kind (child), Auto (car), Hund (dog), Oma (grandma), Mann (man), Papa (daddy), Kuh (cow), Leute (people), Ente (duck), Flugzeug (plane), Hase (rabbit), Pferd (horse), Buch (book), Zug (train), Eisenbahn (railway)

In the *von*-possessive, it is evident that N_1 and N_2 are *bidirectionally* substitutable ($N_1 \leftrightarrow N_2$). There are 39 nouns used in both positions, exceeding the generalization threshold for both $N_1 \mapsto N_2$ (θ_{46}=12) and $N_2 \mapsto N_1$ (θ_{42}=11). Recall the English *of*-possessive (N_1 of N_2), where N_2 shows no apparent restriction but N_1 is categorically inanimate. Because substitutability is bidirectional in the German *von*-possessive, the learner would not even need to investigate the properties of the nouns further. For the German Saxon *s*-possessive the picture is very different: there are 5 N_2 nouns, all of which appear in the N_1 position as well. Accordingly, the child can conclude that the *s*-possessive is not generally recursive but may lexicalize the five attested nouns that do appear in both

[3] Due to smaller corpus size, we depart from the English analysis in this point. The same adjustment was applied to the analysis of the Mandarin data.

positions.[4] Both findings are in line with Pérez-Leroux et al.'s (2021) results from elicted production: when prompted to produce recursive possessives, children at age 5 provided recursive *von*-structures such as *der Ballon von dem Affen von dem Clown* (the balloon of the monkey of the clown) and not a single recursive *s*-possessive. In addition, the Saxon *s*- was used only very rarely even at level one.

For Mandarin Chinese, we analyzed all the CHILDES corpora (1.7 million words of input) and focused on the 56 nouns that were representative of three-year olds' vocabulary (Hao et al. 2008). Similar to the analysis of German, we extracted all the N_2 *de* N_1 and N_2 N_1 possessive structures that contain those words in either N_1 or N_2 position.

Table 3. The distributional properties of the Mandarin possessive structures.

Construction	N (Count)	20 Most Frequent Nouns
with *de* (N_2 de N_1)	N_1 (40)	jiao (foot), yanjing (eye), tou (head), bizi (nose), yifu (clothes) tui (leg), lian (face), shou (hand), fan (rice), shui (water) baba (dad), qian (money), mao (cat), rou (meat), chuang (bed) baobao (baby), beizi (quilt), ya (tooth), cai (vegetable), yu (fish)
	N_2 (39)	baba (dad), mao (cat), nanhai (boy), baobao (baby), didi (brother) chuang (bed), jia (family), ji (chicken), gou (dog), zhu (pig) yu (fish), gege (brother), hua (flower), ma (horse), feiji (plane) yanji (eye), cai (vegetable), yifu (clothes), rou (meat), shui (water)
without *de* (N_2 N_1)	N_1 (29)	baba (dad), didi (brother), jiao (foot), shou (hand) jia (family) yifu (clothes), gege (brother), tui (leg), yanjing (eye), tou (head) bizi (nose), lian (face), rou (meat), baobao (baby), fan (rice) zui (mouth), wazi (sock), ya (tooth), binggan (biscuit), beizi (quilt)
	N_2 (34)	mao (cat), zhu (pig), yu (fish), gou (dog), ji (chicken) cai (vegetable), yifu (clothes), beizi (quilt), baba (dad), baobao (baby) ma (horse), didi (brother), nanhai (boy), gege (brother), yazi (duck) niu (cow), tuzi (rabbit), hua (flower), feiji (plane)

The distribution of the Chinese *de*-possessive is very similar to the German *von*-possessive: bidirectional substitutability ($N_1 \leftrightarrow N_2$) can be maintained, as the 31 nouns used in both N_1 and N_2 are sufficient for $N_1 \mapsto N_2$ (θ_{40}=10) as well as $N_2 \mapsto N_1$ (θ_{39}=10). Again, the bidirectionality of substitutability means that no additional constraints on nouns are necessary. In the construction without *de*, however, only 12 nouns appear in both positions, falling far short of the TSP threshold for either direction of substitutability. Thus children will learn these nouns lexically: close kinship terms (e.g., 'father', 'brother'), body parts (e.g., 'head', 'eye'), and in an interesting parallel to the English *of*-possessive, measure words (e.g., *bei* 'cup' as in *yi bei shui* 'a glass of water'). Indeed, recursion with such limited vocabulary is possible: *baba* (dad) appears in both N_1 (*baba toufa,*

[4] Lexicalization in the absence of productivity requires extensive exposure; it is thus possible that not all speakers allow the recursive embedding of these 5 nouns. Although some kinship terms such as *Mama* 'Mom' and *Papa 'Daddy'* seem to allow recursion for some speakers, it is unclear whether the first N_2 receives a proper name reading in these cases (see also Pérez-Leroux et al. 2021).

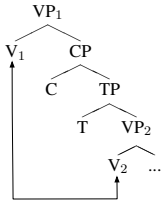
'dad's hair') and N_2 (*baobao baba,* 'baby's dad'), so *baobao baba toufa* ('baby's dad's hair') follows immediately without the use of the possessive *de*.

5. Conclusion

It is worth reiterating that our approach to recursion has two distinct components. The conception of recursion as structural substitutability (e.g., $N_1 \mapsto\!\rightarrow N_2$) transforms the problem of infinite embedding to the problem of productivity that can be resolved on level-one data. This in turn calls for a theory of learning and generalization, with the TSP on offer in our proposal but other independently motivated theories should also be evaluated.

We regard the notion of structural substitutability as crucial for all treatments of recursive structures; see Grohe et al. (2020) for an analysis of adjective embedding in English and German. Likewise, consider the familiar case of CP recursion (with irrelevant details omitted):

(9)



In our view, (9) is not to be regarded as the self-embedding of CP but as the structural substitutability of two structural positions V_1 and V_2, which are occupied by verbs that take CP as complements (e.g., *think* and *believe*, but not *think* and *eat*). In this case, substitutability is bidirectional ($V_1 \leftrightarrow V_2$): every CP-taking verb in our child-directed corpus, about a dozen in all, happens to embed each other, easily clearing the TSP threshold, and the child can conclude that all CP-taking verbs allow recursion. However, forming this generalization is unlikely to be instantaneous. In addition to learning the CP-taking verbs, the learner must also command the formal structure in (9) so that the syntactic positions for structural substitutability can be defined in the first place. To do so requires the acquisition of other components of the grammar, such as the morphosyntax of tense marking, which may follow a protracted development (Yang 2016).

We find it notable that the basic properties of the possessives can be acquired on very simple data such as our child-directed input corpora. This may not be surprising upon reflection. After all, language acquisition is remarkably early and accurate: the necessary distributional evidence must be readily available in the limited input that children receive. More interestingly, the recursion of possessives may be learnable *only* if children have a very small vocabulary of extremely frequent nouns. In order for structural substitutability to hold, the vast majority of the items in one position must also be attested in the other. Given the

sparsity of linguistic distributions, only highly frequent items will have the opportunity to appear in both structural positions. Thus, as the size of the lexicon increases, the intersection of the two sets will shrink as a proportion: the identification of productive processes becomes impossible under the TSP. Indeed, the generalizations in (8) no longer hold when all input nouns are included in the calculation: Less does seem to be more.

In summary, our approach to recursion and its acquisition embodies of a view of language acquisition that prioritizes the formal properties of grammatical structures. Recall the requirement that, as a rule, only inanimate nouns can appear in the N_1 position of the *of*-possessive. This generalization was formed after the establishment of $N_1 \mapsto N_2$, a purely formal and quantitative condition that all N_1 nouns can appear in N_2 but crucially not vice versa. No such semantic refinement is necessary for the German *von*-possessive and the Chinese *de*-possessive where N_1 and N_2 are bidirectionally substitutable. The core semantic properties of the possessives can be identified subsequently, provided that children have the requisite conceptual understanding of possession and ownership (Nancekivell et al. 2019). Indeed, a recent artificial language study demonstrates that structural substitutability and the TSP can learn recursion in a linear structure with nonce words that have no referential contents (Li & Schuler 2021). To be sure, the primacy of formal structure has been recognized since the foundation of modern linguistics (Chomsky 1955): our proposal provides a psychological procedure for the discovery of formal structures, from which their semantic content can be established. Or, as Chomsky remarks (1957, p93), "How are the syntactic devices available in a given language put to work in the actual use of this language?"

***

We end with a fable. It has many variants but this one has become familiar to linguists via Haj Ross (1967).

After a lecture, the great American psychologist William James was accosted by a little old lady who shared her theory that the world was propped up on the back of a giant turtle.

Not wishing to demolish this absurd little theory by bringing to bear the masses of scientific evidence he had at this command, James decided to gently dissuade his opponent by making her see some of the inadequacies of her position.

"If your theory is correct, Madam," James asked, "What does this turtle stand on?"

"You are a very clear man, Mr. James, and that's a very good question," replied the little old lady, "but I have an answer to it. And it's this: the first turtle stands on the back of a second, far larger, turtle, who stands directly under him."

"But what does this second turtle stand on?" persisted James patiently.

To this, the little old lady crowed triumphantly,

"It's no use, Mr. James – It's turtles all the way down."

References

Berwick, Robert & Chomsky, Noam. (2017). *Why only us.* MIT Press.

Carlson, Matthew, Sonderegger, Morgan, & Bane, Max. (2014). How children explore the phonological network in child-directed speech. *Journal of Memory and Language,* 75, 159-180.

Chomsky, Noam. (1955). *The logical structure of linguistic theory*. Unpublished manuscript, Harvard University.

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

Chomsky, Noam. (1970). Remarks on nominalization. In *Readings in English transformational grammar* (pp. 184-121). Ginn & Co.

Emond, Emeryse & Shi, Rushen. (2021). Infant's rule generalization is governed by the Tolerance Principle. (Present volume).

Grohe, Lydia, Schulz, Petra & Yang, Charles. (2020). *How to learn recursive rules: Productivity of prenominal adjective stacking in English and German.* Paper to be presented at the 9th GALANA conference.

Hao, Meiling, Shu, Hua, Xing, Ailing, & Li, Ping. (2008). Early vocabulary inventory for Mandarin Chinese. *Behavior Research Methods*, 40 (3), 728-733.

Huijbregts, M.A.C. (2019). Infinite generation of language unreachable from a stepwise approach. *Frontiers in Psychology*, 10:425. Doi:10.3389/fpsyg.2019.00425

Li, Charles & Thomson, Sandra. (1981). *Mandarin Chinese: A functional reference grammar.* Berkeley University Press.

Li, Daoxin & Schuler, Kathryn. (2021). *Acquiring recursive structures through distributional learning.* Paper presented at the 34th CUNY conference.

MacWhinney, Brian. (2000). *The CHILDES project.* Earlbaum.

Nancekivell, Shaylene, Friedman, Ori & Gelman, Susan (2019). Ownership matters: People possess a naïve theory of ownership. *Trends in cognitive sciences*, 102-113.

O'Connor, Catherine, Maling, Joan & Skarabela, Barbora. (2013). Nominal categories and the expression of possession: A cross-linguistic study of probabilistic tendencies and categorical constraints. In *Morphosyntactic categories and the expression of possession* (pp. 89-122). Benjamins.

Pérez-Leroux, Ana, Roberge, Yves, Schulz, Petra & Lowles, Alex. (2021). *Structural diversity does not affect the development of recursivity: The case of possession in German.* Ms.

Ross, John Robert. (1967). *Constraints on variables in syntax*. MIT dissertation.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey & Svartvik, Jan. (1985). *A comprehensive grammar of the English language.* Longman.

Weiß, Helmut. (2008). The possessor that appears twice? Variation, structure and function of possessive doubling in German. In *Microvariation in syntactic doubling* (pp. 381-401). Emerald.

Yang, Charles. (2005). On productivity. *Language Variation Yearbook,* 5, 265-302.

Yang, Charles. (2013). Ontogeny and phylogeny of language. *PNAS*, 110, 6324-6327.

Yang, Charles. (2016). *The price of linguistic productivity.* MIT Press.

Yang, Charles. (2018). A formalist perspective on language acquisition (Target article with commentary and authors' reply). *Linguistic Approaches to Bilingualism,* 8, 665-706.