Neuroscience and Biobehavioral Reviews 81 (2017) 103–119

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review article

The growth of language: Universal Grammar, experience, and principles of computation

Charles Yang[a,*], Stephen Crain[b], Robert C. Berwick[c], Noam Chomsky[d], Johan J. Bolhuis[e,f]

[a] Department of Linguistics and Department of Computer and Information Science, University of Pennsylvania, 619 Williams Hall, Philadelphia, PA 19081, USA
[b] Department of Linguistics, Macquarie University, and ARC Centre of Excellence in Cognition and its Disorders, Sydney, Australia
[c] Department of Electrical Engineering and Computer Science and Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA
[d] Department of Linguistics and Philosophy, MIT, Cambridge MA, USA
[e] Cognitive Neurobiology and Helmholtz Institute, Departments of Psychology and Biology, Utrecht University, Utrecht, The Netherlands
[f] Department of Zoology and St. Catharine's College, University of Cambridge, UK

ARTICLE INFO

Article history:
Received 13 September 2016
Received in revised form
10 November 2016
Accepted 16 December 2016
Available online 7 January 2017

Keywords:
Language acquisition
Generative grammar
Computational linguistics
Speech perception
Inductive inference
Evolution of language

ABSTRACT

Human infants develop language remarkably rapidly and without overt instruction. We argue that the distinctive ontogenesis of child language arises from the interplay of three factors: domain-specific principles of language (Universal Grammar), external experience, and properties of non-linguistic domains of cognition including general learning mechanisms and principles of efficient computation. We review developmental evidence that children make use of hierarchically composed structures ('Merge') from the earliest stages and at all levels of linguistic organization. At the same time, longitudinal trajectories of development show sensitivity to the quantity of specific patterns in the input, which suggests the use of probabilistic processes as well as inductive learning mechanisms that are suitable for the psychological constraints on language acquisition. By considering the place of language in human biology and evolution, we propose an approach that integrates principles from Universal Grammar and constraints from other domains of cognition. We outline some initial results of this approach as well as challenges for future research.

Contents

* Corresponding author.
E-mail address: charles.yang@ling.upenn.edu (C. Yang).

http://dx.doi.org/10.1016/j.neubiorev.2016.12.023
0149-7634/

1. Internal and external factors in language acquisition

Where does language come from? Legend has it that the Egyptian Pharaoh Psamtik I ordered the ultimate experiment to be conducted on two newborn children. The story was first recorded in Herodotus's *Histories*. At the Pharaoh's instruction, the story goes, the children were raised in isolation and thus devoid of any linguistic input. Evidently pursuing a recapitulationist line, the Pharaoh believed that child language would reveal the historical root of all human languages. The first word the newborns produced was *bekos*, meaning "bread" in the now extinct language, Phrygian. This identified Phrygian, to the Pharaoh's satisfaction, as the original tongue.

Needless to say, this story must be taken with a large pinch of salt. For starters, we now know that consonants such as /s/ in *bekos* are very rare in early speech: the motor control for its articulation requires forcing airflow through partial opening of the vocal tract (hence the hissing sound), and this takes a long time for children to master and is often replaced by other consonants (Locke, 1995; Vihman, 2013). Furthermore, the very first words children produce tend to follow a consonant-vowel (CV) template: even if /s/ were properly articulated, it would almost surely be followed by an epenthetic vowel to maintain the integrity of a CV alternation. But the Pharaoh did get two important matters right. First, he was correct to assume that language is a biological necessity—in Darwin's words, "an instinctive tendency to acquire an art". He was also correct in supposing that, Phrygian or otherwise, a language can emerge simply by putting children together, even in the absence of an external model. Indeed, we now know that sign languages are often spontaneous creations by deaf communities, as recently documented in Nicaragua and in Israel (Kegl et al., 1999; Sandler et al., 2005; Goldin-Meadow and Yang this issue). Second, the Pharaoh was correct in supposing that child language development can reveal a great deal about the nature of language, its place in the mind, and how it emerged as a biological capacity that is unique to our species.

The nature of language acquisition has always been a central concern in the modern theory of linguistics known as generative grammar (Chomsky, 1957; this issue). The notion of a Language Acquisition Device (Chomsky, 1965), a dedicated module of the mind, still features prominently in the general literature on language and cognitive development. In this article, we highlight some major empirical findings that, together with additional case studies, form the foundation of future research in language acquisition. At the same time, we present some leading ideas from theoretical investigations of language acquisition, including promising themes developed in recent years. For several of the additional case studies, see Crain et al. (this issue).

Like the Crain et al. paper, our perspective on the acquisition of language is shaped by the biolinguistic approach (Berwick and Chomsky, 2011). Language is fundamentally a biological system and should be studied using the methodologies of the natural sciences. Although we are still quite distant from fully understanding the biological basis of language, we contend that the development of language, like any biological system, is shaped both by language-particular experience from the external environment and by internal constraints that hold across all linguistic structures. Our discussion is further shaped by considerations from the origin and evolution of language (see Berwick and Chomsky, 2016). Given the extremely brief history of *Homo sapiens* and the subsequent emergence of a species-specific computational capacity for language, the evolution of language must have been built on a foundation of other cognitive and perceptual systems that are shared across species and across cognitive domains. More specifically, the biolinguistic approach is shaped by three factors. These three factors are integral to the design of language and therefore crucial for the acquisition of language (Chomsky, 2005):

a Universal Grammar: The initial state of language development is determined by our genetic endowment, which appears to be nearly uniform for the species. At the initial state, infants interpret parts of the environment as linguistic experience; this is a nontrivial task which infants carry out reflexively and which regulates the growth of the language faculty.
b Experience: This is the source of language variation, within a fairly narrow range, as in the case of other subsystems of human cognition and the formation of the organism more generally.
c Third factors: These factors include principles that are not specific to the language faculty, including principles of hypothesis formation and data analysis, which are used both in language acquisition and in the acquisition of knowledge in other cognitive domains. Among the third factors are also principles of efficient computation as well as external constraints that regulate development in all biological organisms.

In the following Section we discuss several general principles of language and the remarkably early emergence of these principles in children's biological development. Section 3 focuses on the role of language-specific experience and how the latitude in children's experience shapes language variation during the course of acquisition. Section 4 reviews some recent effort to reconceptualize the problem of language acquisition, focusing specifically on how domain-general learning mechanisms and the principles of efficient computation figure into the Language Acquisition Device.

2. The emergence of linguistic structures

As observed long ago by Wilhelm von Humboldt, human language makes "infinite use of finite means". It is clear that to acquire a language is to implement a combinatorial system of rules that generates an unbounded range of meaningful expressions, relying on the biological endowment that determines the range of possible systems and a specific linguistic environment to select among them. von Humboldt's adage underscores a central feature of language; namely, linguistic representations are always hierarchically organized structures, at all levels of the linguistic system (Everaert et al., 2015). These structural representations are formed by the combinatorial operation known as 'Merge.'

2.1. Hierarchy and merge

The hierarchical structure of language was recognized at the very beginning of generative grammar. The complexity of human language lies beyond the descriptive power of finite state Markov processes (Chomsky, 1956), as exemplified in sentences that involve recursive nonlocal dependencies:

(1) a. **If** . . . **Then** . . .
b. **Either** . . . **Or** . . .
c. The **child** who ate the cookies . . . **is** no longer hungry.

The structural relations between the highlighted words can be embedded and so encompass arbitrary numbers of such pairings. In English, the Noun Phrase (NP) and Verb Phrase (VP) mark person/number agreement, and the subject NP may itself contain an arbitrarily long relative clause ("who ate the cookies . . ."), which may contain additional embeddings. Phrase structure rules such as (2) are needed, where the NP and VP in (2a) must show agreement and the recursive application of (2b) produces embedding:

(2) a. $S \rightarrow NP\ VP$
b. $NP \rightarrow NP\ S$

Phrase structure rules also describe the semantic relations among syntactic units. For example, the sentence "they are flying

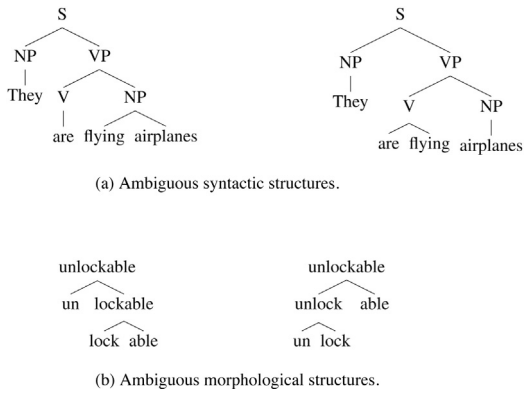


Fig. 1. The composition of phrases and words by Merge, which can lead to structural and corresponding semantic ambiguities.

airplanes" is ambiguous: "flying airplanes" can be understood as a VP, in which "airplanes" is the object of "flying", or the same phrase can be understood as an NP, which is the part of the predicate whose subject is "they" (see Fig. 1a) (Everaert et al., 2015).

Even the most cursory look at these features of English sentences make it evident that human language is set apart from other known systems of animal communication. The best studied case of (nonhuman) animal communication is the structure and learning of birdsongs (Marler, 1991; Doupe and Kuhl, 1999). Birdsongs can be characterized as syllables (elements) arranged in a sequence preceded and followed by silence (Bolhuis and Everaert, 2013; see Prather et al., this issue; Beckers et al., this issue). Despite misleading analogies to human language occasionally found in the literature (e.g., Abe and Watanabe, 2011; Beckers et al., 2012; this issue), birdsongs can be adequately described by finite state networks (Gentner and Hulse, 1998) where syllables are linearly concatenated, with transitions adequately described by probabilistic Markovian processes (Berwick et al., 2011a,b). In fact, the computational analysis of large birdsong corpora suggests that they may be characterized by an even more restrictive type of device known as *k*-reversible finite-state automata (Angluin, 1982; Berwick and Pilato, 1987; Hosino and Okanoya, 2000; Berwick et al., 2011a,b; Beckers et al., this issue). Interestingly, such automata are provably efficient to learn on the basis of positive examples (see Section 3.3 for the nature of positive and negative examples). If so, the formal tools developed for the analysis of human language can also be effectively used to characterize animal communication systems (Schlenker et al., 2016).

The central goal of generative grammar is to understand the compositional process that yields hierarchical structures in human language, the use of this process by language-users in production and comprehension, the acquisition of this process, and its neural implementation in the human brain (Berwick et al., 2013; Friederici et al., in press). In recent years, the operation *Merge* (Chomsky, 1995) has been hypothesized to be responsible for the formation of linguistic structures—the Basic Property of language (Berwick and Chomsky, 2016). Merge is a recursive process that combines two linguistic *terms*, X and Y, to yield a composite term (X, Y) which may itself be combined with another linguistic term Z to form ((X, Y), Z), automatically producing hierarchical structures. For example, the English verb phrase "read the book" is formed by the recursive application of Merge: *the* and *book* are Merged to produce (*the, book*), which is then Merged with *read* to form (*read,* (*the, book*)). The primary function of a syntactic term formed by Merge is determined by one of its terms, traditionally referred to as the *head*—hence (*the*, *book*) is a Noun Phrase and (*read*, (*the*, *book*)) is a Verb Phrase, as shown in (2). Current research has focused on how the headedness of syntactic objects is determined by general principles of efficient computation in the construction of syntactic structures (e.g., Chomsky, 2013).

There is broad structural and developmental evidence that Merge is the fundamental operation of structure building in human language, though it is most often associated with the formation of syntactic structures (Everaert et al., 2015). However, word formation rules (morphology) merge elementary informational units in a stepwise process that is similar to syntax. For instance, the word "unlockable" is doubly ambiguous when describing a lock—it can refer to either a functional lock or to a broken one. Such dualities of meaning can be captured by differences in the combination of sequences of morphemes. The meaning associated with a functional lock is derived by first combining "un" and "lock", followed by "able" — meaning that it is possible to unlock the object. A different derivation is used to refer to broken locks. This meaning is derived by first combining "lock" and "able", followed by the negative marker "un". Such ambiguities in word formation can be represented using a notation similar to arithmetic expressions, where brackets indicate precedence ([un-[lock-able]] vs. [[un-lock]-able]), or by using a tree-like format analogous to syntactically ambiguous structures (Fig. 1b).

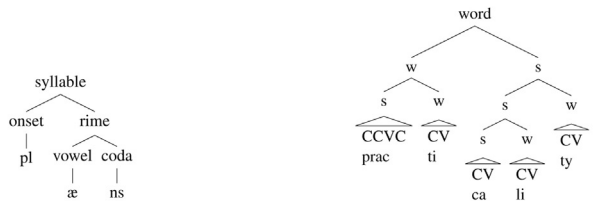
The same holds for phonology. One primary unit of phonological structure is the syllable, which is formed by combining phonemes into structures comprised of (ONSET, (VOWEL, CODA)), as illustrated in Fig. 2a.

The VOWEL is merged with the CODA to form the RIME, which is then merged with the ONSET to form the syllable (Kahn, 1976). The structure of the syllable is universal, whereas the choices of ONSET, VOWEL, and CODA are language-specific and therefore need to be learned. For instance, while neither *clight* nor *zwight* is an actual word of English, the former is a potential English word as the *cl* (/kl/) is a valid onset of English whereas *zw* is not, but is a possible onset in Dutch. Furthermore, the phonological properties of words, such as stress, are determined by assigning different levels of prominence to the hierarchical structures that have been iteratively constructed from the syllables (Liberman, 1975; Halle and Vergnaud, 1987). Syllables are merged to form *feet,* in which one of the syllables is strong (i.e., most prominent), reflecting language-specific choice; see Fig. 2b for the representation of a five-syllable word. The prosodic structures of syntactic phrases have a similar derivation (Chomsky and Halle, 1968): for example, "cars" in the noun phrase "red cars" receives primary prominence in natural intonation (see Everaert et al., 2015, for detailed discussion).

2.2. Merge in early child language

The evidence for Merge can be observed from the very beginning of language acquisition. Newborn infants have been found to be sensitive to the centrality of the vowel in the hierarchical organization of the syllable. After habituation to bisyllablic nonsense words such as "baku", which has four phonemes, babies show no surprise when they encounter new stimuli that are also bisyllabic, even if the positions of constants and vowels and the total number of phonemes are changed (e.g., "alprim", "abstro"). By contrast, infants show a novelty effect when trisyllabic words are introduced following habituation on sequences of bisyllabic words (Bijeljac-Babic et al., 1993).

The compositional nature of Merge is on display as soon as infants start to vocalize. At five to six months, infants spontaneously start to babble, producing rhythmic repetitions of nonsense syllables. Despite the prevalence of sounds such as "mama" and "dada", the combination of consonants and vowels in babbling have no referential meaning. Only a restricted set of consonants and vowels are produced at these early stages, reflecting infants' articulatory limitations, although the combinations generally follow the CV template. By seven to eight month, babbling begins to



(a) Syllable structure of the word *plans*.

(b) The structural representation of stress where feet are constructed from syllables iteratively following language-specific instructions. The symbols *s* and *w* denote the relative prominence of the units that are merge, also reflecting language-specific choices.

Fig. 2. The composition of phonological structures.

exhibit language-specific features including both phonemic inventory and syllable structure (de Boysson-Bardies and Vihman, 1991). Remarkably, deaf infants babble manually, producing gestures that follow the rhythmic pattern of the sign language they are exposed to, which are also devoid of referential meaning, as in vocal babbling (Petitto and Marentette, 1991). These findings suggest that babbling is an essential, and irrepressible, feature of language: despite the absence of semantic content, babbling merges linguistic units (phonemes and syllables) to create combinatorial structures. In marked contrast, other species such as parrots have an impressive ability to mimic human speech (Bolhuis and Everaert, 2013) but there is no evidence that they ever decompose speech into discrete units. The best that Alex, the famous parrot, could offer was a rendition of the word "spool" as "swool", which he evidently picked up when another parrot was taught to label the object (Pepperberg, 2007). This is clearly not the same as the babbling of human infants. The word "swool" is an isolated example rather than the result of free creation. It is most definitely referential in meaning and can only be regarded as an imitation error.

The use of combinatorial structures can be also be observed in infants' development of speech perception. Interestingly, combinatorial structures are revealed in the *decline* of certain aspects of speech perception in infancy. It has long been known that newborns are capable of perceiving nearly all of consonantal contrasts that are manifested across the world's languages (Eimas et al., 1971). Immersion in the linguistic environment eventually leads to the loss of infants' ability to discriminate non-native contrasts at around ten months (Werker and Tees, 1984); only native contrasts are retained after that. The change and reorganization of speech perception are again driven by the combinatorial use of language. Phonetic categories become phonemic if they represent distinctive units of the phonological system- minimal units that distinguish contrasting words. For example, Korean speakers acquiring English as a second language often have difficulty recognizing and producing the contrast between the phonemes /r/ and /l/. While Korean does distinguish the *phonetic* categories [r] and [l]— as in *Ko**r**ea* and *Seou**l***—they are not contrastive, since [r] always appears in the onset of syllables whereas [l] always appears in the coda. In English, the phonetic categories [r] and [l] form a phonemic contrast because they are used to distinguish minimal pairs of words such as *right* ∼ *light, fly-fry, mart-malt*, and *fill-fur*—which are, again, the result of the combinatorial Merge of elementary units. (The notation [] marks phones, and // marks phonemes: thus [r] ∼ [l] in Korean but /r/ ∼ /l/ in English Borden et al., 1983).

The acquisition of native contrasts, sometimes at the expense of the non-native ones, appears to be enabled by contrastive analysis of word meanings, in a process similar to the way that linguists identify the phonemic system of a new language in field work. Two lines of evidence directly support this proposal. First, recent findings suggest that infants start to grasp elements of word meanings before six month of age (Bergelson and Swingley, 2012), which makes the contrastive analysis of words possible. That is, as long as the infant knows that *big* and *pig* have different meanings—even without knowing precisely what these meanings are—they will be able to discover that /b/ and /p/ are phonemes that are used contrastively in English. Second, a minimal-pair based strategy for phonemic learning has been successfully induced in the experimental setting using non-native phonemes. In a study by Yeung and Werker (2009) nine-month-old infants were familiarized with visual objects whose labels were consistently distinguished by non-native phonemic contrasts. After only a brief period of training, infants learned the contrast, which in effect recreated the experience of minimal pairs in phonemic acquisition.

It is worth noting further that children's development of speech perception can be usefully compared with the perceptual abilities of other species. Animals such as chinchillas and pigeons can be trained to discriminate speech categories (Kuhl and Miller, 1975). However, there is no evidence that exposure to one language leads to the *loss* of perceptual ability for categories in other languages. Although we are not aware of any direct studies, it would be highly surprising to find that a monkey raised in Korea would lose the ability to distinguish /r/ from /l/, which would be similar to the developmental changes observed in infants acquiring Korean. The ability to distinguish acoustic differences in speech categories appears to be the stopping point for non-human species; only human infants go on to develop the combinatorial use of these categories. This reveals the influence of the Basic Property of language, Merge, which is responsible for the combinatorial, and contrastive, native-language phoneme language and thus the decline of perceptual abilities for non-native-language units.

In general, the evidence for Merge in syntactic development is only directly observable in conjunction with the acquisition of specific languages: the combinatorial process requires a set of basic linguistic units such as words, morphemes, etc., which are language specific. Note that observing the effects of Merge in syntactic development does not necessarily require speech production: perceptual experiments can reveal syntactic knowledge, sometimes well before children can string together long sequences of words into sentences. For instance, cross-linguistic variation in word order across languages can be described by a *head-directionality* parameter (see Section 3.2), which specifies a language's dominant ordering between the head of a phrase and the complement it is merged with. We noted earlier that the head of a phrase determines its syntactic function. English is a head-initial language. In most English phrases, the head precedes the complement so, for example, the verb "read" precedes the complement "books" in the phrase "read books", and the preposition "on" precedes the complement "deck" in the prepositional phrase "on deck". In contrast, Japanese is a head-final language. In Japanese, the order of the head and complement within the verb phrase is the mirror image of English, and Japanese has post-positional phrases rather than pre-positional phrases. It has been noted (Nespor and Vogel, 1986) that within a phonological phrase, prominence systematically falls on the right in head-initial languages (English, French, Greek, etc.), whereas it

falls on the left in head-final languages (Japanese, Turkish, Bengali, etc.). Perceptual studies (Christophe et al., 2003) have shown that 6–12-week-old infants, well before they have acquired any words, can discriminate between languages that differ in head direction and its prosodic correlate, even in languages that are otherwise similar in phonological properties (e.g., French and Turkish). Thus, very young infants may recognize the combinatorial nature of syntax and make use of its prosodic correlates to generate inferences about language-specific structures.

The productive use of Merge can be detected as soon as children start to create multiword combinations, which begins around the age of two for most children. This is not to say that all aspects of grammar are perfectly acquired, a point we discuss further, in Section 3 for the acquisition of language-specific properties and in Section 4 where we clarify the recursive nature of Merge. Traditionally, evidence for the successful acquisition of an abstract and productive syntactic system comes from the scarcity of errors, especially word order errors, in children's production (Valian, 1986).

A recent proposal, the *usage*-based approach (Tomasello, 2000a,b, Tomasello, 2003), denies the existence of systematic rules in early child language but emphasizes the memorization of specific strings of words; the paucity of errors would be attributed to memorization and retrieval of error-free adult input. For example, English singular nouns can interchangeably follow the singular determiners "a" and "the" (e.g., "a/the car," "a/the story"). In children's speech production, the percentage of singular nouns paired with both determiners is quite low: only 20–40%, and the rest appear with one determiner exclusively (Pine and Lieven, 1997). Low combinatorial diversity measures have been invoked to suggest that children's early language does not have the full productivity of Merge.

However, the usage-based approach fails to provide rigorous statistical support for its assessment of children's linguistic ability. First, quantitative analysis of adult language such as the Brown Corpus of print English (Francis and Kucera, 1967) and infant-directed speech reveals comparable, and comparably low, combinatorial diversity as children (Valian et al., 2009)—but adults' grammatical ability is not in question. Second, since every corpus is finite, not all possible combinations will necessarily be attested. A statistical test (Yang, 2013) was devised using Zipf's law (see Section 3.1) to approximate word probabilities and their combinations. The test was used to develop a benchmark of combinatorial diversity, assuming that the underlying grammar generates fully productive and interchangeable combinations of determiners and nouns. As shown in Fig. 3a, although the combinatorial diversity is low in children's productions, it is statistically indistinguishable from the expected diversity under a rule where the determiner-noun combinations are fully productive and interchangeable—on a par with the Brown Corpus. The test also provides rigorous supporting evidence that Nim, the chimpanzee raised in an American Sign Language environment, never mastered the productive combination of signs (Terrace et al., 1979; Terrace, 1987): Nim's combinatorial diversity falls far below the level expected of productive usage (Fig. 3b). In recent work, the same statistical test has been applied to home signs. Home signs are gestural systems created by deaf children with properties akin to grammatical categories, morphology, sentence structures, and semantic relations found in spoken and sign languages (Goldin-Meadow, 2005). Quantitative analysis of predicate-argument constructions suggests that, despite the absence of an input model, home signs show full combinatorial productivity (Goldin-Meadow and Yang, this issue). Taken together, these statistically rigorous analyses of language, including early child language, reinforce the conclusion that a combinatorial linguistic system supported by the operation Merge is likely an inherent component of children's biological predisposition for language.

2.3. Structure and interpretation

An important strand of evidence for the hierarchical nature of language, attributed to Merge, can be found in children's semantic interpretation of syntactic structures, much like how the semantic ambiguities of words and sentences are derived from syntactic ambiguities (Fig. 1). There is now an abundance of studies suggesting the primacy of the hierarchical organization of language (Everaert et al., 2015). Here we refer the reader to the article in this issue by Crain et al., who review several case studies in the development of syntax and semantics, including the general principle of Structure Dependence, a special case of which is much-studied and much misunderstood problem of auxiliary inversion in English question formation (Chomsky, 1975; Crain and Nakayama, 1987; Legate and Yang, 2002; Perfors et al., 2011; Berwick et al., 2011a,b).

Our remarks start with a simple study dating back to the 1980s. The correct meaning of the simple phrase "the second green ball" is reflected in its underlying syntactic structure (Fig. 4a). The adjective "green" first Merges with "ball", yielding a meaning that refers to a set of balls that are green. This structure is then Merged with "second" that picks up the second member of the previously formed set (this is circled with solid border in Fig. 4c). However, if one were to interpret "second green ball" as a linear string rather than as a hierarchically structure (Fig. 4b) (as in Dependency Grammar, a popular formalism in natural language processing applications), then the meaning of "second" and "green" may be interpreted conjunctively. On this meaning, the phrase would pick out the ball that is in the second position as well as green (this is circled with the dotted border in Fig. 4c). However, young children can correctly identify the target object of the "second green ball"—the solid, not, dotted border—producing only 14% non-adult-like errors (Hamburger and Crain, 1984).

The contrast between linear and hierarchical organization was also witnessed in studies of children's interpretation of sentences in which pronouns appeared in one of several different structural positions; see (Crain et al., this issue) for a wide range of related linguistic examples and empirical findings from experimental studies of child language. Consider the experimental setup of Crain and McKee (1985) and Kazanina and Phillips (2001), where Winnie-the-Pooh ate an apple and read a book—and the gloomy Eeyore ate an banana instead of an apple. Children were divided into four groups, and each group heard a puppet describing the situation using one of the sentences in (3).

(3) a. While Winnie-the-Pooh was reading the book, he ate an apple.
b. While he was reading the book, Winnie-the-Pooh ate an apple.
c. Winnie-the-Pooh ate an apple while he was reading the book.
d. He ate an apple while Winnie-the-Pooh was reading the book.

The child participants' task was to judge if the puppet got the story right. This experimental paradigm, known as the Truth Value Judgment Task (Crain and McKee, 1985; Crain and Thornton, 1998), only requires the child to answer Yes or No, thereby sidestepping performance difficulties that may limit young children's speech production.

If the interpretation of the pronoun *he* is Winnie-the-Pooh in all four versions of the test sentences, then clearly all four descriptions of the situation were factually correct. However, as every English speaker can readily confirm, *he* cannot refer to Winnie-the-Pooh in sentence (3d). In the experimental context, the pronoun *he* could only refer to Eeyore, who did not eat the apple. Therefore, if children represent the sentences in (3) using the same hierarchical structures as adults, then children should say that the puppet got it wrong when he said (3d), but the puppet got it right when he used the other test sentences. Even children younger than three

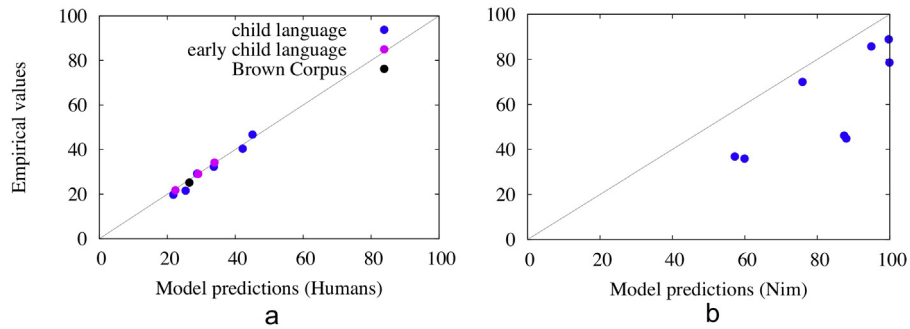


Fig. 3. Syntactic diversity in human language (a) and Nim (b). The human data consists of determiner-noun combinations from 9 corpora, including 3 at the very beginning of multiple-word utterances. Nim's sign data is based on Terrace et al. (1979). The diagonal line denotes identity between empirical and predicted values of diversity. Adapted from Yang (2013).

produced adult-like responses: they overwhelmingly rejected the description in (3d) but readily accepted the other three descriptions (3a-c). Note that the linear order between the pronoun and the NP is not at issue: in both (3b) and (3d), the pronoun *he* precedes the name *Winnie-the-Pooh*, but coreference is unproblematic in (3b), whereas it is impossible in (3d). The underlying constraint of interpretation has been studied extensively, and appears to be a linguistic universal, with acquisition studies similar to (3) carried out with children in many languages. Briefly, a constraint of coreference (Chomsky, 1981) states that a referential expression such as *Winnie-the-Pooh* cannot have the same referent as another expression that "c-commands" it. The notion of c-command is purely a formal relation defined in terms of syntactic hierarchy:

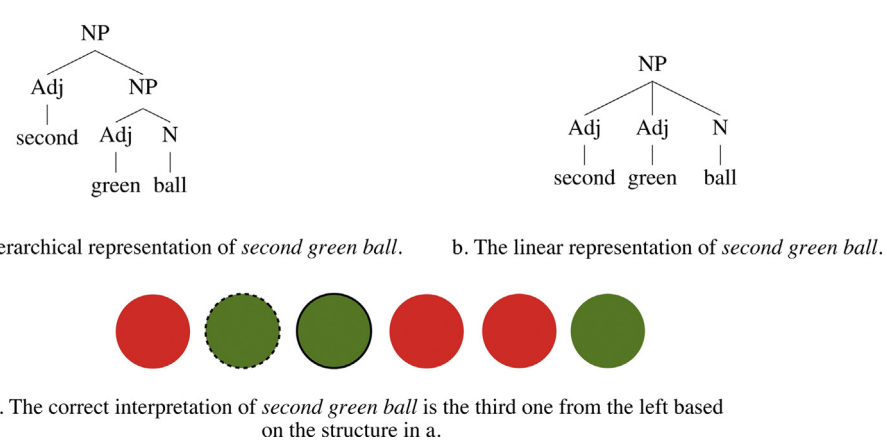
(4) X c-commands Y if and only if
 a. Y is contained in Z,
 b. X and Z are terms of Merge.

As schematically illustrated in Fig. 5, the offending structure in (3d) has the pronoun *he* in a c-commanding relationship with the name *Winnie-the-Pooh*, making coreference impossible. By contrast, the pronoun *he* in (3b) is contained in the "When . . ." clause, which in turn modifies the main clause "Winnie-the-Pooh ate an apple": there is no c-command relation, so coreference is allowed in (3b).

The hierarchical and compositional structure of language, encapsulated in Merge and commanded by children at a very early age, suggests that it is an essential feature of Universal Grammar, our biological capacity for language (Crain et al., this issue). But we also stress that the study of generative grammar and language acquisition, like all sciences, is constantly evolving: theories are refined which in turn offer new ways of looking at data. For example, not very long ago, languages such as Japanese and German, due to their relative flexibility of word order, were believed to have flat syntactic structures. However, advances in syntactic theories, and new diagnostic tests developed by these theories, convincingly revealed that these languages generate hierarchical structure (see Whitman, 1987 for a historical review). Even Australian languages such as Warlpiri, once considered an outlier as it appears to allow any permutation of word order, have been shown to adhere to the principle of Structure Dependence and the constraint on coreference just reviewed (Legate, 2002). The principles posited in the past and current linguistic research have considerable explanatory power as they have been abundantly supported in structural, typological, and developmental studies of language. However, as we discuss in Section 4, they may follow from far deeper and more unifying principles, some of which may be domain general and not unique to language, once the evolutionary constraints on language are taken into consideration.

3. Experience, induction, and language development

It is obvious that language acquisition requires experience. While the computational analysis of linguistic data, including probabilistic and information-theoretical methods, was recognized as an important component of linguistic theory from the very beginning of generative grammar (Chomsky, 1955, 1957; Miller and Chomsky, 1963), it must be acknowledged that the generative study of language acquisition has not paid sufficient attention to the role of the input until relatively recently. Part of the reason is empirical.



a. The hierarchical representation of *second green ball*.

b. The linear representation of *second green ball*.

c. The correct interpretation of *second green ball* is the third one from the left based on the structure in a.

Fig. 4. Semantic interpretation is based on hierarchical structures.

a. *Winnie-the-pooh* is not c-commanded by *he* and coreference is possible. b. *Winnie-the-pooh* is c-commanded by *he* and coreference is prohibited.

Fig. 5. Constraints on coreference are hierarchical not linear.

As we have seen, much of children's linguistic knowledge appears fully intact when it is assessed using appropriate experimental techniques, even at the earliest stages of language development. This is not surprising if much of child language reflects inherent and general principles that hold across all languages and require no experience-dependent learning. Even when it comes to language-specific properties, children's command has been generally found to be excellent. For example, the first comprehensive, and still highly influential, survey of child language by Roger Brown (1973) concludes that errors in child language are "triflingly few" (p. 156). These findings leave the impression that the contribution of experience, while clearly undeniable, is nevertheless minimal. Furthermore, starting from the 1970s, the mode of adult-child interactions in language acquisition became better understood (see Section 3.3). Contrary to still popular belief, controlled experiments have shown that "motherese" or "baby talk", the special register of speech directed at young children (Fernald, 1985), has no significant effect on the progression of syntactic development (Newport et al., 1977). Furthermore, studies of language/dialect variation and change show convincingly that despite the obvious differences in learning styles, level of education, and socio-economic status, all members in a linguistic community show remarkable uniformity in the structural aspects of language—which is deemed an "enigma" in sociolinguistics (Labov, 2010). Finally, to seriously assess the role of input evidence requires relatively large corpora of child-directed speech data, which have only become widely available in the past two decades following technological developments.

In this section, we reconsider the role of experience in language acquisition. After all, even if language acquisition were in fact instantaneous, it would still be important to determine the extent to which experience so quickly and accurately is used by child language learners in figuring out the specific properties of the local language. We first review the statistical properties of language. These statistical properties highlight the challenges the child faces during language acquisition. We then summarize some quantitative and cross-linguistic findings in the longitudinal development of language, with special reference to the theory of Principles and Parameters (Chomsky, 1981). We show that children do not simply match the expressions in the input. Instead, children are found to spontaneously produce linguistic structures that could be analyzed as errors with respect to the target language.

3.1. Sparsity and the necessity of generalization

Sparsity is the most remarkable statistical property of language: When we speak or write, we use a relatively small number of words very frequently, while the majority of words are hardly used at all. In a one-million-word Brown Corpus (Kŭcera and Francis, 1967), the top five words alone (*the, of, and, to, a*) account for 20% of usage, and 45% of word types appear exactly once. More precisely, the frequency of word usage conforms to what has become known as Zipf's Law (1949): the rank order of words and their frequency of use are inversely related, with the product approximating a constant. Zipf's Law has been empirically verified in numerous languages and genres (Baroni, 2009), although its underlying cause is by no means clear, or even informative about the nature of language: as pointed out long ago, random generating processes that combine letters also yield Zipf-like distributions (Mandelbrot, 1953; Miller, 1957; Chomsky, 1958).

Language, of course, is not just about words; it is also about rules that combine words and other elements to form meaningful expressions. Here the long tail of word usage that follows from Zipf's Law grows even longer: the frequencies of combinatorially formed expressions drop off even more precipitously than single words do. For example, as compared to the 45% of words that appear exactly once in the Brown Corpus, 80% of word pairs appear once in the corpus, and a full 95% of word triplets appear once (Fig. 6).

Past research has shown that, as new data comes in, so do new sequences of words (Jelinek, 1998). Indeed, the sparsity of language can be observed at every level of linguistic organization. In modestly complex morphological systems (e.g., Spanish), even the most frequent verb stems are paired with only a portion of the inflections that are licensed in the grammar. It follows that language learners never witness the whole conjugation table—helpfully provided in textbooks—fully fleshed out, for even a single verb. To acquire a language with a limited amount of data, estimated at on average 20–30 million words (Hart and Risley, 1995). The main challenge is to form accurate generalizations. Of the innumerable combinations of words that the learner will not observe, some will be grammatical while others are not, as illustrated in the pair "John speaks frequently French" versus "John frequently speaks French": How should a learning system tease apart the grammatical from the ungrammatical? We return to some of the relevant issues in Section 3.3.

The sparsity of language speaks against the usage-based approach to language acquisition, in view of its emphasis on the role of memorization in development. Granted, the brain has an impressive storage capacity, such that even minute details of linguistic signals can be retained. At the same time, the limitation of the usage-based approach is highlighted by the absence of a viable and realistic storage-based model of natural language processing despite the unimaginable amount of data storage and data mining capabilities. Indeed, research in natural language processing affords support for the indispensability of combinatorial rules in language acquisition. Statistical models of language that have been designed for engineering purposes can provide a useful way to assess how different conceptions of linguistic structures contribute to broader coverage of linguistic phenomena. For instance, a statistical model of grammar such as a probabilistic parser (Charniak and Johnson, 2005; Collins, 1999) can encode several types of grammatical rules: a phrase "drink water" may be represented in multiple forms ranging from categorical ($VP \rightarrow V\ NP$) to lexically specific ($VP \rightarrow V_{drink}\ NP$) or bilexically specific ($VP \rightarrow V_{drink}\ NP_{water}$). These multiple representations can be selected and combined to test

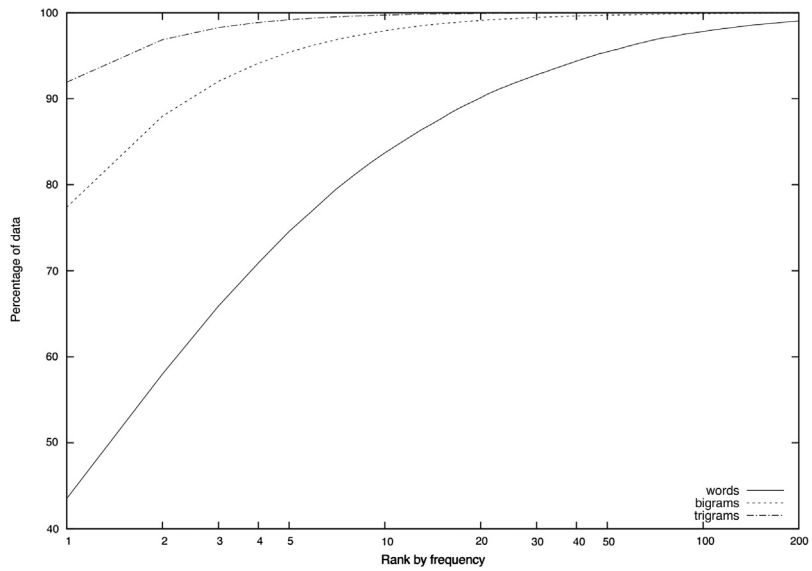


Fig. 6. The vast majority of words and word combinations (*n*-grams) are rare events. The x-axis denotes the frequency of the gram, and y-axis denotes the cumulative% of the grams that appear at that frequency or lower. Based on the Brown Corpus.

their descriptive effectiveness. It turns out that the most generalizing power comes from categorical rules; lexicalization plays an important role in resolving syntactic ambiguities (Collins, 1999) but bilexical rules – the lexically specific combinations that form the cornerstone of usage-based theories (Tomasello, 2003) – offer virtually no additional coverage (Bikel, 2004). The necessity of rules and the limitation of storage are illustrated in Fig. 7.

As seen in the figure, a statistical parser is trained on an increasing amount of data (the x-axis) from the Penn Treebank (Marcus et al., 1999) where sentences are annotated with tree structure diagrams (e.g., Fig. 1). Training involves the statistical parameters in the grammar model, and its performance (the y-axis) is measured by its successful performance in attempting to parse a new data set. An increase in the amount of data presented to the parser during training does improve its performance, but the most striking feature of the figure is found toward the lower end of the x-axis. The vast majority of data coverage is gained on a very small amount of data, between 5% and 10%. The model's impressive success using a small amount of data can only be attributed to highly abstract and general rules, because the parser will have seen very few lexically specific combinations. Thus lessons from language engineering are strongly convergent with the conclusions from the structural and psychological study of child language that we reviewed in Section 2.2: memorization of specific linguistic forms as proposed in the usage-based theories is of very limited value and cannot substitute for the overarching power of a productive grammar even in early child language.

3.2. The trajectories of parameters

The complexity of language was recognized very early in the study of generative grammar. It may appear that complex linguistic phenomena must require equally complex descriptive machinery. However, the rapid acquisition of language by children, despite the poverty and sparsity of linguistic input, led generative linguists to infer that the Language Acquisition Device must be highly struc-

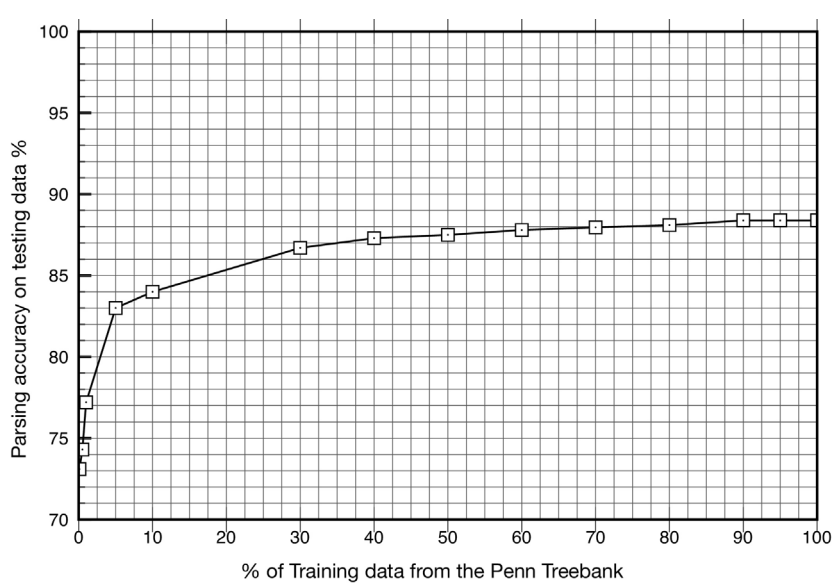


Fig. 7. The quality of statistical parsing as a function of the amount of the training data. (Courtesy of Robert Berwick and Sandiway Fong).

Table 1
Statistical correlates of parameters in the input and output of language acquisition. Very early acquisition refers to cases where children rarely, if ever, deviate from the target form, which can typically be observed as soon as they start producing multiple word combinations. Adapted from Yang (2012).

Parameter	Target	Signature	Input Frequency	Acquisition
Wh-fronting	English	Wh questions	25%	Very early
Topic drop	Chinese	Null objects	12%	Very early
Pro drop	Italian	Null subjects in questions	10%	Very early
Verb raising	French	Verb adverb/*pas*	7%	1;8
Obligatory subject	English	Expletive subjects	1.2%	3;0
Verb second	German/Dutch	OVS sentences	1.2%	3;0–3;2
Scope marking	English	Long-distance questions	0.2%	>4;0

tured in order to promote rapid and accurate acquisition (Chomsky, 1965). The Principles and Parameters framework (Chomsky, 1981) was an attempt to resolve the descriptive and explanatory problem of language simultaneously (see Thornton and Crain, 2013 for a review). The original motivation for parameters comes from comparative syntax. The variation across languages and dialects appear to fall in a recurring range of options (see Baker, 2001 for an accessible introduction): the head-directionality reviewed earlier is a case in point. Parameters provide a more compact description of grammatical facts than construction-specific rules such as those that create cleft structures, subject-auxiliary inversion, passivization, etc. Broadly speaking, the parameterization of syntax can be likened to the problem of dimension reduction in the familiar practice of principal component analysis. Ideally, parameters should have far-ranging implications, such as the combination of a small number of parameters can yield a much larger set of syntactic structures (see Sakas and Fodor, 2012 for an empirical demonstration): the determination of the parameter values will thus simplify the task of language learning.

Some, perhaps most, linguistic parameters are set very early. One of the best-studied cases is the placement of finite verbs in the main clause. The finite verb appears before certain adverbs in French (e.g., *Paul travaille toujours*) but follows the corresponding adverbs in English (e.g., *Paul always works*); the difference between these languages is a frequent source of error for adult second language learners (White, 1990). Children, however, almost never get this wrong. For instance, Pierce (1992) finds virtually no errors in French-learning children's speech starting at the 20th month, the very beginning of multi-word combinations when verb placement relative to adverbs becomes observable. Studies of languages with French-like verb placement show similarly early acquisition by children (Wexler, 1998). Likewise, English children almost never produce verb placement errors (e.g., "I eat often pizza").

Clearly, the differences between English and French must be determined on the basis of children's input experience. Many sentences that children hear, however, will not bear on this parametric decision: a structure such as "Jean voit Paul" or "John sees Paul" contains no adverb landmarks that signify the position of the verb. If children are to make a binary choice for verb placement in French, they can only do so on examples such as *Paul travaille toujours.* This in turn invites us to examine language-specific input data to determine the quantity of disambiguating data available to language learners. In the case of French, about 7% of sentences contain finite verbs followed by adverbs; this provides an empirical benchmark for the volume of evidence necessary for very early acquisition of grammar. Table 1 summarizes the developmental trajectories of several major syntactic parameters, which clearly show the quantitative role of experience.

Several additional remarks can be made about parameter setting. First, the sensitivity to the quantity of evidence suggests that parameter setting is gradual and probabilistic and may involve domain-general learning processes: we return to this issue in Section 4.2. Second, consider a well-known problem in the acquisition of English, a language that in general requires the use of the grammatical subject ("obligatory subject" in Table 1; Bloom, 1970; Hyams, 1986). Despite the consistent use of subjects in the child-directed input data, English-learning children frequently omit subjects—up to 30% of the time—and they occasionally omit objects as well, although the intended meanings are generally deducible from the context:

(6) a. __want cookies.
b. Where __ going?
c. How __ wash it?
d. Erica took __
e. I put __ on

This so-called subject/object drop stage persists until around a child's third birthday, when children begin to use subjects and objects consistently like adults (Valian, 1991).

If children are merely replicating the input data, the subject/object drop stage would be puzzling since the expressions in (6) are ungrammatical and thus not present in the input. The acquisition facts are also difficult to account for if the grammar consists of construction-specific rules such as (probabilistic) phrase structure grammar. For instance, the rule "S → NP VP" is consistently supported by the data, and children should have no difficulty acquiring the presence of the NP subject, or learning that the rule has a probability close to 1.0. As discussed at length by , cases such as the acquisition of subject use in English provide strong support for the theory of parameters. There are types of languages for which the omitted subject is permissible as it can be recovered on the basis of verb agreement morphology (*pro-drop*, as in Spanish and Italian) or discourse context (*topic-drop*, as in Chinese and Japanese). There are also languages including some in the family of Slavic languages and languages native to Australia for which both *pro-drop* and *topic-drop* are possible. Indeed, the omission of subjects and objects by English-learning children shows striking parallels with the omission of subjects and objects by Chinese-speaking adults (Hyams, 1991; Yang, 2002). For instance, where English children omit subjects in *wh*-questions, the question words that accompany children's omissions are *where, how, when,* as in *"Where __ going"* and "*How __ wash it*". Children almost never omit subjects in questions with the question words *what or who*; that is, children do not produce non-adult questions such as "*Who __ see"* which corresponds to the adult form "Who (did) you see?" where the subject has been omitted and the question word *who* has been extracted from the object position, following the verb *see*.

The pattern of omission in child English is exactly the same as the pattern of omission in topic-drop languages such as Chinese. In those languages, subject/object omission is permissible if and only if its identity can be recovered from the discourse. If a modifier of a predicate (e.g., describing manner, time) has a prominent position in the discourse, the identification of the omitted subject/object is unaffected. But if a new argument of a predicate (e.g., concerning who and what) becomes prominent, then discourse linking is disrupted and subject/object omission is no longer possible. In other words, English-learning children drop subjects/objects only in the same discourse situation when Chinese-speaking adults may do so. This suggests that topic-drop, a grammatical option that is exercised by many languages of the world, is spontaneously available to children, but it is an option that needs to be winnowed out during

the course of acquisition for languages such as English: a classic example of *use it or lose it*. The telltale evidence for the obligatoriness of subjects is the use of *expletive* subjects. In a sentence such as "There is a toy on the floor", the occupant of the subject position is purely formal and non-thematic, as the true subject is "a toy". Indeed, only obligatory subject languages such as English have expletive subjects, and the counterpart of "There is a toy on the floor" in Spanish and Chinese leave the subject position phonetically empty. In other words, while the overwhelming majority of English input sentences contain a thematic subject, such as a noun phrase or a pronoun, the input serves no purpose whatever in helping children learn the grammar, because thematic subjects are universally allowed and do not disambiguate the English type languages from the Spanish or the Chinese type. It is only the cumulative effect of expletive subjects, which are used in only about 1% of all English sentences, that gradually drives children toward their target parametric value (Yang, 2002).

Children's non-adult linguistic structures often follow the same pattern, such that children differ from adult speakers of the local language just in ways that adult languages differ from each other. In the literature on language acquisition, this is called the Continuity Assumption. Other examples of the Continuity Assumption are discussed in Crain et al. (this issue). To cite just one example, Chinese and English differ in how disjunction words are interpreted in negative sentences. In English, the negative sentence *Al didn't eat sushi or pasta*, entails that Al didn't eat sushi and that he didn't eat pasta. Negation takes scope over the English disjunction word *or*. However, the word-by-word analogue of the English sentence does not exhibit the same scope relations in Chinese, so adult speakers of Chinese judge a negative sentence with disjunction to be true as long as one of the disjunction is false, for example when Al didn't eat sushi, but did eat pasta. To express this interpretation, English speakers use a cleft structure, where the scope relations between negation and disjunction are reversed, as in the sentence *It was sushi or pasta that Al didn't eat*. Children acquiring Chinese, however, assign the English interpretation to sentences in which disjunction appears in the scope of negation, such as *Al didn't eat sushi or pasta*. Children acquiring Chinese seem to be speaking a fragment of English, rather than Chinese. The examples we have cited in this section are clearly not "errors" in the traditional sense. Rather children's non-adult linguistic behavior is compelling evidence of their adherence to the principles and parameters of Universal Grammar.

3.3. Negative evidence and inductive learning

Together with the general structural principles of language reviewed in Section 2, the theory of parameters has been successfully applied to comparative and developmental research. The motivation for parameters is convergent with results from machine learning and statistical inference (Gold, 1967; Valiant, 1984; Vapnik, 2000), all of which point to the necessity of restricting the hypothesis space to achieve learnability (see Niyogi, 2006, for an introduction).

It remains true, however, that the description and acquisition of language cannot be accounted for entirely by the universal principles and parameters. Language simply has too many peripheral idiosyncrasies that must be learned from experience (Chomsky, 1981). The acquisition of morphology and phonology most clearly illustrates the inductive and data-driven nature of these problems. Not even the most diehard nativist would suggest that the "add -ed" rule for the English past tense is available innately, waiting to be activated by regular verbs such as *walk-walked*. Furthermore, linguists are fond of saying all grammars leak (Sapir, 1928). Rules often have unpredictable exceptions: English past tense, of course, has some 150 irregular verbs that do not take "-ed" and they must be rote-learned and memorized (Chomsky and Halle, 1968). In the domain of syntax, the need for inductive learning is also clear. A well-studied problem concerns the acquisition of dative constructions (Baker, 1979; Pinker, 1989):

(7) a. John gave the ball to Bill.
John gave Bill the ball.
b. John assigned the problem to Bill.
John assigned Bill the problem.
c. John promised the car to Bill.
John promised Bill the car.
d. John donated the picture to the museum.
*John donated the museum the picture.
e. *John guaranteed a victory to the fans.
John guaranteed the fans a victory.

Examples such as (7a-c) seem to suggest that the double-object construction (Verb NP NP) and the *to*-dative construction (Verb NP to NP) are mutually interchangeable, only to be disconfirmed by the counterexamples of *donate* (7d) and *guarantee* (7e) for which only one of the constructions is available; see Levin (1993, 2008) for many similar examples in English and in other languages.

How do children avoid these traps of false generalizations? Not by parents telling them off. The problem of inductive learning in language acquisition has unique challenges that set it apart from learning problems in other domains. Most significantly, language acquisition must succeed in the absence of negative evidence. Study after study of child-adult interactions have shown that learners do not receive (effective) correction of their linguistic errors (Brown and Hanlon, 1970; Braine, 1971; Bowerman, 1988; Marcus, 1993). And there are cultures where children are not treated as equal conversational partners with adults until they are linguistically and socially competent (Heath, 1983; Schieffelin and Ochs, 1986; Allen, 1996). Thus, the learner must succeed to acquire a grammar based on a set of positive examples produced by speakers of that grammar. This is very different from many other domains of learning and skill acquisition—be it bike-riding, mathematics, and many studies in the psychology of learning literature—where explicit instructions and/or negative feedback are available. For example, most applications of neural networks including deep learning (LeCun et al., 2015) are "supervised": the learner's output can be compared with the target form such that errors can be detected and used for correction to better approximate the target function. Viewed in this light, linguistic principles such as Structure Dependence and the constraint on coreference reviewed in Section 2 are most likely accessible to children innately. These are *negative* principles, which prohibit certain linguistic structures or interpretations (e.g., certain expressions cannot have the same referential content). They were discovered through linguistic analysis that relies on the creation of ungrammatical examples using highly complex structures, which are clearly absent in the input.

Starting with Gold (1967), the special conditions of language acquisition have inspired a large body of formal and empirical work on inductive learning (see Osherson et al., 1986 for a survey). The central challenge is the problem of generalization: How does the learner determine the grammatical status of expressions not observed in the learning data? Consider the *Subset Problem* illustrated in Fig. 8.

Suppose the target grammar is the subset *g* but the learner has mistakenly conjectured *G*, a superset. In the case of the dative constructions, *g* would be the correct rules of English but *G* would be the grammar that allows the double object construction for all relevant verbs (e.g., the ungrammatical *I donated the library a book*). If the learner were to receive negative evidence, then the instances marked by "+"—possible under G but not under g—would be greeted with adult disapproval or some other kind of feedback. This would inform the learner of the need to zoom in on a smaller grammar. But the absence of negative evidence in language acquisition makes it impossible to deploy this useful learning strategy.

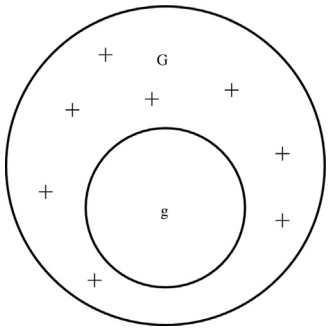


Fig. 8. The target, and smaller, hypothesis *g* is a proper subset of the larger hypothesis *G*, creating learnability problems when no negative evidence is available.

One potential solution to the Subset Problem is to provide the learner with additional information about the nature of the examples in the input (Gold, 1967). For instance, if the child surmises that the absence of a certain string in the input entails its ungrammaticality, then positive examples may substitute for negative examples—dubbed *indirect negative evidence* (Chomsky, 1981)—and the task of learning is greatly simplified. Of course, the absence of evidence is *not* the evidence of absence, and the use of indirect negative evidence and its formally similar formulations (Wexler and Culicover, 1980; Clark, 1987) have to be viewed with suspicion (see Pinker, 1989, for review). For example, usage-based language researchers such as Tomasello (2003) assert that children do not produce "John donated the museum the picture" because they consistently observe the semantically equivalent paraphrase "John donated the picture to the museum": the double-object structure is thus "preempted". But this account cannot be correct. The most frequent errors in children's acquisition of the dative constructions documented by Pinker (1989) and in fact other usage-based researchers (Bowerman and Croft, 2005) are utterances such as "I said her no": the verb "say" only, and very frequently, appears in a prepositional dative construction (e.g., "I said Hi to her") in adult input and should have preempted the double-object usage in children's language.

More recently, indirect negative evidence has been reformulated in the Bayesian approach to learning and cognition (e.g., Xu and Tenenbaum, 2007). In the Subset Problem, if the grammar *G* is expected to generate certain kinds of examples (the "+" examples) which will not appear in the input data generated by *g*, then the learner may have good reasons to reject *G*. In Bayesian terms, the *a posteriori* probability of *G* is gradually reduced when the sample size increases. But it remains questionable whether indirect negative evidence can be effectively used in the practice of language acquisition. On the one hand, indirect negative evidence is generally computationally intractable to use (Osherson et al., 1986; Fodor and Sakas, 2005): for this reason, most recent models of indirect negative evidence explicitly disavow claims of psychological realism (Chater and Vitányi, 2007; Xu and Tenenbaum, 2007; Perfors et al., 2011).

A second consideration has to do with the statistical sparsity of language reviewed earlier. As we discussed, examples of both ungrammatical and grammatical expressions are often absent or equally (im)probable in the learning sample and would all be rejected (Yang, 2015): the use of indirect negative evidence has considerable collateral damage, even if the computational efficiency issue is overlooked. As such, it is in the interest of the learner as well as the theorist to avoid the postulation of over-general hypotheses to the extent possible.

In some cases, it appears that language has universal default states that places the learner's starting point in the subset grammar (the Subset Principle; Berwick, 1985); see Crain (2012) and Crain et al. (this issue) for studies of the acquisition of semantic properties. Thus, in Fig. 8, the learner will start at *g* and remain there if *g* is indeed the target. If the target grammar is in fact the superset *G*, the learner will consider it and only positive evidence for it is presented (e.g., "+" examples). The Subset Problem does not arise. However, children *do* occasionally over-generalize during the course of language acquisition. In the acquisition of dative constructions, for example, many young children produce sentences such as "I said her no" (Pinker, 1989), clearly ungrammatical for the adult grammar. Thus children may indeed postulate an over-general grammar, only to retreat from it as learning proceeds: the Subset Problem must be solved when it arises.

Here we briefly review a recent proposal of how children acquire productive rules in the inductive learning of language, the Tolerance Principle (Yang, 2016). Following a traditional theme in the research on linguistic productivity (Aronoff, 1976), a rule is deemed productive if it applies to a sufficiently large number of items to which it's applicable. But if a rule has too many exceptions, then the learner will decide against its productivity. The Tolerance Principle provides a calculus of quantifying the cost of exceptions:

(8) Tolerance Principle

If R is a productive rule applicable to N candidates in the learning sample, then the following relation holds between N and e, the number of exceptions that could, but do not, follow R:

$$e \leq \theta_N \text{ where } \theta_N = N/\ln(N)$$

The Tolerance Principle is motivated by the computational mechanism of how language users process rules and exceptions in real time, and the closed form solution makes use of the fact that the probabilities of *N* words can be well characterized by Zipf's Law, where the *N-th* Harmonic number can be approximated by the function 1/lnN. The slow growth of the N/ln(N) function suggests that a productive rule must be supported by an overwhelming number of rule-following items to overcome the exceptions.

Experiments in the acquisition of artificial languages (Schuler et al., 2016) have found near-categorical support for the Tolerance Principle. Children between the age of 5 and 7 were presented with 9 novel objects with labels. The experimenter produced suffixed "singular" and "plural" forms of those nouns as determined by their quantity on a computer screen. In one condition, five of the nouns share a plural suffix and the other four have individually specific suffixes. In another condition, only three share a suffix and the other six are all individually specific. The nouns that share the suffix can be viewed as the regulars and the rest of the nouns are the irregulars. The choice of 5/4 and 3/6 was by design: the Tolerance Principle predicts the productive extension of the shared suffix in the 5/4 condition because 4 exceptions fall just below the threshold ($\theta_9 = 4.1$), but it predicts that there will be no generalization in the 3/6 conditions. In the latter case, despite the statistical dominance of the shared suffix as the most frequent suffix, the six exceptions exceed the threshold. After training, children were presented with novel items in the singular and were prompted to produce the plural form. Nearly all children in the 5/4 condition generalized the shared suffix on 100% of the test items in a process akin to the productive use of English past tense "-ed". In the 3/6 condition, almost no child showed systematic usage of any suffix: none cleared the threshold for productivity.

The Tolerance Principle has been applied to many empirical cases in language acquisition. As a parameter-free model, there is no need to statistically fit any empirical data: corpus counts of rule-following and rule-defying words alone can be used to predict, leading to sharp and well-confirmed predictions (Yang, 2016). Here we summarize its application to dative acquisition. The acquisition process unfolds in two steps. First, the child observes verbs that appear in the double object usage ("Verb NP NP"). Quantitative

analysis was carried out for a five-million-word corpus of child-directed English, which roughly corresponds to a year's input data. The overwhelming majority of double-object verbs—38 out of 42, easily clearing the productivity threshold—have the semantic property of "transferred possession" (Levin, 1993), be it physical object, abstract entity, or information (*give X books, give X regards, give X news*). Second, the child tests the validity of this semantic generalization: of the verbs that have the transferred possession semantics, how many are in fact attested in the double object construction. Again, the productivity threshold was met: although the five-million-word corpus contains 11 verbs that have the appropriate semantics but do not appear in the double object construction, 38 out of 49 is sufficient. This accounts for children's overgeneralization errors such as "I said her no": *say*, as a verb of communication, meets the semantic criterion and is thus automatically used in the double-object construction. However, this stage of productivity is only developmentally transient. Analysis of larger speech corpora, which are reflective of the vocabulary size and input experience of more mature English speakers, shows that the once productive mapping between the transferred possession semantics and the double-object syntax cannot be productively maintained. The lower-frequency verbs such as *donate*, which are unlikely to be learned by young children, thus will not participate in the double-object construction. For *say,* the learner will only follow its usage in the adult input, thereby successfully retreating from the overgeneralization of "I said her no".

4. Language acquisition in the light of biological evolution

We hope that our brief review has illustrated the kinds of rich empirical results that have been forthcoming from decades of cross-linguistic research. Like all scientific theories, our understanding of language acquisition will always remain a work in progress as we seek a more precise and deeper understanding of child language. At the same time, theories of linguistic structures and child language acquisition also contribute to the study of second language acquisition and teaching (see White, 2003; Slabakova, 2016). In this section, we offer some specific suggestions and speculations on the future prospects of acquisition research from the biolinguistic perspective.

4.1. Merge, cognition, and evolution

While anatomically modern humans diverged from our closest relatives some 200,000 years ago (McDougall et al., 2005), the emergence of language use was probably later—80–100,000 years, as suggested by the dating of the earliest undisputed symbolic artifacts (Henshilwood et al., 2002; Vanhaeren et al., 2006) although the Basic Property of language (i.e., Merge) may have been available as early as 125,000 years ago, when the ancestral Khoe-San groups diverged from human lineages and have remained genetically largely isolated (see Huybregts, this issue). In any case, the emergence of language is a very recent evolutionary event: it is thus highly unlikely that all components of human language, from phonology to morphology, from lexicon to syntax, evolved independently and incrementally during this very short span of time. Therefore, a major impetus of current linguistic research is to isolate the essential ingredient of language that is domain specific and can plausibly be attributed to the minimal evolutionary changes in the recent past. At the same time, we need to identify principles and constraints from other domains of cognition and perception which interact with the computation of linguistic structure and language acquisition and which may have more ancient evolutionary lineages (Chomsky, 1995, 2001; Hauser et al., 2002)

As reviewed earlier, the findings in language acquisition support Merge as the Basic Property of language (Berwick and Chomsky, 2016). A sharp discontinuity is observed across species. Children create and use compositional structures as seen in babbling, phonemic acquisition, and the first use of multiple word combinations, including home signers who do not have a systematic input model. Our closest relative such as Nim can only store and retrieve fragments of fixed expressions, even though animals (Terrace et al., 1979; Kaminski et al., 2004; Pepperberg, 2009) are capable of forming associations between sounds and meanings (which are likely quite different from word meanings in human language; see Bloom, 2004). This key difference, which we believe is due to the species-specific capacity that is conferred on humans by Merge, casts serious doubt on proposals that place human language along a continuum of computational systems in other species, including recapitulationist proposals that view children's language development as retracing the course of language evolution (e.g., Bickerton, 1995; Studdert-Kennedy, 1998; Hurford, 2011). Similarly, the hierarchical properties of Merge stand in contrast with the properties in the visual and motor systems, which have been suggested to be related to linguistic structures and evolution (Arbib, 2005; Fitch and Martins, 2014). It is not sufficient to draw analogies between visual representation or motor planning to the structure of language. The hierarchical nature of language has been established by rigorous demonstrations—for instance, with tools from the theory of formal language and computation—that simpler systems such as finite state concatenation or context-free systems are in principle inadequate (Chomsky, 1956; Huybregts, 1976; Shieber, 1985). Even putting the structural aspects of language aside and focusing only on the properties of strings, we have seen a convergence of results from very different formal models of grammar to suggest that human language lies in the complexity class known as mildly context-sensitive languages (Joshi et al., 1991; Stabler, 1996; Michaelis, 1998). At the very minimum, the reductionist accounts of language evolution need to formalize the structural properties of the visual or motor system and demonstrate their similarities with language.

It is also important to clarify that the Basic Property refers to the *capacity* of recursive Merge. As a matter of logic, this does not imply that infinite recursion must be observed in every aspect of every human language. One needn't wander far from the familiar Indo-European languages to see this point. Consider the contrast in the NP structures between English and German:

(9) a. Maria's neighbor's friend's house
b. Marias Haus
c. *Marias Nachbars Freundins Haus

In English, a noun phrase allows unlimited embedding of possessives as in (9a). In German, and in most Germanic languages (Nevins, Pesetsky, and Rodrigues, 2009), the embedding stops at level 1 (9b) and does not extend further, as illustrated by the ungrammaticality of (9c). Presumably, German children do not allow the counterpart to (9a) because they do not encounter sufficient evidence that enables such an inductive generalization, whereas English children do: this constitutes an interesting language acquisition problem in its own right. But no one would suggest that German, or a German-learning child, lacks the capacity for infinite embedding on the basis of (9). Similar patterns can be found in the numeral systems across languages. The languages familiar to the readers all have an infinite number system, perhaps a fairly recent cultural development, but one that has deep structural similarities across languages (Hurford, 1975). At the same time, there are languages with a small and finite number of numerals that in turn affect the native speaker's numerical performance (Gordon, 2004; Pica et al., 2004; Dehaene, 2011). All the same, the ability to acquire an infinite numerical system (in a second language; Gelman and Butterworth, 2005) or infinite embedding in

any linguistic domain (Hale, 1976) is unquestionably present in all language speakers.

Once Merge became available, all the other properties of language should fall into place. This is clearly the simplest evolutionary scenario, and it invites us to search for deeper and more general principles, including constraints from nonlinguistic domains, that underlie the structural properties of language uncovered by the past few decades of generative grammar research, including those that play a crucial role in the acquisition of language reviewed earlier. However, to avoid Panglossian fallacies, the connection between language and other domains of cognition is compelling only if it offers explanations for very specific properties of language. Empirical details matter, and they must be assessed on a case-by-case basis. For example, it has been suggested that social development (e.g., theory of mind) plays a causal role in the development and evolution of language (Tomasello, 2003). To be credible, however, this approach must provide a specific account of the structural and developmental properties of language currently attributed to Merge and Universal Grammar, such as those reviewed in these pages. In fact, the social approach to language does not fare well on much simpler problems. For instance, it has been suggested that social and pragmatic inference may serve as an alternative to constraints specifically postulated for the learning of word meanings (Bloom, 2000). Social cues may indeed play an important role in word learning: for example, young children direct attention to objects by following the eye-gaze of the caretaker (Baldwin, 1991). At the same time, blind children without access to similarly richly social cues nevertheless manage to acquire vocabulary in strikingly similar ways as sighted children (Landau and Gleitman, 1984). Over the years, accounts based on social and pragmatic learning (e.g., Diesendruck and Markson, 2001) have been suggested to replace the powerful Mutual Exclusivity constraint (Markman and Wachtel, 1988), according to which labels are expected to be uniquely associated with objects. In our assessment, the constraint-based approach still provides a broader account of word learning than the social/pragmatic account (Markman et al., 2003; Halberda, 2003), especially in light of studies involving word learning by children with autism spectrum disorders (de Marchena et al., 2011).

More challenging for the research program suggested here is the fact of language variation. Why are there so many different languages even though the cognitive capacities across human individuals are largely the same? Why do we find the locus of language variation along certain specific dimensions but not others? For example, why does verb placement across languages interact with tense, as we have seen in the structure and acquisition of English and French, but not with valency (the number of arguments for a verb)? What is the conceptual basis for the rich array of adverbial expressions that are nevertheless rigidly structured in specific syntactic positions (Cinque, 1999)? The noun gender system across languages, to varying degrees, is based on biological gender, even though it has been completely formalized in some languages where gender assignment is arbitrary. Conceivably, gender is an important distinction rooted in human cognition. However, we are not aware of any language that has grammaticalized colors to demarcate adjective classes, even though color categorization is also a profound feature of cognition and perception. Indeed, while parameters are very successful in the comparative studies of language and have direct correlates in the quantitative development of language (Section 3.2), they are unlikely to have evolved independently and thus may be the reflex of Merge and the property of the sensorimotor system that must impose a sequential order when realizing language in speech or signs. We can all agree that the simpler conception of language with a reduced innate component is evolutionarily more plausible—which is exactly the impetus for the Minimalist Program of language (Chomsky, 1995). But this requires working out the details of specific properties so richly documented in the structural and developmental studies of languages. Le biologiste passe, la grenouille reste.

4.2. The mechanisms of the language acquisition device

In the concluding paragraphs, we briefly discuss three computational mechanisms that play significant roles in language acquisition. They all appear to follow more general principles of learning and cognition, including principles that are not restricted to language, such as principles of efficient computation (Chomsky, 2005).

First, language acquisition involves the *distributional analysis* of the linguistic input, including the creation of linguistic categories. It was suggested long ago (Harris, 1955; Chomsky, 1955) that higher-level linguistic units such as words and morphemes might be identified by the local minima of transitional probabilities over adjacent low-level units such as syllables and phonemes. Indeed, eight-month-old infants can use transitional probability to segment words in artificial language (Saffran et al., 1996). While similar computational mechanisms have been observed in other learning tasks and domains (Saffran et al., 1999; Fiser and Aslin, 2001), they seem constrained by domain-specific factors (Turk-Browne et al., 2005) including prosodic structures in speech (Johnson and Jusczyk, 2001; Yang, 2004; Shukla et al., 2011). Similarly, in the much-studied acquisition of English past tense acquisition, recent work (Yang, 2002; Stockall and Marantz, 2006) suggests that the irregular verbs are learned and organized into classes whose membership is arbitrary and closed (Chomsky and Halle, 1968). For instance, an irregular verb class is characterized by the rule that changes the rime to "ought", which applies to *bring, buy, catch, seek, teach,* and *think* and nothing else. As such, children's acquisition of past tense contains virtually no errors of the irregular patterns despite phonological similarities: they do not produce *gling-glang/glung* following *sing-sang/sung* in experimental studies using novel words (Berko, 1958), nor do they produce errors such as *hatch-haught* following *catch-caught* in natural speech (Xu and Pinker, 1995). The only systematic errors are the extension of productive rules such as *hold-holded* (Marcus et al., 1992); see Yang (2016) for proposals of how productivity is acquired. The rule-based approach to morphological learning seems surprising because the irregular verbs in English form very small classes: the direct association between the stem and the past-tense form is a simpler approach (Pinker and Ullman, 2002; McClelland and Patterson, 2002). But children's persistent creation of arbitrary classes recalls strategies from the study of categorization. Specifically, there is striking similarity between morphological learning and the one-dimensional sorting bias (e.g., Medin et al., 1987), where all experimental subjects categorically group objects by some attribute rather than seeking to construct categories on the basis of overall similarity.

In the case of English past tense, the privileged dimension for category creation seems to be the shared morphological process: *bring, buy, catch, seek, teach,* and *think* are grouped together because they undergo the same change in past tense ("ought"). Such cases are commonplace in the acquisition of language. For example, Gagliardi and Lidz (2014) studied the acquisition of noun class in Tzez, a northeastern Caucasian language spoken in Dagestan. Tzez nouns are divided into four classes, each with distinct agreement and case reflexes in the morphosyntactic system. Analysis of speech corpus shows that the semantic properties of nouns provide statistically strong cues for noun class membership. However, children choose the statistically weak phonological cues to noun classes. Taken together, distributional learning is broadly implicated in language acquisition while its effectiveness relies on adapting to the specific constraints in each linguistic domain.

Second, language acquisition incorporates *probabilistic learning* mechanisms widely used in other domains and species. This is especially clear in the selection of parametric choices in the acquisition of syntax. Traditionally, the setting of linguistic parameters was conjectured to follow discrete and domain-specific learning mechanisms (Gibson and Wexler, 1994; Tesar and Smolensky, 1998). For example, the learner is identified with one set of parameter values at any time. Incorrect parameter settings, which will be contradicted by the input data, are abandoned and the learner moves on to a different set of parameter values. This on-or-off ("triggering") conception of parameter setting, however, is at odds with the gradual acquisition of parameters (Valian, 1991; Wang et al., 1992). Indeed, the developmental correlates of parameters that were reviewed in Section 2.2, including the quantitative effect of specific input data that disambiguate parametric choices, only follow if parameter setting is probabilistic and gradual. A recent development, the variational learning model (Yang, 2002), proposes that the setting of syntactic parameters involves learning mechanisms first studied in the mathematical psychology of animal learning (Bush and Mosteller, 1951). Parameters are associated with probabilities: parameter choices consistent with the input data are rewarded and those inconsistent with the input data are penalized. On this account, the developmental trajectories of parameters are determined by the quantity of disambiguating evidence along each parametric dimension, in a process akin to Natural Selection where the space of parameters provide an adaptive landscape. Children's systematic errors that defy the input data, as in the phenomenon of subject/object omission by children acquiring English and the English scope assignments that are generated by children acquiring Chinese, are attributed to non-target parametric options before their ultimate demise. The computational mechanism is the same—for a rodent running a T-maze and for a child setting the head-directionality parameter—even though the domains of application cannot be more different. Along similar lines, recent work suggests that probabilistic learning mechanisms also add robustness to word learning. In a series of studies, Gleitman, Trueswell, and colleagues have demonstrated that when subjects learn word-referent associations, they do not keep track of all co-occurrence relations in the environment as previously supposed (e.g., Yu and Smith, 2007) but selectively attend to a small number of hypotheses (Medina et al., 2011; Trueswell et al., 2013). Adapting the variational learning model for word learning, Stevens et al. (2016) show that associating probabilities with these hypotheses significantly broadens the coverage of experimental findings, and the resulting model outperforms much more complex computational models of word learning (e.g., Frank et al., 2009) when tested on corpora of child-directed English input.

Third, the inductive learning mechanism for language appears grounded in the principle of *computational efficiency*. Examples of computational efficiency include computing the shortest possible movement of a constituent, and keeping to a minimum the number of copies of expressions that are pronounced. Where efficient computation competes with parsing considerations, efficient computation apparently wins. For example, in resolving ambiguities, providing a copy of the displaced constituent in filler-gap dependencies would clearly help the parser and therefore aid the listener or reader in identifying the intended meanings of ambiguous sentences. However, the need to minimize copies, to satisfy computational efficiency, apparently carries more weight than does the ease of comprehension, so copies of moved constituents are deleted despite the potential difficulties in comprehension that may ensue. Similar considerations of parsimony can be traced back to the earliest work in generative grammar (Chomsky, 1955/1975; Chomsky, 1965): the target grammar is selected from a set of candidates by the use of an *evaluation metric*. One concrete formulation favors the most compact descriptions of the input corpus, which has been developed in the machine learning literature as the Minimum Description Length (MDL) principle (Rissanen, 1978). The Subset Principle (Angluin, 1980; Berwick, 1985) requires the learner to generalize as conservatively as possible. It thus also follows the principle of simplicity, by forcing the learner to consider the hypothesis with the smallest extension possible.

The Tolerance Principle (Yang, 2016) approaches the problem of efficient computation from the perspective of real-time language processing. The calculus for productivity is motivated by reaction time studies of how language speakers process rules and exceptions in morphology. There is evidence that to generate or process words that follow a productive rule (e.g., *cat* which takes the regular plural −s), the exceptions (i.e., *foot, people, sheep* etc.) must be evaluated and rejected—much like an IF-THEN-ELSE statement in programming languages. An increase in the number of exceptions correspondingly leads to processing delay for the rule-following items. The productive threshold (8) is analytically derived by comparing the expected processing cost of having a productive rule with a list of exceptions and that of having no productive rule at all where all items are listed. In other words, the learner chooses a grammar that results in *faster* processing time. In addition to the empirical cases discussed in Yang (2016), the Tolerance Principle provides a plausible account of why children are such adept language learners. The threshold function (N/lnN) entails that if a rule is evaluated on a smaller set of words (N), the number of tolerable exceptions is a larger proportion of N. That is, the productivity of rules is easier to detect if the learner has a small vocabulary—as in the case of young children. The approach dovetails with the suggestion from the developmental literature that it may be to the learner's advantage to "start small" (Elman, 1993; Newport, 1990). Again, language learning needs to be simple.

We hope to have conveyed some new results and syntheses, although necessarily selectively, from the many exciting research developments that continue to enhance our understanding of child language acquisition. While the connection with linguistics and psychology has always been central, language acquisition is now increasingly engaged with computational linguistics, comparative cognition, and neuroscience. The growth of language is nothing short of a miracle: the full scale of linguistic complexity in a toddler's grammar still eludes linguistic scientists and engineers alike, despite decades of intensive research. Considerations from the perspective of human evolution have placed even more stringent conditions on a plausible theory of language. The integration of formal and behavioral approaches, and the articulation of how language is situated among other cognitive systems, will be the defining theme for language acquisition research in the years to come.

Acknowledgments

We are grateful to Riny Huybregts for valuable comments on an earlier version of the manuscript. J.J.B. is part of the Consortium on Individual Development (CID), which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO; grant number 024.001.003).

References

Allen, S., 1996. Aspects of Argument Structure Acquisition in Inuktitut. John Benjamins Publishing, Amsterdam.
Angluin, D., 1980. Inductive inference of formal languages from positive data. Inf. Contr. 45, 117–135.
Angluin, D., 1982. Inference of reversible languages. J. ACM 29, 741–765.
Arbib, M.A., 2005. From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. Behav. Brain Sci. 28, 105–124.
Aronoff, M., 1976. Word Formation in Generative Grammar. MIT Press, Cambridge, MA.

Baker, C.L., 1979. Syntactic theory and the projection problem. Ling. Inq. 10, 533–581.
Baker, M., 2001. The Atoms of Language: The Mind's Hidden Rules of Grammar. Basic Books, New York.
Baldwin, D.A., 1991. Infants' contribution to the achievement of joint reference. Child development 62, 874–890.
Baroni, M., 2009. Distributions in text. In: Lüdeling, A., Kyöto, M. (Eds.), Corpus Linguistics: An International Handbook. Mouton de Gruyter, Berlin, pp. 803–821.
Bergelson, E., Swingley, D., 2012. At 6–9 months, human infants know the meanings of many common nouns. Proc. Natl. Acad. Sci. U. S. A. 109, 3253–3258.
Berko, J., 1958. The child's learning of English morphology. Word 14, 150–177.
Berwick, R.C., Chomsky, N., 2011. The biolinguistic program: the current state of its development. In: Di Sciullo, A.M., Boeckx, C. (Eds.), The Biolinguistic Enterprise. Oxford University Press, pp. 19–41.
Berwick, R.C., Chomsky, N., 2016. Why Only Us: Language and Evolution. MIT Press, Cambridge, MA.
Berwick, R.C., Pilato, S., 1987. Learning syntax by automata induction. Machine Learn. 2, 9–38.
Berwick, R.C., Okanoya, K., Beckers, G.J., Bolhuis, J.J., 2011a. Songs to syntax: the linguistics of birdsong. Trends Cogn. Sci. 15, 113–121.
Berwick, R.C., Pietroski, P., Yankama, B., Chomsky, N., 2011b. Poverty of the stimulus revisited. Cogn. Sci. 35, 1207–1242.
Berwick, R.C., Friederici, A.D., Chomsky, N., Bolhuis, J.J., 2013. Evolution, brain, and the nature of language. Trends Cogn. Sci. 17, 89–98.
Berwick, R., 1985. The Acquisition of Syntactic Knowledge. MIT Press, Cambridge, MA.
Bickerton, D., 1995. Language and Human Behavior. University of Washington Press, Seattle, WA.
Bijeljac-Babic, R., Bertoncini, J., Mehler, J., 1993. How do 4-day-old infants categorize multisyllabic utterances? Dev. Psych. 29, 711–721.
Bikel, D.M., 2004. Intricacies of Collins' parsing model. Comput. Ling. 30, 479–511.
Bloom, L., 1970. Language Development: Form and Function in Emerging Grammar. MIT Press, Cambridge, MA.
Bloom, P., 2000. How Children Learn the Meanings of Words. MIT Press, Cambridge, MA.
Bloom, P., 2004. Can a dog learn a word? Science 304, 1605–1606.
Bolhuis, J.J., Everaert, M. (Eds.), 2013. Birdsong, Speech, and Language. Exploring the Evolution of Mind and Brain. MIT Press, Cambridge, MA.
Borden, G., Gerber, A., Milsark, G., 1983. Production and perception of the/r/-/l/contrast in Korean adults learning English. Lang. Learn. 33 (4), 499–526.
Bowerman, M., 1988. The 'no negative evidence' problem: how do children avoid constructing an overly general grammar? In: Hawkins, J.A. (Ed.), Explaining Language Universals. Basil Blackwell, Oxford, pp. 73–101.
Braine, M.D., 1971. On two types of models of the internalization of grammars. In: Slobin, D.I. (Ed.), The Ontogenesis of Grammar: A Theoretical Symposium. Academic Press, New York, pp. 153–186.
Brown, R., Hanlon, C., 1970. Derivational complexity and the order of acquisition in child speech. In: Hayes, J.R. (Ed.), Cognition and the Development of Language. Wiley, New York, pp. 11–53.
Brown, R., 1973. A First Language: The Early Stages. Harvard University Press, Cambridge, MA.
Bush, R.R., Mosteller, F., 1951. A mathematical model for simple learning. Psychol. Rev. 68, 313–323.
Charniak, E., Johnson, M., 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics Association for Computational Linguistics.
Chater, N., Vitányi, P., 2007. Ideal learning of natural language: positive results about learning from positive evidence. J. Math. Psychol. 51, 135–163.
Chomsky, N., 1956. Three models for the description of language. IRE Trans. Inf. Theor. 2, 113–124.
Chomsky, N., 1957. Syntactic Structures. The Hague, Mouton.
Chomsky, N., 1955. The Logical Structure of Linguistic Theory Ms. Harvard University and MIT. Revised version published by Plenum, New York, 1975.
Chomsky, N., 1958. [Review of belevitch 1956]. Language 34, 99–105.
Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.
Chomsky, N., 1975. Reflections on Language. Pantheon, New York.
Chomsky, N., 1981. Lectures in Government and Binding. Foris, Dordrecht.
Chomsky, N., 1995. The Minimalist Program. MIT Press, Cambridge, MA.
Chomsky, N., 2001. Beyond Explanatory Adequacy MIT Working Papers in Linguistics. MIT Press, Cambridge, MA.
Chomsky, N., 2005. Three factors in language design. Ling. Inq. 36, 1–22.
Chomsky, N., 2013. Problems of projection. Lingua 130, 33–49.
Chomsky, N., Halle, M., 1968. The Sound Pattern of English. MIT Press, Cambridge, MA.
Christophe, A., Nespor, M., Guasti, M.T., Van Ooyen, B., 2003. Prosodic structure and syntactic acquisition: the case of the head-direction parameter. Devl. Sci. 6, 211–220.
Cinque, G., 1999. Adverbs and Functional Heads: A Cross-linguistic Perspective. Oxford University Press, Oxford.
Clark, E.V., 1987. The principle of contrast: a constraint on language acquisition. In: MacWhinney, B. (Ed.), Mechanisms of Language Acquisition. Erlbaum, Hillsdale, NJ, pp. 1–33.
Collins, M., 1999. Head-driven Statistical Models for Natural Language Processing. PhD Thesis. University of Pennsylvania.
Crain, S., McKee, C., 1985. The acquisition of structural restrictions on anaphora. Proc. NELS 15, 94–110.
Crain, S., Nakayama, M., 1987. Structure dependence in grammar formation. Language 63, 522–543.
Crain, S., Thornton, R., 2011. Syntax acquisition. WIREs Cogn. Sci, http://dx.doi.org/10.1002/wcs.1158.
Crain, S., 2012. The Emergence of Meaning. Cambridge University Press, Cambridge.
Dehaene, S., 2011. The Number Sense: How the Mind Creates Mathematics. Oxford University Press, New York.
de Boysson-Bardies, B., Vihman, M.M., 1991. Adaptation to language: evidence from babbling and first words in four languages. Language 67, 297–319.
de Marchena, A., Eigsti, I.-M., Worek, A., Ono, K.E., Snedeker, J., 2011. Mutual exclusivity in autism spectrum disorders: testing the pragmatic hypothesis. Cognition 119, 96–113.
Diesendruck, G., Markson, L., 2001. Children's avoidance of lexical overlap. A pragmatic account. Dev. Psychol. 37, 630–641.
Doupe, A.J., Kuhl, P.K., 1999. Birdsong and human speech: common themes and mechanisms. Annu. Rev. Neurosci. 22, 567–631.
Eimas, P.D., Siqueland, E.R., Jusczyk, P., Vigorito, J., 1971. Speech perception in infants. Science 171, 303–306.
Elman, J.L., 1993. Learning and development in neural networks: the importance of starting small. Cognition 48, 71–99.
Everaert, M.B., Huybregts, M.A.C., Chomsky, N., Berwick, R.C., Bolhuis, J.J., 2015. Structures, not strings: linguistics as part of the cognitive sciences. Trends Cogn. Sci. 19, 729–743.
Fernald, A., 1985. Four-month-old infants prefer to listen to Motherese. Infant Behav. Dev. 8, 181–195.
Fiser, J., Aslin, R.N., 2001. Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychol. Sci. 12, 499–504.
Fitch, W., Martins, M.D., 2014. Hierarchical processing in music, language, and action. Lashley revisited. Ann. N.Y. Acad. Sci. 1316, 87–104.
Fodor, J.D., Sakas, W.G., 2005. The subset principle in syntax: costs of compliance. J. Ling. 41, 513–569.
Frank, M.C., Goodman, N.D., Tenenbaum, J.B., 2009. Using speakers' referential intentions to model early cross-situational word learning. Psychol. Sci. 20, 578–585.
Gagliardi, A., Lidz, J., 2014. Statistical insensitivity in the acquisition of Tsez noun classes. Language 90, 58–89.
Gelman, R., Butterworth, B., 2005. Number and language: how are they related? Trends Cogn. Sci. 9, 6–10.
Gentner, T.Q., Hulse, S.H., 1998. Perceptual mechanisms for individual vocal recognition in european starlings, sturnus vulgaris. Anim. Behav. 56, 579–594.
Gibson, E., Wexler, K., 1994. Triggers. Ling. Inq. 25, 407–454.
Gold, E.M., 1967. Language identification in the limit. Inf. Contr. 10, 447–474.
Goldin-Meadow, S., 2005. The Resilience of Language: What Gesture Creation in Deaf Children Can Tell Us About How All Children Learn Language. Psychology Press, New York.
Gordon, P., 2004. Numerical cognition without words: evidence from amazonia. Science 306, 496–499.
Halberda, J., 2003. The development of a word-learning strategy. Cognition 87, B23–B34.
Hale, K., 1976. The adjoined relative clause in Australia. In: Dixon, R.M.W. (Ed.), Grammatical Categories in Australian Languages. Australian National University, Canberra, pp. 78–105.
Halle, M., Vergnaud, J.-R., 1987. An Essay on Stress. MIT Press, Cambridge, MA.
Hamburger, H., Crain, S., 1984. Acquisition of cognitive compiling. Cognition 17, 85–136.
Harris, Z.S., 1955. From phoneme to morpheme. Language 31, 190–222.
Hart, B., Risley, T.R., 1995. Meaningful Differences in the Everyday Experience of Young American Children. Paul H Brookes Publishing, Baltimore, MD.
Hauser, M.D., Chomsky, N., Fitch, W.T., 2002. The faculty of language: what is it, who has it, and how did it evolve? Science 298, 1569–1579.
Heath, S.B., 1983. Ways with words: language. In: Life and Work in Communities and Classrooms. Cambridge University Press, Cambridge.
Henshilwood, C., d'Errico, F., Yates, R., Jacobs, Z., Tribolo, C., et al., 2002. Emergence of modern human behavior: middle stone age engravings from South Africa. Science 295, 1278–1280.
Hosino, T., Okanoya, K., 2000. Lesion of a higher-order song nucleus disrupts phrase level complexity in bengalese finches. Neuroreport 11, 2091–2095.
Hurford, J.R., 1975. The Linguistic Theory of Numerals. Cambridge University Press, Cambridge.
Hurford, J.R., 2011. The Origins of Grammar: Language in the Light of Evolution, vol 2. Oxford University Press, Oxford.
Huybregts, M.A.C., 1976. Overlapping dependencies in dutch. Utrecht working papers in linguistics. 1, 24–65.
Hyams, N., 1986. Language Acquisition and the Theory of Parameters. Reidel, Dordrecht.
Hyams, N., 1991. A reanalysis of null subjects in child language. In: Weissenborn, J., Goodluck, H., Roeper, T. (Eds.), Theoretical Issues in Language Acquisition: Continuity and Change in Development. Psychology Press, pp. 249–268.
Jelinek, F., 1998. Statistical Methods for Speech Recognition. MIT Press, Cambridge, MA.
Johnson, E.K., Jusczyk, P.W., 2001. Word segmentation by 8-month-olds: when speech cues count more than statistics. J. Mem. Lang. 44, 493–666.

Joshi, A.K., Shanker, K.V., Weir, D., 1991. The convergence of mildly context-sensitive grammar formalisms. In: Sellers, P.H., Shieber, S., Wasow, T. (Eds.), Foundational Issues in Natural Language Processing. MIT Press, Cambridge, MA, pp. 31–81.
Kahn, D., 1976. Syllable-based Generalizations in English Phonology PhD Thesis, MIT. Garland, New York, pp. 1980.
Kaminski, J., Call, J., Fischer, J., 2004. Word learning in a domestic dog: evidence for fast mapping. Science 304, 1682–1683.
Kazanina, N., Phillips, C., 2001. Coreference in child Russian: distinguishing syntactic and discourse constraints. In: Do, A.H., Domínguez, L., Johansen, A. (Eds.), Proceedings of the 25th Annual Boston University Conference on Language Development. Cascadilla Press, Somerville, MA, pp. 413–424.
Kegl, J., Senghas, A., Coppola, M., 1999. Creation through contact: sign language emergence and sign language change in Nicaragua. In: DeGraff, M. (Ed.), Language Creation and Language Change: Creolization, Diachrony, and Development. MIT Press, Cambridge, MA, pp. 179–237.
Kŭcera, H., Francis, W.N., 1967. Computational Analysis of Present-day American English. Brown University Press, Providence.
Kuhl, P.K., Miller, J.D., 1975. Speech perception by the chinchilla: voiced voiceless distinction in alveolar plosive consonants. Science 190, 69–72.
Labov, W., 2010. Principles of linguistic change, cognitive and cultural factors. Wiley-Blackwell, Malden.
Landau, B., Gleitman, L.R., 1984. Language and Experience: Evidence from the Blind Child, vol 8. Harvard University Press, Cambridge, MA.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Legate, J.A., 2002. Warlpiri: Theoretical Implications. PhD Thesis. Massachusetts Institute of Technology, Cambridge, MA.
Legate, J.A., Yang, C., 2002. Empirical re-assessment of stimulus poverty arguments. Ling. Rev. 19, 151–162.
Levin, B., 1993. English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago.
Levin, B., 2008. Dative verbs: a crosslinguistic perspective. Lingv. Investig. 31, 285–312.
Liberman, M., 1975. The Intonational System of English. PhD Thesis. MIT.
Locke, J.L., 1995. The Child's Path to Spoken Language. Harvard University Press.
Mandelbrot, B., 1953. An informational theory of the statistical structure of language. In: Jackson, W. (Ed.), Communication Theory, vol 84. Betterworth, pp. 486–502.
Marcus, G., Pinker, S., Ullman, M.T., Hollander, M., Rosen, J., Xu, F., 1992. Overregularizationn Language Acquisition Monographs of the Society for Research in Child Development. University of Chicago Press, Chicago.
Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., Taylor, A., 1999. Treebank-3. Linguistic Data Consortium: LDC99T42.
Marcus, G.F., 1993. Negative evidence in language acquisition. Cognition 46, 53–85.
Markman, E.M., Wachtel, G.F., 1988. Children's use of mutual exclusivity to constrain the meanings of words. Cogn. Psychol. 20, 121–157.
Markman, E.M., Wasow, J.L., Hansen, M.B., 2003. Use of the mutual exclusivity assumption by young word learners. Cogn. Psychol. 47, 241–275.
Marler, P., 1991. The instinct to learn. In: Carey, S., Gelman, R. (Eds.), The Epigenesis of Mind: Essays on Biology and Cognition. Psychology Press, New York, pp. 37–66.
McClelland, J.L., Patterson, K., 2002. Rules or connections in past-tense inflections: what does the evidence rule out? Trends Cogn. Sci. 6, 465–472.
McDougall, I., Brown, F.H., Fleagle, J.G., 2005. Stratigraphic placement and age of modern humans from Kibish. Ethiopia Nat. 433, 733–736.
Medin, D.L., Wattenmaker, W.D., Hampson, S.E., 1987. Family resemblance, conceptual cohesiveness, and category construction. Cogn. Psychol. 19, 242–279.
Medina, T.N., Snedeker, J., Trueswell, J.C., Gleitman, L.R., 2011. How words can and cannot be learned by observation. Proc. Natl. Acad. Sci. U. S. A. 108, 9014–9019.
Michaelis, J., 1998. Derivational minimalism is mildly context–sensitive. In: International Conference on Logical Aspects of Computational Linguistics. Springer, pp. 179–198.
Miller, G.A., 1957. Some effects of intermittent silence. Am. J. Psychol. 70, 311–314.
Miller, G.A., Chomsky, N., 1963. Finitary models of language users. In: Luce, R.D., Bush, R.R., Galanter, E. (Eds.), Handbook of Mathematical Psychology, vol II. Wiley, New York, pp. 419–491.
Nespor, M., Vogel, I., 1986. Prosodic Phonology. Foris, Dordrecht.
Nevins, A., Pesetsky, D., Rodrigues, C., 2009. Pirahã exceptionality: a reassessment. Lang 85, 355–404.
Newport, E., 1990. Maturational constraints on language learning. Cogn. Sci. 14, 11–28.
Newport, E., Gleitman, H., Gleitman, L., 1977. Mother, I'd rather do it myself: some effects and non-effects of maternal speech style. In: Ferguson, C.A., Snow, C.E. (Eds.), Talking to Children: Languge Input and Acquisition. Cambridge University Press, Cambridge, pp. 109–149.
Niyogi, P., 2006. The Computational Nature of Language Learning and Evolution. MIT Press, Cambridge, MA.
Osherson, D.N., Stob, M., Weinstein, S., 1986. Systems That Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists. MIT Press, Cambridge, MA.
Pepperberg, I.M., 2007. Grey parrots do not always 'parrot': the roles of imitation and phonological awareness in the creation of new labels from existing vocalizations. Lang. Sci. 29, 1–13.
Pepperberg, I.M., 2009. The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots. Harvard University Press, Cambridge, MA.
Perfors, A., Tenenbaum, J.B., Regier, T., 2011. The learnability of abstract syntactic principles. Cognition 118, 306–338.
Petitto, L.A., Marentette, P.F., 1991. Babbling in the manual mode: evidence for the ontogeny of language. Science 251, 1493–1496.
Pica, P., Lemer, C., Izard, V., Dehaene, S., 2004. Exact and approximate arithmetic in an amazonian indigene group. Science 306, 499–503.
Pierce, A., 1992. Language Acquisition and Syntactic Theory: A Comparative Analysis of French and English. Dordrecht, Kluwer.
Pine, J.M., Lieven, E.V., 1997. Slot and frame patterns and the development of the determiner category. Appl. Psychol. 18, 123–138.
Pinker, S., Ullman, M.T., 2002. The past and future of the past tense. Trends Cogn. Sci. 6, 456–463.
Pinker, S., 1989. Learnability and Cognition: The Acquisition of Argument Structure. MIT Press, Cambridge, MA.
Rissanen, J., 1978. Modeling by shortest data description. Automatica 14, 465–471.
Saffran, J.R., Aslin, R.N., Newport, E., 1996. Statistical learning by 8-month-old infants. Science 274, 1926–1928.
Saffran, J.R., Johnson, E.K., Aslin, R.N., Newport, E.L., 1999. Statistical learning of tone sequences by human infants and adults. Cognition 70, 27–52.
Sakas, W.G., Fodor, J.D., 2012. Disambiguating syntactic triggers. Lang. Acq. 19, 83–143.
Sandler, W., Meir, I., Padden, C., Aronoff, M., 2005. The emergence of grammar: systematic structure in a new language. Proc. Natl. Acad. Sci. U. S. A. 102, 2661–2665.
Sapir, E., 1928. Language: An Introduction to the Study of Speech. Harcourt Brace, New York.
Schieffelin, B.B., Ochs, E., 1986. Language Socialization Across Cultures. Cambridge University Press, Cambridge.
Schlenker, P., Chemla, E., Schel, A.M., Fuller, J., Gautier, J.-P., Kuhn, J., Veselinoviíc, D., Arnold, K., Cäsar, C., Keenan, S., et al., 2016. Formal monkey linguistics. Theor. Ling. 42, 1–90.
Schuler, K., Yang, C., Newport, E., 2016. Testing the Tolerance Principle: children form productive rules when it is more computationally efficient to do so. In: The 38th Cognitive Society Annual Meeting, Philadelphia, PA.
Shieber, S., 1985. Evidence against the context-freeness of natural language. Ling. Philos. 8, 333–343.
Shukla, M., White, K.S., Aslin, R.N., 2011. Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. Proc. Natl. Acad. Sci. U. S. A. 108, 6038–6043.
Slabakova, R., 2016. Second Language Acquisition. Oxford University Press, Oxford.
Stabler, E., 1996. Derivational Minimalism. In International Conference on Logical Aspects of Computational Linguistics. Springer, pp. 68–75.
Stevens, J., Trueswell, J., Yang, C., Gleitman, L., 2016. The pursuit of word meanings. Cogn. Sci. (in press).
Stockall, L., Marantz, A., 2006. A single route, full decomposition model of morphological complexity : MEG evidence. Mental Lexic. 1, 85–123.
Studdert-Kennedy, M., 1998. The particulate origins of language generativity: from syllable to gesture. In: Hurford, J., Studdert-Kennedy, M., Knight, C. (Eds.), Approaches to the Evolution of Language. Cambridge University Press, Cambridge, pp. 202–221.
Terrace, H.S., Petitto, L.-A., Sanders, R.J., Bever, T.G., 1979. Can an ape create a sentence? Science 206, 891–902.
Terrace, H.S., 1987. Nim: A Chimpanzee Who Learned Sign Language. Columbia University Press.
Tesar, B., Smolensky, P., 1998. Learnability in optimality theory. Ling. Inq. 29, 229–268.
Thornton, R., Crain, S., 2013. Parameters: the pluses and the minuses. In: den Dikken, M. (Ed.), The Cambridge Handbook of Generative Syntax. Cambridge University Press, Cambridge, pp. 927–970.
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. Cogn. Ling. 11, 61–82.
Tomasello, M., 2003. Constructing a Language. Harvard University Press, Cambridge, MA.
Trueswell, J.C., Medina, T.N., Hafri, A., Gleitman, L.R., 2013. Propose but verify: fast mapping meets cross-situational word learning. Cogn. Psychol. 66, 126–156.
Turk-Browne, N.B., Jungé, J.A., Scholl, B.J., 2005. The automaticity of visual statistical learning. J. Exp. Psychol: Gen. 134, 552.
Valian, V., 1986. Syntactic categories in the speech of young children. Dev. Psychol. 22, 562.
Valian, V., 1991. Syntactic subjects in the early speech of American and Italian children. Cognition 40, 21–81.
Valiant, L.G., 1984. A theory of the learnable. Commun. ACM 27, 1134–1142.
Valian, V., Solt, S., Stewart, J., 2009. Abstract categories or limited-scope formulae? The case of children's determiners. J. Child Lang. 36, 743–778.
Vanhaeren, M., D'Errico, F., Stringer, C., James, S.L., Todd, J.A., 2006. Middle paleolithic shell beads in Israel and Algeria. Science 312, 1785–1788.
Vapnik, V., 2000. The Nature of Statistical Learning Theory. Springer, Berlin.
Vihman, M.M., 2013. Phonological Development: The First Two Years. John Wiley & Sons.
Wang, Q., Lillo-Martin, D., Best, C.T., Levitt, A., 1992. Null subject versus null object: some evidence from the acquisition of Chinese and English. Lang. Acq. 2, 221–254.
Werker, J.F., Tees, R.C., 1984. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. Inf. Behav. Dev. 7, 49–63.

Wexler, K., Culicover, P., 1980. Formal Principles of Language Acquisition. MIT Press, Cambridge, MA.
Wexler, K., 1998. Very early parameter setting and the unique checking constraint: a new explanation of the optional infinitive stage. Lingua 106, 23–79.
White, L., 1990. The verb-movement parameter in second language acquisition. Lang. Acq. 1, 337–360.
White, L., 2003. Second Language Acquisition and Universal Grammar. Cambridge University Press, Cambridge.
Whitman, J., 1987. Configurationality parameters. In: Takashi, I., Saito, M. (Eds.), Japanese linguistics. Foris, Dordrecht, pp. 351–374.
Xu, F., Pinker, S., 1995. Weird past tense forms. J. Child Lang. 22, 531–556.
Yang, C., 2002. Knowledge and Learning in Natural Language. Oxford University Press, Oxford.
Yang, C., 2004. Unversal grammar, statistics or both? Trends Cogn. Sci. 8, 451–456.
Yang, C., 2012. Computational models of syntactic acquisition. Wiley Interdiscipl. Rev Cogn. Sci. 3, 205–213.
Yang, C., 2013. Ontogeny and phylogeny of language. Proc. Natl. Acad. Sci. 110, 6324–6327.
Yang, C., 2015. Negative knowledge from positive evidence. Language 91, 938–953.
Yang, C., 2016. The Price of Linguistic Productivity: How Children Learn to Break Rules of Language. MIT Press, Cambridge, MA.
Yeung, H.H., Werker, J.F., 2009. Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. Cognition 113, 234–243.
Yu, C., Smith, L.B., 2007. Rapid word learning under uncertainty via cross-situational statistics. Psychol. Sci. 18, 414–420.