SECTION A: COGNITIVE PERSPECTIVES ON FORMULAIC LANGUAGE

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

Annual Review of Applied Linguistics (2012), **32**, 3–16.
© Cambridge University Press, 2012, 0267-1905/12 $16.00
doi: 10.1017/S0267190512000062

Formulaic Language in L1 Acquisition

Colin Bannard and Elena Lieven

The recognition that speech formulas play a role in first language acquisition—that children reuse sequences of words taken directly and seemingly unanalyzed from the input—goes back to the earliest days of the field. Until fairly recently, however, such formulaic language was considered part of an early and soon-superseded stage of development. The last decade has seen the rise of a perspective on language development in which such formulas are central to language acquisition across development. According to this perspective, which is often known as the *usage-based theory of language development*, acquisition begins when children identify, infer a communicative function for, and start to utilize pieces of language of different sizes (single words and multiword sequences). Generalization, and as a result grammar, is an emergent property resulting from the ongoing coexistence of such sequences in a shared representational space. The growth in popularity of such an account, which represents a radical break from traditional models of grammatical development, has resulted in large part from the appearance of very large corpora of child–caregiver interactions. Such corpora have supported a new understanding of the challenges and opportunities facing the learner, as well as allowing new naturalistic analyses of children's productions and the creation of stimuli for experiments, all of which offer considerable support for the usage-based position. This article offers a review of these developments.

Children learn language from the speech of others. Many accounts of language development assume that children proceed by (a) acquiring words, (b) classifying these words as being of particular categories (which may be innately specified, or simply arise from distributional patterns in the input), (c) inferring abstract rules for combining words based on their categories, and then (d) learning to produce and understand language by recombining or analyzing novel word sequences on the basis of these categories and rules. The research presented in this article comes mostly from an ever-growing tradition that gives social learning a much greater role in language development. Rather than breaking others' utterances down and building their own representations brick by brick (inferring abstract categories and rules, and then regenerating language), children construct a language on rich scaffolding provided by adult speakers.

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

The process of grammatical development begins when children identify, infer a communicative function for, and start to utilize pieces of language. Rather than identifying only words in the input (a task that despite the advances in artificial-language-based research, remains, for an arbitrary sample of real language, a process for which no fully reliable mechanistic account exists), children acquire whatever pieces of language are of use to them (either in achieving instrumental goals or taking part in the recurrent routines of human social life). The nature of the pieces of language they detect will of course depend largely on what word sequences occur and reoccur in the language they hear, and their size will thus depend on the typology of the language they are hearing.

As these pieces of language (which are form-meaning mappings) become established, they take their place in a developing network of other pieces, with the network being organized by the overlap in form or in meaning of the network content. The regularities of grammar arise as an emergent property of this network. The coexistence in this shared representation space of fragments of language with overlapping phonological and semantic content gives rise to the creation of schematic patterns (often conceptualized as slot-and-frame patterns). The points of variability in overlapping word sequences, filled with material that is often similar in function, gives rise to what are usually observed to be lexical and phrasal categories. As the child gets older, the schematic patterns (equivalent in many respects to the argument structures of traditional grammatical description) that arise, and the connections between them (which are the basis of the categories of traditional linguistic analysis) become more and more entrenched, and are eventually made available to conscious reflection and manipulation (through a process often referred to as representational redescription; see Karmiloff-Smith, 1995; Tomasello, 1999). Nonetheless there is much evidence, both from the study of child language acquisition (which we will review in depth here) and adult processing (which we will not), that the original representations that are the basis of the grammar that emerges do not completely atrophy and remain a vital part of a person's knowledge of grammar across the life span, giving rise to the formulaicity that is so characteristic of language use and is the topic of this volume.

This article will review the evidence for this account. In the first section we look at the new availability of data that have so boosted the plausibility of an account of development in which language reuse plays such a central part, and we offer a snapshot of the linguistic input children hear. In the next two sections, we provide evidence that children have dedicated representations for word sequences that they frequently encounter and that these sequences form the basis of their developing productive grammars. Next, we explore the role that such sequences play in the development of complex syntax. We look at the role that such sequences play in the systematic errors that children make and in the development of some specific construction types—questions and sentential complements. Our focus in the first four sections is on the learning of English. In the final section we turn to the current research and the outstanding questions concerning the role of language reuse in the acquisition of other languages.

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

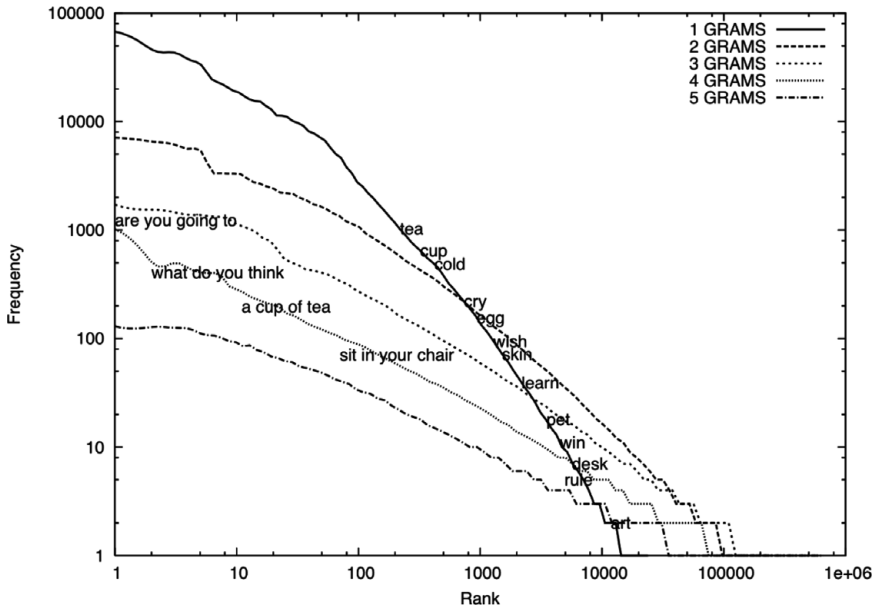


Fig. 1. A rank-frequency plot of the words and word sequences in 1.9 million words of one mother's speech. From Bannard and Matthews (2008). Reprinted with permission.

WHAT CHILDREN HEAR AND WHAT IT MEANS FOR LANGUAGE DEVELOPMENT

An extremely important development in the study of language acquisition in recent years has been the collection and distribution of data sets of children's interactions with their caregivers. While the collecting of such data has been underway since at least the 1960s, samples were often very limited in size and sometimes spread thinly over extended periods. The last 5–10 years have seen a massive increase in the size of the data sets available. This has had many important effects on the field. One has been that for the first time we have enough data to start really understanding the nature of the language children hear. This new availability of data has confirmed that the language children hear is (like all natural language use) highly repetitive. See, for instance, the study by Cameron-Faulkner, Lieven, and Tomasello (2003) showing that in the child-directed speech of 12 English-speaking mothers to their two-year-olds, 51 percent of all maternal utterances began with one of 52 item-based phrases, mostly consisting of two words or morphemes (45 percent began with one of just 17 words). The extent of this is perhaps best shown by a picture of the input. Figure 1 plots the frequency of words and sequences of words (up to length five) in 1.9 million words of one mother's speech, plotted against their rank. It can be seen that the most frequent word (the leftmost point of the solid line) occurs around 70,000 times. If we follow that line to the right, we can see that the 100th most frequent word occurs around 10,000 times. Similar lines are

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

also included for sequences of length two to five. The most frequent four-word sequence occurs around 1,000 times, and the 100th most frequent four-word sequence occurs around 100 times (for further details of this kind of plot, see, e.g., Bannard & Lieven, 2009). What can be seen here is that while the most frequent word (1-gram) is much more frequent than the most frequent unit of any other length, there are many sequences of more than one word that are more frequent than words that we might consider a part of a child's core vocabulary.

What then are the consequences of this for language acquisition? What it means is that there a great many multiword sequences that the child repeatedly encounters. A popular account of how children segment words out of the speech stream says that they do so by tracking the probability with which sounds follow one another, and form segments by identifying high probability combinations (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996). The frequency of many of these word combinations means that applying such an approach to the speech stream is going to result in the extraction of sequences of longer than one word.

The other thing that this plot tells us is that, rather than simply being an unfortunate side product of a word-identification procedure, the extraction of multiword sequences from the input and the formation and retention of dedicated representations is actually an efficient path to take. Given that the child will want eventually to identify words, this might seem counterintuitive to linguists who are used to pursuing minimal and nonredundant descriptions. That this kind of redundancy is optimal can be shown mathematically via coding theory, and Bannard and Lieven (2009) provided a more extended argument than we will here. However, it will hopefully suffice to point out a contemporary real world example—the establishment of the conventional acronyms that are so widely deployed in modern electronic communication. In order to be able to send and unpack the meaning of such strings as LOL or TMI, English speakers have to form dedicated representations for the phrases and a mapping from them to the acronyms. This might seem redundant, as we already have shared knowledge of the words and the capacity to produce them, but the efficiency that they allow in the production and reception of messages justifies the formation of such representations, and hence such acronyms continue to proliferate. Similarly for a child constructing representations that will support the efficient processing and production of language by an immature system, it should be clear that having a prebuilt representation for those sequences that will be frequently required or encountered will be well worth the representational cost.

EVIDENCE FOR THE ENTRENCHMENT OF MULTIWORD STRINGS IN CHILDREN'S LANGUAGE

In this section we look not at the language that children hear but at the language they produce. We have described two possible models of language acquisition. According to one model, children infer rules and categories from the input and then use these to generate new strings. According to the other, usage-based, model, children use sequences of words taken directly from the input, with

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

productivity arising from the coexistence in a shared representational space of forms that overlap in phonological content or in meaning. One way to compare these models is to ask which model is most commensurate with children's productions. A number of studies have done this by utilizing very densely sampled corpora that document children's interactions with others in their home for five hours a week over six-week periods. One approach has been the so-called traceback method. Traceback proceeds by taking a child's unique multiword utterances in the final session or sessions of the six-week period and then evaluating how well this can be accounted for in terms of stored lexical sequences, by searching the data from the preceding weeks for overlapping word sequences (Dąbrowska & Lieven, 2005; Lieven, Behrens, Speares, & Tomasello, 2003; Lieven, Salomo, & Tomasello, 2009). In one recent example of this kind of work, Lieven et al. (2009) looked at the speech of four two-year-old children. There were 30 hours of recordings for each child. The traceback was performed by taking the final two hours of recordings for each and searching for related utterances in the preceding 28 hours of recordings. They found that 20–40 percent of children's utterances were word sequences that the children had used before, and 40–50 percent of their utterances were identical to a previous utterance except for one single point of variation, which could be produced by a single productive operation of inserting material into a prefabricated frame to produce a novel utterance. Building on this, the study also looked at the productive process that might support the creation of these novel utterances, by classifying the kinds of words or word sequences that seemed to have been inserted into the frames. In most cases the inserted material referred to people, places, or objects. As already stated, one claim of the usage-based model is that the formation of the slot-and-frame patterns leads to the creation of lexical and phrasal categories, and it seems that in the two-year-old children studied, the network of stored strings and the resulting development of frames is giving rise to a noun category. The children with a higher mean length of utterance, produced more of the apparently productive utterances and, also showed greater variety in the kinds of elements that were included in the slots.

In addition to enabling corpus analyses, the availability of large samples of the input has allowed researchers to conduct lab-based experiments in which the frequency of word sequences in such databases has been used as an estimate of participants' exposure to the strings of words. In one study, Bannard and Matthews (2008) asked two- and three-year-old children to repeat back either frequent sequences of words (e.g., *a drink of milk*), or less frequent sequences of words that were identical except for the last word (e.g., *a drink of tea*), with the last word itself and the final bigram having the same frequency in both phrase types. The children made fewer errors and took less time to produce the matched part of the string (e.g., *a drink of*) in the frequent than in the infrequent sequence. It seems, then, that children have dedicated representations for sequences that recur frequently in their input. In a later study, Matthews and Bannard (2010) looked at the input again in order to explore when children would generalize over lexical patterns. By collapsing over the kind of four-word sequences deployed in their earlier article, they extracted frequent patterns of the form WORD1 WORD2 WORD3 SLOT, such as *a piece of X*. They then tested the children's

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

ability to repeat back novel instantiations of these patterns (e.g., *a piece of toast*, *a piece of meat*, *a piece of brick*). They found that children were more fluent in their repetition of such sequences when the sequences were observed with very high statistical variability in the linguistic types (operationalized using the information-theoretic measure of entropy over the distribution of items) seen in the slot position and when there was also high semantic coherence to the items seen in that position (when they were maximally similar to one another). It seems, then, that children's generalization over the input arises from the overlap in the lexical content and the meaning of the strings in their input. The finding that overlap leads to generalization like this is consistent with the artificial grammar learning findings of Onnis, Waterfall, and Edelman (2008).

BUILDING GRAMMARS OUT OF LEXICALLY SPECIFIC MULTIWORD STRINGS

The work described here extends the corpus analysis work outlined earlier in this article. The goal is to look at whether children's productions can be accounted for with lexically specific grammars. This work differs from traceback in that it reverses the mode of analysis: Whereas the traceback studies described here took children's utterances and asked whether they could be accounted for using the formulas and repetitions found in the children's prior speech, in the study described in this section grammatical representations were inferred from the prior utterances, and the value of grammar was then tested by looking at how well it was able to predict later utterances. Bannard, Lieven, and Tomasello (2009) performed a computational analysis of the transcribed speech of two children, Brian and Annie, for six weeks immediately following their second and third birthdays. (Brian's entire corpus is now deposited on the CHILDES database [MacWhinney, 2000] under the name "Thomas": http://childes.psy.cmu.edu/manuals/.) The researchers first set aside the transcriptions from the last two weeks of speech from each of the corpora, then used Bayesian grammar induction to infer grammars from each of the remaining data sets. All production rules in the grammars had some lexically specific content, with categories being distributionally similar classes of items that were discovered from the data by the inference process. These grammars were then used to parse the set-aside samples of data. For the age two data, it was found that the fully lexically specific grammars covered 84 percent of the utterances for Brian and 75 percent of the utterances for Annie; at age three, they covered 70 percent and 81 percent, respectively.

These grammars thus seemed to account very well for the children's speech. At age two, nearly 60 percent of Brian's utterances were accounted for by analyses in which only one combinatorial operation (the insertion of material in a slot) was performed; by age three nearly 50 percent of his utterances were accounted for by analyses including only two operations. For Annie at age two, 60 percent of utterances were accounted for with only two operations, and by age three her language was far more productive, with five operations being necessary to account for the same coverage. This suggests that at age two the

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

children are relying on substantially formulaic grammars, but by age three they are far more productive. This impression is also supported by the finding that each child's grammar did a bad job of accounting for the other's data at age two, but a much better job at age three. In a subsequent step, words were replaced with human-annotated categories (first nouns and then verbs), and the selected grammars were used to parse test sentences in which the words had similarly been replaced. Nouns were found to significantly improve coverage at age two for both children, while the verb category offered little improvement in coverage for Brian. At age three the addition of the verb category had an impact for both children, indicating that linguistic categories are emergent rather than pregiven.

FROM SIMPLE TO MORE COMPLEX SYNTAX

The account of grammatical development we are presenting here is one in which children begin by taking sequences of words directly from the input, with productivity emerging from the overlap (phonological and semantic) between these sequences. While this knowledge becomes more available to awareness and manipulation (through a kind of representational redescription) as children develop, the sequences continue to play a central role in children's productions, even as they develop more complex syntax. Directly tracing the representations that underlie the children's productions becomes a far less tractable task as their knowledge develops. However, the existence of lexically specific representations at later stages can be inferred by other means. One way in which such representations clearly show through in older children's speech is in the errors that they make. Heavily entrenched word sequences compete with the emerging generalizations in production, and this leads to the production of errors. In the next section, we review some of the main error types children make and describe how they can be explained as arising from the activation of formulaic sequences.

Inflectional Errors

One aspect of grammar that children need to master is the case system. The very limited use of case marking in English means that the task facing English learners is comparatively straightforward, and yet English-speaking children still often make striking errors in using *me* as a pronoun in subject position, in place of the correct nominative form *I*, in even very simple sentences. One explanation for the pervasiveness of this phenomenon is that it results from children's extraction of stored sequences from more complex sentence forms. In a study of 17 children, Kirjavainen, Theakston, and Lieven (2009) found that the number of me-for-I errors a child made correlated with the number of forms in their caregiver's speech in which *me* was followed by a nonfinite verb form (e.g., *let me do that*, *did you hear me singing*?). Furthermore, the children were found to be more likely to make me-for-I errors for sentences containing verbs that were used in such contexts by their caregivers.

A second pervasive inflectional error that appears to result from the child's erroneous deployment of a stored form is children's use of nonfinite forms when a finite form is obligatory—so-called optional infinitive errors (e.g., saying *Mummy*

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

go to work instead of *Mummy goes to work*). Such errors are more common in some languages than others, and a number of accounts have been proposed to explain this. One very compelling explanation was provided by Freudenthal, Pine, and Gobet (2010), who in an analysis of Dutch, English, French, German, and Spanish, found that the rate of optional infinitive errors in children's speech varied depending on the frequency with which nonfinite verb forms occurred at the end of utterances in caregivers' speech. Based on the independent observation that word sequences that are utterance-final have a greater tendency to become entrenched in children's speech, it seems highly plausible that the observed differences between languages is the result of the activation of stored sequences containing the nonfinite forms. This account was further supported by the fact that optional infinitive errors were more likely to occur with verbs that occurred in nonfinite forms in complex verb phrases in the input.

Inversion in English questions

The formation of questions is an aspect of grammar for which different grammatical formalisms take radically different approaches. Take, for example, a question like *What does Daddy want?* The WH word (here, *what*) indicates the thing that is being inquired about. A great deal of work has gone into accounting for how this so-called missing element is represented by speakers such that it affects their grammatical decisions. In a simple English declarative sentence this element would come at the end of the sentence (e.g., *Daddy wants a biscuit*), and early transformational grammar (see Radford, 1988) proposed that questions are derived from a deep structure in which the WH element occurs in this position and is moved to the beginning of the sentence when deriving the surface form. While the notion of deep structure has been dropped from later theories in this tradition, the notion of movement has remained. These movement rules are assumed to be specified in universal grammar.

In usage-based models of grammar development, by contrast, children acquire knowledge of questions in the same way that they acquire knowledge of other constructions—they start by mapping frequently encountered sequences of words to functions, and productivity emerges via the phonological and semantic overlap between these sequences. The identification of categories emerges via the detection of systematic points of variability in the sequences, giving rise eventually to more abstract question constructions (the way that such constructions are usually described seems to be compatible with the nonderivational generative analyses specified in, for example, Ginzburg & Sag, 2000; Sag, 2010).

Children frequently make errors known as noninversion errors, where they appear to put words in the declarative order and simply place the interrogative word at the front (e.g., *What Daddy is doing? What does she doesn't want?*). In nativist accounts of questions, errors occur as the result of mistaken parameter setting, whereas in usage-based accounts they occur because of the entrenchment of the declarative form and the absence of an entrenched form for the correct question form. Rowland and Pine (2000) looked the noninversion errors in the WH questions of one child in 55 hours of recordings between the ages

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

of two and almost five. They looked at the WH+AUX combinations seen in the language that the child heard in these recordings and found that the child was less likely to make an inversion error on those combinations that were frequently heard in the input, supporting the usage-based account that the errors are the result of the absence of the correct entrenched string in the child's developing linguistic representations.

This result was confirmed in a larger-scale corpus study by Rowland (2007), who looked at the errors of 10 children between the ages of two and five. For each of the four question types (yes-no questions or WH questions, both with either a modal verb or with *do*), about 60–80 percent of the error-free questions were produced with one of the three most frequent frames for that question type. Ambridge, Rowland, Theakston, and Tomasello (2006) provided experimental evidence that children have lexical-sequence-specific representations. They elicited different kinds of WH questions from three- and four-year-olds and found that the best predictor of the number of noninversion errors was not the WH word or the auxiliary alone, but rather the specific word combination.

Dąbrowska and Lieven (2005) performed a traceback analysis of two children's speech and proposed a grammar of children's early questions that was built around schematic sequences with productive slots into which material of a particular semantic class could be inserted (e.g., *What's a THING?* or *What can we PROCESS?*). Evidence that children construct their questions in such a manner was provided by Ambridge and Rowland (2009), who looked at auxiliary doubling errors in children's negative questions (e.g., *What does she doesn't want*). According to a usage-based account, such errors would be predicted to occur because of the erroneous activation of entrenched forms. Following the schema-based model, we might propose that children would form such utterances by combining the entrenched positive question schema "What does" and the entrenched declarative form "she doesn't want." Ambridge and Rowland elicited what, why, and yes-no questions from 50 children aged three–four in positive (e.g., *What does she want?*) and negative form (*What doesn't she want?*). Consistent with usage-based predictions, they found that the children's tendency to make auxiliary doubling errors in the negative question correlated with their ability to produce the positive form of the sentence.

Dąbrowska, Rowland, and Theakston (2009) turned their attention to questions in which the dependency between the filler and the gap spans two or more clausal boundaries (e.g., *What do you think is in the box?*). Responding to the observation that children seem able to correctly produce such complex questions quite early in development (which has been used to argue for the existence of innate knowledge), they proposed that such questions are produced using schematic patterns such as *WH do you think S-GAP*. They tested this in a repetition study where the children were required to reproduce two kinds of questions—those based on the prototypical schema just listed, such as *What do you think the boys will really like?*, and another nonprototypical set using a different verb, such as *What does the man really hope they will like*? Children were found to make more errors when repeating the nonprototypical questions, suggesting that they might be utilizing entrenched schematic patterns when producing the typical questions.

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

Complement-Clause Constructions

It seems, then, that children's production of even quite complex questions involves the use of formulaic, schematic patterns. Similar findings exist for other complex constructions. Diessel and Tomasello (2001) looked at the complement-clause constructions of seven English-learning children between ages one and five. They argued that early on most of the children's apparent uses of complement clauses are in fact simply a single proposition (contained in the complement), with the main verb being simply an "epistemic marker, attention getter, or marker of illocutionary force" (p. 97). Later on, they do find real complement clauses, but the vast majority of these occur with the same highly frequent verbs (*think*, *say*, *see*), suggesting that they are part of a formulaic frame and not generated by an abstract syntactic rule. Brandt, Lieven, & Tomasello (2010) reported similar findings for an analysis of a very dense corpus for a single German-speaking child across the same age range.

Children's early complement-clause constructions, then, appear to utilize entrenched schematic patterns. Experimental evidence for this was provided by Brandt, Verhagen, Lieven, and Tomasello (2011). They prompted four- and five-year-old German-speaking children to change the subject of main (matrix) verbs in complement-clause constructions. A toy dog would produce such a construction (e.g., *Ich glaube*, *dass Emma morgen ein Haus baut*; "I believe that Emma is building a house tomorrow"), and the experimenter would then ask the child what was up with the dog. The children were found to be more likely to produce the appropriately altered utterance when the matrix verb occurred less frequently in complement-clause constructions in a large corpus of caregiver speech. This suggests that the entrenchment of the frequent pattern may have inhibited the children's ability to be productive.

CROSS-LINGUISTIC CONSIDERATIONS

This article has reviewed evidence that language acquisition is driven to a very large extent by children's reuse of word sequences taken directly from the speech that they hear. One prediction that follows from this is that the course of development should be heavily dependent on the nature of that input. It is therefore significant that the nature of that input will vary hugely between languages. Most of the evidence that we have discussed in this article has concerned the acquisition of English. In what ways might we expect learning to be different in other languages?

As we have discussed throughout, a very important factor in English-speaking children's early language reuse is the sheer amount of repetition to be observed in the language that they hear. They are repeatedly exposed to the same strings of words in familiar social routines, allowing them to identify the form and the conventional function to which it is put. It might, however, be reasonable to question whether this situation would be found in all other languages. English has a fixed word order, while in other languages there is a much greater degree of flexibility in this regard. English also has a relatively impoverished morphology,

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

meaning that the forms of the words observed are subject to rather minimal variation. As a result, the formal repetitiveness of English might not extend to languages with freer word order and richer morphology.

Stoll, Abbot-Smith, and Lieven (2009) looked at this question by applying an adaptation of the methods of Cameron-Faulkner et al. (2003) to three different languages: English, German, and Russian. They looked specifically for recurrent frames (defined as a sequence that occurred four times or more in the corpus) at the beginning of utterances. The typological differences between the languages meant that they did observe some differences. The greater word-order flexibility of German meant that it had fewer frames than English; Russian, with its even freer word order as well as its absence of determiners and the absence of the present-tense copula form, had the fewest frames of all three languages. Nonetheless, at least 70 percent of utterances began with a frame in all three languages. Thus although there is greater diversity, the input still offers the opportunity for formulaicity in all three languages.

While talking about the formulaicity in children's speech might take attention away from productivity, the presence of invariant structure is a necessary precondition of generalization—without any repetition, language would be unlearnable. An important consequence of the lexically specific frames that we have posited is that they are the basis of children's induction of categories. However, the kind of frames that support this kind of generalization seem to differ cross-linguistically. The idea that lexical frames might be central to the induction of categories was explored by Mintz (2003). A frame for Mintz was any pair of words that recurred with a single word intervening, so two words, A and B, might form a frame A_X_B, where X is a slot into which a variety of words can be inserted. Looking over a corpus of English caregiver speech, Mintz identified the 45 most frequent frames and found that of these, the words observed in their slot were largely of the same category, such that a given frame could be used to predict the category of its intervening word with 90–98 percent accuracy.

Thus a particular kind of so-called frequent frame would seem to be a very useful cue to lexical categories in English. Comparable results have been found for French (Chemla, Mintz, Bernal, & Christophe, 2009), but less compelling results have been reported for Dutch (Erkelens, 2009) and German (Stumper, Bannard, Lieven, & Tomasello, 2011). While some frames in the German analysis were found to be useful for categorizing words, these seemed to be frames in which there was very limited variability in the intervening word. In order to find reasonable coverage, the child would need to consider frames that would lead to very low classification accuracy. It seems, then, that the kinds of frames that will emerge from a child's generalization process will necessarily vary greatly cross-linguistically.

Children's language learning therefore likely proceeds slightly differently in other languages. One possibility also acknowledged by Mintz (2003) is that the schematic patterns utilized might be based upon morphemes rather than words. This leads to the interesting prediction that the discovery of morphology might be more closely bound up with the discovery of syntax than is traditionally assumed. There is evidence from the study of Spanish verb inflection (Aguado-Orea, 2004) and Polish noun inflection (Krajewski, Lieven, & Theakston, 2012)

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

that the development of morphology proceeds in a similarly lexically specific and piecemeal manner to that of grammar. Existing literature thus indicates that the segmentation of the speech stream, and the discovery of morphology and grammar, might emerge from the same cognitive processes.

CONCLUSIONS

This article has reviewed a growing literature suggesting that formulaic speech is a core component of language acquisition. Far from being a peripheral phenomenon, or an alternative to productive grammar, children's direct reuse of the speech of others is the very basis of their generalizations. The sequences that children copy contain a latent structure that, when encoded in a shared representational space, leads to the emergence of grammatical knowledge.

Grammatical development begins when children identify and infer a holistic communicative function for sequences of words in the speech of others. By identifying such form-meaning mappings, they begin to build representations for phonologically and semantically overlapping sequences that become entrenched through repeated exposure and use. However, rather than being an alternative to generalization, as a critical mass of such formulaic phrases builds up, this phonological and semantic overlap gives rise to generalization. This generalization is first based on pieces of invariant structure combined with points of variation, giving rise to the low-scope generalizations that seem to underlie so much of children's early speech.

As development progresses, the original representations of speech formulas are redescribed and begin to look more like the rule-based system of adult grammar. Children end up much less constrained by the input. The clearest example of this is the range of errors that they make that are not the result of erroneous reuse of the input. However, there is good reason to believe that the sequences that are the starting point of the system are not completely driven out, and that the most heavily entrenched sequences persist in the range of formulaic phenomena discussed elsewhere in this volume.

ANNOTATED BIBLIOGRAPHY

Ambridge, B., & Lieven, E. (2011). *Child language acquisition: Contrasting theoretical approaches.* Cambridge, UK: Cambridge University Press.

This textbook provides an evidence-based review of the central issues in first-language acquisition research, including many of the examples discussed in this article.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241–248.

This article reports on a repetition experiment in which two- and three-year-old children were found to be better (fewer errors, shorter duration) at producing the

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

first three words of frequent four-word sequences (e.g., *a drink of milk*) than they were at producing the same three words when part of infrequent sequences (e.g., *a drink of tea*). It provides the first experimental evidence that young children have dedicated representations for frequent multiword sequences.

Matthews, D., & Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child directed speech. *Cognitive Science*, *34*, 465–488.

This article provides experimental evidence that two- and three-year-old children extract generalizations from the input when sequences of repeated words occur with points of significant variation (e.g., they extract the schematic phrase *a piece of X* when exposed to multiple sequences such as *a piece of toast*, *a piece of string*, etc.). The article also explores some distributional factors that appear to determine when children will identify such patterns.

Rowland, C. F. (2007). Explaining errors in children's questions. *Cognition*, *104*, 106–134.

This article reports on a corpus analysis of the questions produced by 10 children. When a very large proportion of the error-free questions were structured around highly frequent formulas, errors occurred when children were required to deviate from such patterns. The article concluded that children's early questions are produced by reusing word sequences taken directly from the input.

Stoll, S., Abbot-Smith, K., & Lieven, E. (2009). Lexically restricted utterances in Russian, German, and English child directed speech. *Cognitive Science*, *33*, 75–103.

This article describes an analysis of the word sequences children hear in three different languages and provides evidence that opportunities for formulaic learning abound in even languages with freer word order and richer morphology than English.

REFERENCES

Aguado-Orea, J. J. (2004). *The acquisition of morpho-syntax in Spanish: Implications for current theories of development* (Unpublished doctoral thesis). University of Nottingham, Nottingham, UK.

Ambridge, B., & Rowland, C. F. (2009). Predicting children's errors with negative questions: testing a schema-combination account. *Cognitive Linguistics*, *20*, 225–266.

Ambridge, B., Rowland, C. F., Theakston, A. L., & Tomasello, M. (2006). Comparing different accounts of inversion errors in children's non-subject wh-questions: "What experimental data can tell us?" *Journal of Child Language*, *33*, 519–557.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Bannard, C., & Lieven, E. (2009). Repetition and reuse in child language learning. In R. Corrigan, E. Moravcsik, H. Ouali, & K. Wheatley (Eds.), *Formulaic language: Vol. 2. Acquisition, loss, psychological reality, functional explanations* (pp. 297–321). Amsterdam, the Netherlands: John Benjamins.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*, 17284–17289.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241–248.

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press

Brandt, S., Lieven, E., & Tomasello, M. (2010). Development of word order in German complement-clause constructions: Effects of input frequencies, lexical items, and discourse function. *Language*, *86*, 583–610.
Brandt, S., Verhagen, A., Lieven, E., & Tomasello, M. (2011). German children's productivity with simple transitive and complement-clause constructions: Testing the effects of frequency and variability. *Cognitive Linguistics*, *22*, 325–357.
Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction-based analysis of child-directed speech. *Cognitive Science*, *27*, 843–873.
Chemla, E., Mintz, T. H., Bernal, S., & Christophe, A. (2009). Categorizing words using "frequent frames": What cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science*, *12*, 396–406.
Dąbrowska, E., & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics*, *16*, 437–474.
Dąbrowska, E., Rowland, C., & Theakston, A. (2009). The acquisition of questions with long-distance dependencies. *Cognitive Linguistics*, *20*, 571–597.
Diessel, H., & Tomasello, M. (2001). The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics*, *12*, 97–141.
Erkelens, M. A. (2009). *Learning to categorize verbs and nouns: Studies on Dutch*. Utrecht, the Netherlands: LOT.
Freudenthal, D., Pine, J. M., & Gobet, F. (2010). Explaining quantitative variation in the rate of Optional Infinitive errors across languages: A comparison of MOSAIC and the Variational Learning Model. *Journal of Child Language, 37*, 643–669.
Ginzburg, J., & Sag, I. A. (2000). *Interrogative investigations: The form, meaning, and use of English interrogatives*. Stanford, CA: CSLI.
Karmiloff-Smith, A. (1995). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press/Bradford Books.
Kirjavainen, M., Theakston, A., & Lieven, E. (2009). Can input explain children's me-for-I errors? *Journal of Child Language*, *36*, 1091–1114.
Krajewski, G., Lieven, E., & Theakston, A. (2012). Productivity of a Polish child's inflectional noun morphology: a naturalistic study. *Morphology, 22*, 1, 9–34.
Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, *30*, 333–370.
Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*, 481–508.
MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk: Vol. 2. The database* (3rd ed.). Mahwah, NJ: Erlbaum.
Matthews, D., & Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child directed speech. *Cognitive Science*, *34*, 465–488.
Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.
Onnis, L., Waterfall, H., & Edelman, S. (2008). Learn locally, act globally: Learning language with variation set cues. *Cognition*, *109*, 423–430.
Radford, A. (1988). *Transformational grammar: A first course*. Cambridge, UK: Cambridge University Press.
Rowland, C. F. (2007). Explaining errors in children's questions. *Cognition*, *104*, 106–134.
Rowland, C. F., & Pine, J. M. (2000). Subject-auxiliary inversion errors and *wh*-question acquisition: "What children do know!" *Journal of Child Language*, *27*, 157–181.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.
Sag, I. A. (2010). English filler-gap constructions. *Language*, *86*, 486–545.
Stoll, S., Abbot-Smith, K., & Lieven, E. (2009). Lexically restricted utterances in Russian, German, and English child directed speech. *Cognitive Science*, *33*, 75–103.
Stumper, B., Bannard, C., Lieven, E., & Tomasello, M. (2011). "Frequent frames" in German child-directed speech: A limited cue to grammatical categories. *Cognitive Science*, *35*, 1190–1205.
Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

https://doi.org/10.1017/S0267190512000062 Published online by Cambridge University Press