I'm are what I'm are: The acquisition of first-person singular present BE

THEA CAMERON-FAULKNER and EVAN KIDD*

Abstract

The present study investigates the development of am *in the speech of one English-speaking child, Scarlett (aged 4;6–5;6). We show that* am *is infrequent in the speech addressed to children; the acquisition of this form of BE presents a unique insight into the processes underlying language development because children have little evidence regarding its correct use. Scarlett produced a pervasive error where she overextended* are *to first-person singular contexts where* am *was required* (*e.g.,* I'm are trying, When are I'm finished?). Am *gradually emerged in her speech on what appears to be a construction-specific basis. The findings of the study are used in support of a usage-based, constructivisit approach to language development.*

Keywords: *BE; constructions; errors; input; usage-based.*

1. Introduction

Researchers working in what is broadly defined as a usage-based (UB) approach to language development have highlighted the importance of input frequency in the early stages of language acquisition. Within this approach the early stages of development are characterised by the acquisition of lexically specific constructions (e.g., *It's a X, want Y*) which are initially represented independently of one another within the child's linguistic system. Children are argued to build up increasingly abstract knowledge of the language over time. The frequency of a construction in the input is argued to be one driving force behind acquisition, and there is now a body of literature that indicates the frequency with which constructions are attested in the input correlates with their emergence and use in the speech of young children (e.g., Farrar 1990; Rowland and Pine 2000; Theakston et al. 2001, 2005; Kidd 2006; Kidd and Cameron-Faulkner, in press).

Cognitive Linguistics 18–1 (2007), 1–22
DOI 10.1515/COG.2007.001
0936–5907/07/0018–0001

Although frequency of form as a determinant of acquisition has been a well-studied aspect of the UB approach, the issue of how children acquire less frequently heard structures has yet to be addressed. Clearly children acquire features of language that are rarely attested in the input. In these cases researchers who work from a nativist perspective invoke the poverty of the stimulus argument, and attribute to the child innate knowledge of language to explain the acquisition of rare structural patterns. Functionalist researchers argue that such explanations beg the developmental question. However, there exist very few studies from a functionalist perspective that explicitly address the acquisition of rarely attested forms (but see Reali and Christiansen 2005). This presents a challenge to the UB approach, since there are clearly aspects of the language that children do not hear directly. The present study addresses this issue by documenting and analysing one child's acquisition of the first-person present singular of BE (i.e., *am*). This particular form and its associated constructions were signalled out due to the presence of a pervasive error in the speech of our subject, which can best be described as an overgeneralization of *are* in all constructions where the 1SG.PRES form of BE was required, as in (1):

(1) *I'm are trying.*
Are I'm going to school tomorrow?
What are I'm having for mains?

The constructions encoding 1SG.PRES BE are, for pragmatic reasons, likely to be infrequent in the direct input to children (see also Dąbrowska 2004). The types of communicative intents expressed by such constructions (e.g., questioning of future activities in which the speaker has limited control, asking for affirmation, emphatic protesting) are not the kinds of intents commonly expressed by adults to young children. Moreover, in a lexically based study of child-directed speech, Cameron-Faulkner, Lieven and Tomasello (2003) showed that the 1SG.PRES form of BE only occurred in one syntactic frame, and even then not in its full form: the only instantiation of the 1SG.PRES form of BE that children heard was in the *I'm V NP* construction (e.g., *I'm eating a cake*), a frame that accounted for only 0.01 percent of all transitive utterances. The acquisition of 1SG.PRES BE thus presents an opportunity to examine the types of strategies a child may use when acquiring structures heard infrequently in the input.

The acquisition of BE has been the focus of a number of child language studies (e.g., Brown 1973; Kuczaj 1985/1986; Joseph et al. 2002; Wilson 2003; Theakston et al. 2005). A major issue relates to whether BE exists as an instantiation of the larger grammatical category of "inflection" from

the onset of development (e.g., Hyams 1992) or whether knowledge of BE is built up on a piecemeal basis (e.g., Wilson 2003; Theakston et al. 2005). Typically, studies of BE focus on its use and non-use. This emphasis reflects a developmental stage in which production of BE appears to be optional in a child's grammar. Lexically based studies indicate that the consistency of BE provision is dependent upon on the actual pronoun +BE unit in question (e.g., *I'm*, *he's*, *it's*), suggesting that use of BE is item-specific in the early stages of development and not part of a larger, more abstract grammatical category (Wilson 2003; Theakston et al. 2005). Both of these studies suggest that the development of BE begins with the storage of lexically specific frames (e.g., *I'm V-ing*, *she's V-ing*), which, over time, are involved in a gradual process of abstraction, resulting in a more schematic representation of BE. Theakston, Lieven, Pine and Rowland (2005) relate the trajectory of BE development to the characteristics of the input and suggest that the order of emergence of pronoun+BE units (e.g., *I'm*, *she's*) is determined by the frequency of the units in the input. This trend, however, is not without exception. Theakston and coauthors reported that the most frequent pronoun+BE unit in their input sample was *you're*, which was amongst the latest to emerge in the speech of the children sampled in the study. The late emergence of *you're* in child speech argues against a purely probabilistic explanation of the acquisition of BE. Instead, the late emergence appears to be pragmatically motivated: descriptions of caregivers' ongoing activity (e.g., *You're singing*) are presumably rare in the speech of children around the age of two or three (the age at which the children were sampled), whereas descriptions of the child's ongoing activities by the caregiver are extremely frequent, as attested by the frequency of *you're V-ing* utterances in the input sample. When utterances containing *you're* were omitted from the input analysis a significant correlation was identified between the frequency of pronoun+BE units in the input and their order of emergence in the speech of the sampled children. The availability of individual forms thus appears to be a necessary but insufficient condition for early acquisition.

Both Wilson (2003) and Theakston et al. (2005) focused on the use of pronoun+BE units, and both papers concluded that these units are tied to sentence-level constructional units. Consequently it could be argued that utterances such as *I'm running* and *She's sad* do not display any knowledge of BE. In the same way that child language researchers view units such as *don't* and *can't* as unanalysed units which do not indicate any knowledge of *do* and *can* in a child's linguistic system (e.g., Klima and Bellugi 1966; Choi 1988), so too should we be cautious of discussing the acquisition of BE based on the production of pronoun+BE units. Instead,

it may be more accurate to describe units such as *I'm* and *she's* as instances of allomorphic variation triggered by specific types of constructions. For example, the child could learn that when discussing ongoing activities the present progressive construction is required and in this construction the subject pronoun + contracted BE (*I'm*, *you're*, *he's*, etc.) functions as a complex allomorph in subject position. Some studies have investigated the acquisition of BE through the analysis of full forms. For example, Kuczaj reported a study of BE development in which contracted forms were omitted "because of the difficulties involved in determining the status of such forms in young children's speech" (1985/1986: 111). Additionally, Theakston and Lieven (2005) presented findings of an imitation and elicitation study that focused on the full forms of BE and HAVE. In both cases the studies once again indicated that acquisition of the associated forms was piecemeal and closely related to specific constructions.[1]

The suggestion that knowledge of BE may be tied to constructions of varying levels of schematicity in the early stages of development leads naturally to the issue of how constructions are related. In traditional linguistic theory interrogatives are claimed to be formed through subject-auxiliary inversion. In contrast, non-transformational approaches invoke a range of non–movement based explanations; declaratives and interrogatives are argued to be partially independent. In usage-based approaches to acquisition, declaratives and interrogatives are viewed as initially independent structures, between which links gradually emerge over time (Tomasello 2003). Thus, in UB approaches, question formation does not involve any kind of operation on a declarative utterance, but instead the activation of the appropriate interrogative construction. Accounts invoking subject-auxiliary inversion have been challenged directly by Rowland and Pine (2000) and Rowland, Pine, Lieven, and Theakston (2005), who suggested that *wh*-question formation involves the production of lexically specific constructions fronted by *wh*-word auxiliary combinations (e.g., *What are*, *Where's*), which are extracted from the input. The study of question formation leads to a testable hypothesis: if interrogatives are formed through a process of movement then the representation of BE in interrogatives should be the same as for declaratives. If, however, interrogative constructions develop independently of declarative constructions then it may be possible to observe qualitative differences in the representation of BE between the constructions. The errors presented in the present study provide an ideal opportunity to evaluate the hypothesis.

Unlike previous studies of BE development, the present analysis is based on error data and the subsequent recovery from the erroneous

representation. Specifically, we investigate the patterns of the attested error in order to ascertain the child's knowledge of BE. The error in question can be categorised as a form of overgeneralisation in which the morpheme *are* is overextended and used with the 1SG.PRES pronoun. Overextension and, more broadly, overgeneralisation errors are part and parcel of language development (e.g., Bowerman 1988). These types of error suggest that the child is forming generalisations based on stored exemplars of speech, and is thus on the way to becoming a competent speaker of their target language. Overgeneralisation/overextension errors are attested in many areas of linguistic competence. In English the most notable examples are often found in expression of irregular past-tense forms (e.g., *runned*, *goed*) and irregular plural forms (e.g., *feets* and *sheeps*), but are also found at the construction level. For example, overgeneralisation of the resultative construction (e.g., *The doggie bited him untied*), as reported by Bowerman (1988). Although the errors are well documented the process by which children recover from these novel expressions is contentious.

The main mechanism responsible for the recovery from overgeneralisation errors within the UB approach is *competition*. According to MacWhinney (2004), overgeneralisation/overextension errors occur as a result of tension between two pressures. Firstly, the process of analogy results in the creation of an overgeneralisation/overextension error. Analogy is a powerful cognitive process which plays a pivotal role within UB theories of language development. The process is responsible for the formation of generalisations across constructions of similar form and function and leads to the eventual formulation of more schematic representations of linguistic knowledge (Tomasello 2003). Sometimes this analogic force results in the creation of schematic constructions attested in adult speech (e.g., the transitive construction, the passive construction and so on), but it can also result in the creation of novel constructions which, though logical from the child's perspective, are not found in the target language (i.e., overgeneralisations). Competing with this analogic force, in the case of systematic errors, is the stored representation of input forms in declarative memory. Thus errors formed through analogy (e.g., *runned*) compete directly with the "correct" form, and eventually drop out of the child's linguistic system as the strength of the input form is entrenched (MacWhinney 2004). The speed at which recovery takes place is argued to depend on the frequency of the "correct" form in the input and the extent to which the error has become entrenched in the linguistic system of the speaker.

The present study had three aims. The first aim was to document the type of strategy used by one child to express the 1SG.PRES form of BE.

The second aim was to explore reasons why the child adopts the particular strategy attested in her speech. The final aim was to investigate the processes used by the child to curtail *are* overextension.

2. Method

The paper consists of two studies. Firstly we present a case study documenting one child's 1SG.PRES utterances containing full forms of BE. The case study is based on diary data collected by the child's mother. The second study documents the frequency of all full forms of BE in a sample of child-directed speech in order to present a picture of the types of forms that the child may be hearing on a daily basis.

2.1. *Study 1: Case study*

Participant. The target child (Scarlett) was aged 4;6 at the beginning of data collection and 5;6 at its completion. Scarlett is a monolingual English speaker and the oldest of three children. She has attended part-time nursery since the age of 1 and both her mother and father have been the primary caregiver in the home at different stages. The CELF-Preschool UK (Clinical Evaluation of Language Fundamentals, Semel et al. 2000) test was administered at age 5;3. At the time of testing Scarlett had a receptive language age of 6;0 and an expressive language age of 7;2 (overall language age 6;10), placing her well above the average for her chronological age. Therefore it cannot be argued that the errors we report on here were due to any kind of language impairment.

Data collection. The mother collected all spontaneous utterances in which 1SG.PRES BE was encoded by a full form (i.e., *am* or *are*). These were written down on paper and then entered into a database at regular intervals. Thus the present study is based on diary data, which brings with it both the benefit of a natural communicative setting and the limitation that not all target utterances may have been captured.

Sampling of child's speech. The corpus consists of two data sets. The first data set (Sample 1) consists of two months of data collected between August and September 2004. Collection of the second data set began in February 2005 and continued until the end of July 2005.[2] This period comprises Samples 2 to 4. During Sample 4, only question forms were noted since 1SG.PRES BE errors were only attested in this particular construction type by this point. Table 1 summarises the corpus.

Table 1. *Corpus summary*

	Period of sample	Age of child	No. of utterances
Sample 1	August–September 2004	4;7–4;8	26
Sample 2	February–March 2005	5;1–5;2	45
Sample 3	April–May 2005	5;3–5;4	78
Sample 4	June–July 2005	5;5–5;6	17

Analysis. In the analysis utterances were coded for construction type using the following broad taxonomy:

i. Declaratives, e.g., *I am poorly.*
ii. Presentationals, e.g., *Here I am*
iii. *wh*-questions, e.g., *Where am I going tomorrow?*
iv. *yes/no*-questions, e.g., *Am I good too?*

The frequency of *am* and *are* was calculated for the utterances within each construction type for each sample. The form of the subject pronoun used in the target utterances was also analysed due to the observation that two pronoun forms (*I'm* and *I*) occurred in the target utterances.

Negated utterances and tags were omitted from all analyses. Negated utterances were excluded since the uncontracted form of 1SG.PRES BE (i.e., *I am not*) did not appear in Scarlett's speech. Instead Scarlett favoured *I'm aren't* and later *I'm not*. Tag questions were omitted because the issue of polarity added a level of complexity which detracts from the general aims of the study.

2.2. *Study 2: Analysis of BE in child-directed speech*

Traditionally, input samples are taken from the speech of the child's mother. However, since the present study is based on diary data as opposed to naturalistic recordings, it is not possible to present this form of analysis. Furthermore, like many children, Scarlett had been exposed to large quantities of linguistic input from a variety of sources (i.e., her father, siblings, a range of caregivers in the day-care environment, and, later, primary school teachers) and thus an input sample based solely on the mother would not be representative of the child's ambient speech. Since it would be well beyond the scope of this paper to analyse such a wide range of speakers, a compromise was reached which involved pooling the speech of a number of mothers' from the Wells corpus (Wells 1981). The Wells corpus was selected for three reasons. Firstly, the corpus consists of the speech of 32 mother–child dyads which can be pooled to present a more general picture of child-directed speech than would be

possible given a corpus sampling fewer dyads. Secondly, the method of data collection used in the corpus is more naturalistic than that of many other corpora of child language. The data in the Wells corpus were collected using a child-attached microphone that turned on at random times during the day, resulting in a sample that cut across daily activities. The microphone picked up ambient speech and thus recorded the caregiver in addition to the child. Finally the corpus consisted of samples taken from British English speakers and therefore was considered to be similar to the linguistic environment encountered by Scarlett on a daily basis.

All files in which the children were aged 3;6 and over were incorporated into the input analysis, resulting in a sample of 33 files taken from 24 mothers. Therefore while the data do not reflect the specific nature of the input directed to Scarlett, the analysis should present a reasonably accurate indication of the forms of BE (and their frequencies) typically heard by a child of Scarlett's age.

Analysis. All utterances containing full forms of BE were extracted from the input corpus and pooled across the mothers. The frequency of each form of BE was then calculated within the four broad construction types employed in the case study.

3. Results

3.1. *Study 1*

Study 1 focuses on the development of constructions containing forms of 1SG.PRES BE in Scarlett's speech sample. The analysis of the pronoun form used with BE is also incorporated since a preliminary analysis of the data indicated that Scarlett used *I* and *I'm* interchangeably in her expression of the 1SG.PRES pronoun during the sampling period. The analysis aimed to discover any developmental trends underlying this variation.

Sample 1. Figure 1 presents Scarlett's representation of 1SG.PRES BE in the Sample 1 data set.

In the first sample Scarlett used *are* in all but two constructions where the first-person present singular form of BE was required. This pattern is attested in all construction types, as shown in (2).

(2) *I'm are an astronaut.* (declarative)
Here I'm are. (presentational)
Are I'm snuggly? (*yes*/*no*-question)
What are I'm going to wear? (*wh*-question)

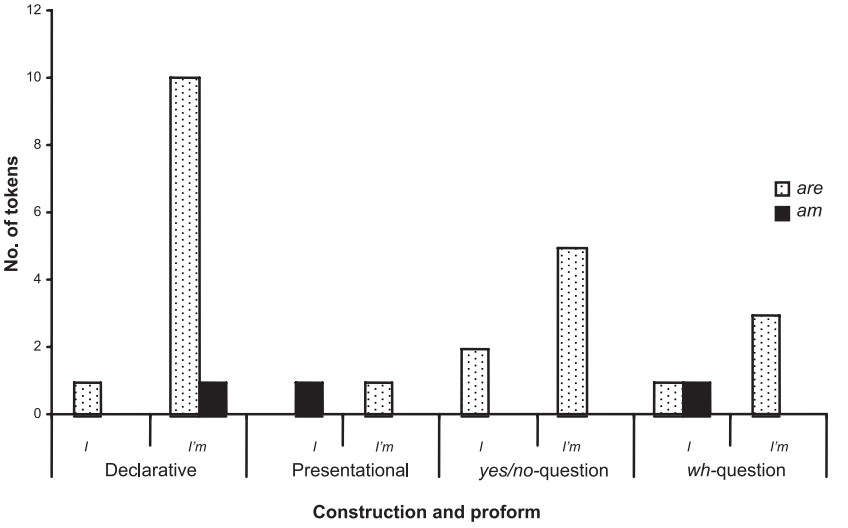

Figure 1. *Token frequency of* am *and* are *in the Sample 1*

Two forms of the first-person pronoun occur in the sample, *I* and *I'm*.[3] As Figure 1 indicates, *are* co-occurs with both *I'm* and *I*, but the former is much more frequent and is found in 83 percent of 1SG.PRES BE **are* utterances. *I'm* and *I* also occur with *am*, but as there is only one token of the former and two of the latter it is not possible to make any generalisations about preferential pairings of *am* and pronoun forms at this stage. A significant McNemar test of homogeneity for dependent observations showed that the pronoun distribution was beyond that expected by chance ($\chi^2 = 16.2$, df $= 1$, $p < 0.001$). This result is largely carried by Scarlett's use of different permutations of *I'm* and *are* (e.g., **I'm are* ..., **Are I'm* ...), suggesting that this was a pervasive error at this sampling period.

Sample 2. Figure 2 presents the findings for the Sample 2 data set and indicates asymmetry between the forms used to express 1SG.PRES BE across the different construction types.

Declaratives and presentationals are now expressed by *I am*, resulting in utterances such as those shown in (3):

(3) *I am deep red.*
Here I am.

Conversely, all *yes/no*-questions and the majority (80 percent) of *wh*-questions are still expressed by *are*. Thus it appears that Scarlett does not

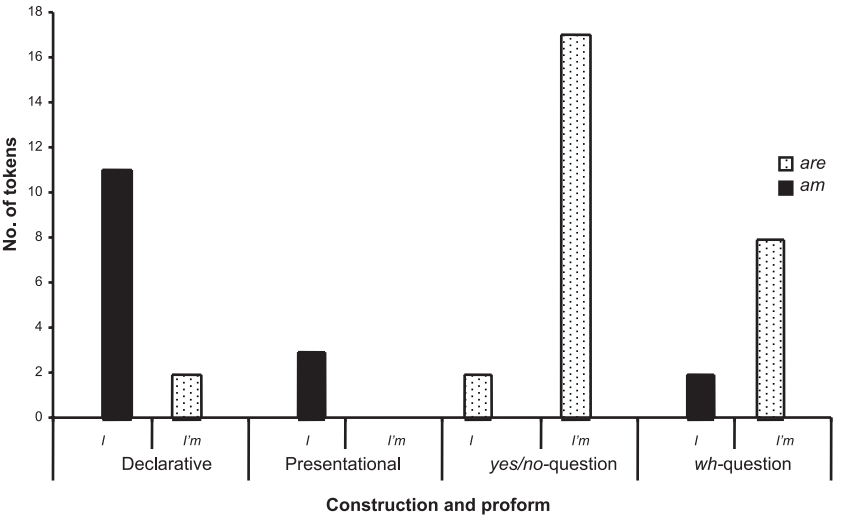

Figure 2. *Token frequency of* am *and* are *in the Sample 2 data set*

possess a single adult-like representation of BE, and instead development towards an adult-like use appears to be construction-based.

The pronoun analysis of Sample 2 indicates a correlation between *I'm are*/*are I'm* and *I am*/*am I* pairings. This finding indicates that in some cases Scarlett has broken *I'm* down into its constituents (resulting in constructions containing *I am*) but in others *I'm* is still produced as an unanalysed unit and consequently *are* overextension persists. A McNemar test of homogeneity for dependent observations showed that this distribution approached significance ($\chi^2 = 2.81,\ \mathrm{df} = 1,\ p < 0.10$). Despite this marginal result, Scarlett is still producing more errors than correct utterances.

Sample 3. In the Sample 3 data set, *am* finally emerges in Scarlett's *yes*/*no*-questions, though *are* is still the predominant realisation of 1SG.PRES BE in both *yes*/*no*- and *wh*-questions (see Figure 3).

The pronoun analysis indicates that *I'm* is still strongly linked to *are* while *I* co-ocurs predominantly with *am*. A significant McNemar test of homogeneity for dependent observations showed that this distribution was beyond that expected by chance ($\chi^2 = 12.16,\ \mathrm{df} = 1,\ p < 0.001$). This result is driven by the fact that there is a striking asymmetry between Scarlett's use of different forms of BE and the pronouns with which these different forms (*am* and *are*) co-occur. *Are* occurs most often with *I'm*, suggesting that the error observed in Sample 1 is still pervasive.

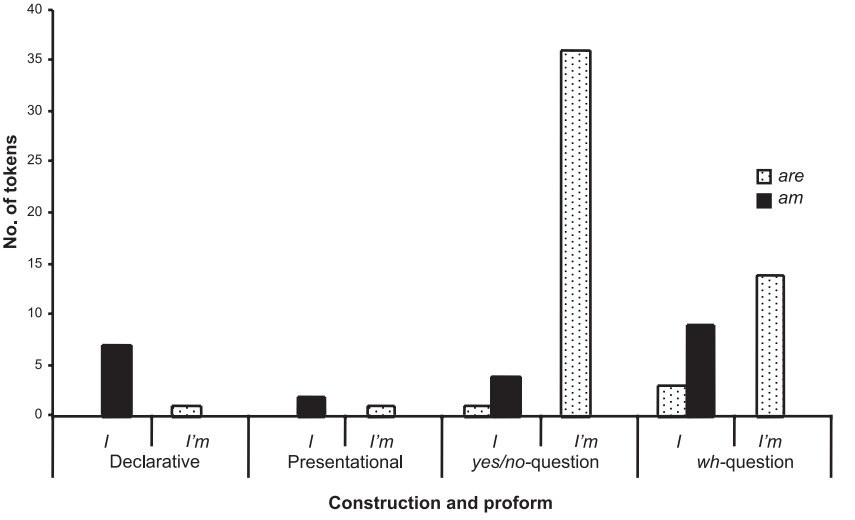

Figure 3. *Token frequency of* am *and* are *in the Sample 3 data set*

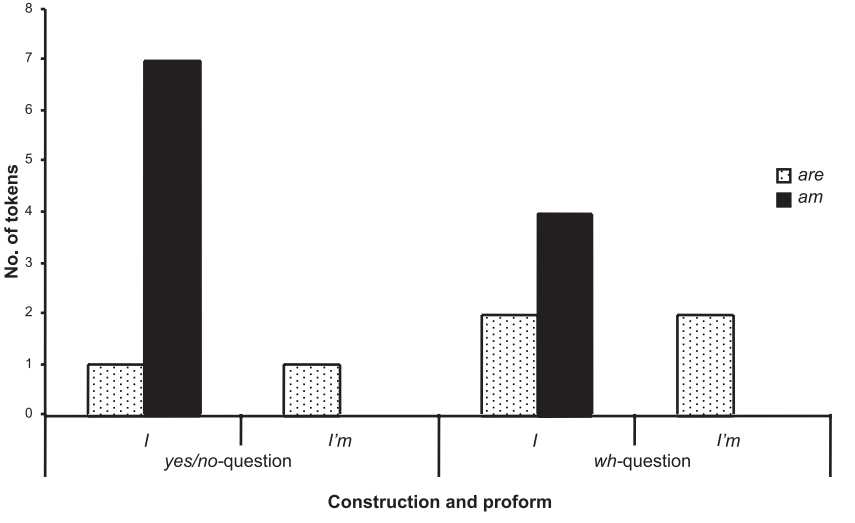

Figure 4. *Token frequency of* am *and* are *in the Sample 4 data set*

Sample 4. Since Scarlett had been using *am* correctly in declaratives and presentationals in the previous two samples, only interrogatives (*wh*- and *yes/no*-questions) were analysed in the Sample 4 data set. As Figure 4 indicates, both *am* and *are* are attested in both question types. However, by

this stage *am* has taken over as the predominant form of 1SG.PRES BE in *yes/no*-questions and is equal in frequency to *are* in *wh*-questions. Thus, even though *am* is now used productively in all construction types, including *yes/no*-questions, Scarlett is still producing some errors involving *are*.

The pronoun analysis indicates the dominance of *I* in the utterances attested in the final sample (82 percent). Of these utterances, 78 percent co-occur with *am*, suggesting that Scarlett is beginning to produce fewer errors. A significant McNemar test of homogeneity for dependent observations showed that this distribution was beyond that expected by chance ($\chi^2 = 4.57,\ \mathrm{df} = 1,\ p < 0.05$). Unlike the previous analyses, where the effects were driven by the production of errors, this significant result is driven by the correct combination of *I* and *am*.

3.2. *Study 2*

Study 2 focuses on the frequency of BE morphemes in the input sample. Only full forms were included in the analysis. The results are shown in Figure 5.

In total, 259 tokens of full-form BE were identified in an input sample of 4136 utterances. Full-form BE was therefore attested in just 6.26 percent of utterances in the sample. Only 18 tokens of *am* were found in the corpus, accounting for 6.10 percent of all BE tokens. These tokens were restricted to declaratives and presentations. Thus it appears that there is very little positive evidence of *am* in the input sample selected for the present study, and, if this observation is representative of the input addressed to Scarlett, it may account for her late acquisition of the

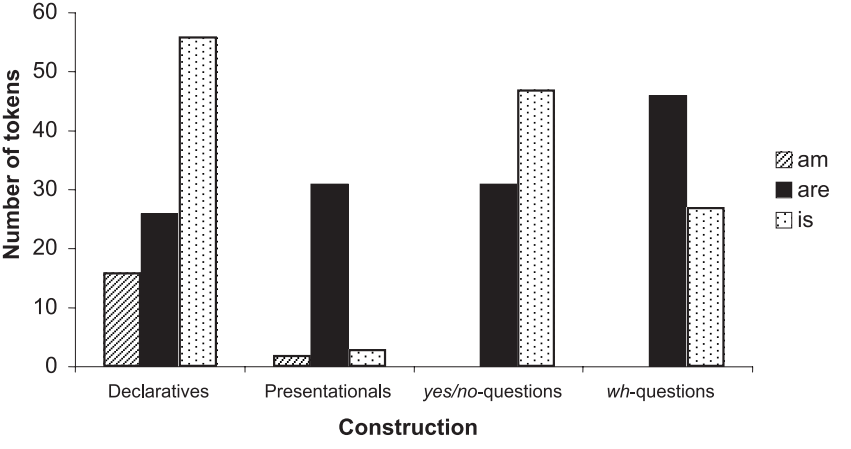

Figure 5. *Token frequency of* am, are, *and* is *in the input sample*

morpheme. *Are* and *is* were much more frequent, accounting for 45.42 and 45.08 percent of all BE tokens in the input sample respectively. Input frequency may thus play a role in Scarlett's overextension of *are* to first-person contexts. However, if input frequency is a factor then the question emerges of why Scarlett overextends *are* and not *is* in her utterances. The possible role played by input frequencies of BE forms is discussed in the next section.

4. Discussion

In this paper we documented a very specific error which to our knowledge has not yet been reported in child-language literature.[4] Most studies relating to the acquisition of BE focus on the use and non-use of the morpheme, possibly since agreement errors are considered to be rare (Joseph et al. 2002).[5] The error has implications for the representation of BE during language development and also for the cognitive processes that play a role in language development.

As in other studies of overextensions (e.g., Bowerman 1988, 1996; Gropen et al. 1996; Marcotte 2006), the findings of the present study provide a unique insight into the acquisition process and Scarlett's representation of BE. Scarlett's correct use of *am* was construction dependent, which we suggest can be explained in one of two ways. First, the data could suggest that even at age five, BE was not a paradigmatic linguistic category for Scarlett, and consequently she did not have interconnected knowledge of the use of *am*. This explanation is consistent with arguments made in the UB acquisition literature (e.g., Tomasello 2003), where acquisition is characterised by the gradual development of independent constructions. This explanation takes Scarlett's productions as directly indicative of her grammatical knowledge. Although we take Scarlett's errors as suggesting her knowledge is non–adult like, at the time she was producing this error she was relatively old and had quite a lot of experience with language.

An alternative interpretation grants Scarlett more interconnected knowledge of BE and explains the errors as reflecting an interaction between weak abstract linguistic knowledge and her developing production system. That is, Scarlett possessed implicit knowledge of the distribution and function of the different forms of BE but made errors when required to explicitly use low-frequency forms (i.e., *am*). MacDonald (1999) argued that sensitivity to distributional information is an important link between the comprehension and production systems and the acquisition of language. Children must attend to distributional information to learn the syntactic regularities of their input language, but these regularities will be

more or less difficult to identify depending on their frequency of occurrence. Therefore, since *am* is not as well attested as other forms of BE, its distribution is difficult to identify and hence, as we observe with Scarlett, its use is prone to error.

Let us consider this hypothesis in more detail. Adult psycholinguistic research investigating sentence production suggests that production is lemma driven and incremental (Chang et al. 2000; Ferreira 1996; Levelt 1989). Additionally, there is evidence to suggest that the structural choices speakers make in production result from the interplay between lexical availability and syntactic production mechanisms (e.g., Arnold et al. 2000; Bock 1986, 1987; Ferreira and Dell 2000). Recent computational work by Chang, Dell, and Bock (2006) suggests that even very young children's production systems can be assumed to be qualitatively similar to adult ones, such that there is a degree of continuity across development. Therefore we assume that children also produce sentences using both their syntactic knowledge and, given the real time pressures of production, the most available lexical items. However, as many UB studies of child language argue, there are crucial differences between children's and adults' lexical and syntactic knowledge. Indeed, children have a smaller vocabulary than adults, but more importantly, their initial grammatical systems are less abstract than those of adults; they rely much more on concrete complex lexical formulae during sentence production (e.g., Lieven et al. 2003). Rowland (2006) has argued that young (two to five-year-old) children's errors in question formation are best captured by a UB account, where errors occur most often when children do not possess pre-packaged lexical formulae to aid production. For example, children who produce **Where does he does go?* do so because they lack the lexical frame *where does* + X. That is, errors occur when children attempt to go beyond their entrenched syntactic knowledge.

We suggest that a similar process could explain Scarlett's errors. The input data presented in Study 2 indicated the infrequency of *am* in the input sample. This should come as no surprise given the types of functions expressed by Scarlett's *am* constructions, such as, for example, protesting, questioning future activities which to some degree are not within her control, and asking for affirmation (examples [4] to [6], respectively):

(4) *I'm aren't laughing.*
(5) *What are I'm going to wear?*
(6) *Are I'm a groovy chick too?*

These types of functions are not typically expressed by caregivers during their interaction with young children. Consequently, Scarlett lacks substantial positive evidence for *am* and is forced to fill the gap for herself.

The next question this raises is why the morpheme *are* is chosen to fill the gap. As Study 2 indicates, *is* occurs with almost the same frequency as *are*, but regardless of this comparable frequency, *are* is selected for overextension.

The most obvious and straightforward explanation is that the error results from first and second-person role reversal within constructions containing BE. That is, Scarlett inserts the first-person pronoun into frames expressing second-person agreement (e.g., *NP are V-ing*, *Are NP Adj?*). A number of studies suggest that role reversal is a key strategy used in the early stages of acquisition, as in (7) and (8).

(7) a. *You will go tomorrow.*
    b. *I will go tomorrow.*
(8) a. *You like cake.*
    b. *I like cake.*

However, in the case of BE, different allomorphs are required depending on the person and number of the subject, and role reversal would therefore result in the erroneous use of *are*, as in (9).

(9) a. *You really are trying.*
    b. **I really are trying.*

Although appealing, this rather straightforward explanation does not fully account for the data attested in the present study. As mentioned previously, the pronoun form used most frequently with **are* is *I'm*, not *I*. Therefore role reversal alone cannot be the answer. The studies of BE development outlined in the introduction highlight the construction-based nature of early pronoun-auxiliary units (e.g., *I'm V-ing*). It is thus possible that Scarlett's *I'm are/are I'm* utterances originate from constructions in which *I'm* occurs as a lexically specified unit, as in the hypothetical examples shown in (10):

(10) a. *I'm trying.* [*I'm V-ing*]
     b. *I'm an astronaut.* [*I'm a/an X*]

Scarlett appears unaware that *I'm* encodes BE in utterances such as those in (10) and therefore, when a form of BE is required (due either to the placement of stress or as a marker of interrogatives), she chooses and inserts a frequently heard form from the input: *are*. However, frequency of *are* in the input cannot be the only factor determining its selection; *is* appears to be just as frequent in the Wells input sample presented in Study 2, and *is* is also the most frequent form found in the input sample analysed by Joseph, Serratrice, and Conti-Ramsden (2002).[6]

We suggest that the adoption of *are* could be motivated by two other factors. Firstly, *are* occurs with the second-person singular (e.g., y*ou are running*). Therefore, given the fact that role reversal of *you* and *I* works for a number of auxiliaries and verbs, as mentioned above, it is possible that the strategy is a contributing factor in *are* being selected for the 1SG.PRES form of BE over and above other forms. Secondly, in addition to the second-person singular, *are* is also used with both the first-person and third-person plural, and is thus the most "flexible" form of BE. Once again, therefore, adoption of *are* would appear to be the best bet, given a lack of evidence to the contrary. Within the conceptual parlance of the Bates and MacWhinney (1989) Competition Model, *are* has both high cue availability and high cue reliability. Given this fact, we suggest that rather than resulting from the process of role reversal alone (i.e., substitution of *I* for *you* in target constructions), Scarlett's **are* constructions are formed through a process of structure building whereby existing constructions and units are combined in order to form novel constructions in the child's speech, as also observed, for instance, by Rowland (2006). Role reversal plays a role in the selection of the "appropriate" form of BE, which is then combined with *I'm X* constructions. The exact process of formation would be dependent upon the target construction. For example, in the case of declaratives, *are* would be inserted after *I'm* (Figure 6), while in *yes*/*no*-questions, *are* would be placed at the front of the construction (Figure 7). The situation is slightly more complex with regard to *wh*-questions, since these utterances may be formed by adding a *wh*-word+*are* unit to the front of the utterance (Figure 8).

How does Scarlett recover from these errors? Scarlett's data indicate that her recovery was slow and gradual, which is indicative of a period

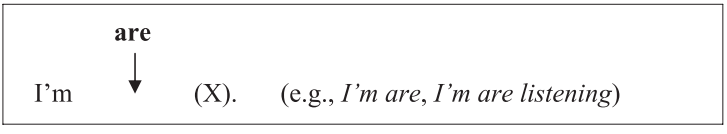

Figure 6. *Structure building of declaratives and presentationals*

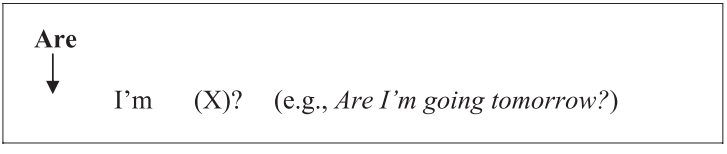

Figure 7. *Structure building of* yes/no-*questions*

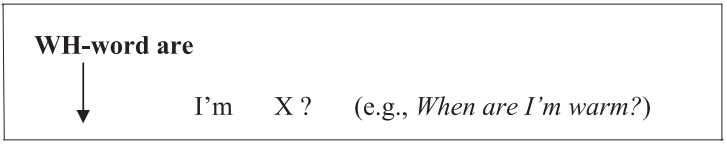

Figure 8. *Structure building of* wh-*questions*

of competition between two forms (*are* and *am*). The use of *are* as a marker of 1SG.PRES BE is firmly entrenched in Scarlett's speech, so much so that it is even found in utterances that would usually be described as "rote learned" (e.g., *Here I am* is expressed as *Here I'm are*). At the same time, *am* would appear to be highly infrequent in the input, resulting in a very weak representation. Thus the competition between the forms **are* and *am* is heavily biased in favour of **are* at the onset of the data sample.

The data suggest that recovery from *are* overextension occurs at different rates, dependent on the construction in question. If we were to make a generalisation, we could say that recovery occurred first in constructions where the auxiliary follows the subject pronoun (i.e., in declaratives and presentationals) and then in constructions where the auxiliary precedes the pronoun (i.e., in interrogatives). A possible reason for this asymmetry may be found in the input. Although *am* is infrequent in the input sample, it does occur in declarative and presentational utterances. It is in these types of constructions that *am* first emerges in Scarlett's speech. Thus the types of constructions for which positive evidence of *am* is attested in the input sample are the same types of constructions which display the first emergence of *am*. The constructions which displayed the longest delay with regard to *am* production are *yes/no*-questions; it is not until Sample 3 that *am* is attested in this particular construction type. This pattern is also attested in Kuczaj (1985/1986) and Theakston and Lieven (2005), where most forms of BE were mastered in declaratives before they emerged in *yes/no*-questions.

The delay in *am* production in interrogatives could well be due to the lack of positive evidence of *am*. Take, for instance, the case of *yes/no*-questions, which showed the longest delay in recovery. Study 2 showed that there is little evidence in the input that *am* occurs in *yes/no*-questions, which provides a potential explanation for Scarlett's errors. However, another factor causing prolonged overextension of *are* in *yes/no*-questions may be the frequency of *are* within these constructions in the input: the fact that *are* occurs frequently within *yes/no*-questions in the input sample may further entrench the use of *are* in Scarlett's *yes/no*-questions. It is interesting to note that in their lexically based study of

child-directed speech, Cameron-Faulkner, Lieven, and Tomasello (2003) reported that questions beginning with *Are you . . .* were the most frequent form of *yes/no*-question found in the sample, accounting for 17 percent of all *yes/no*-questions. Thus the combined factors of a lack of competition from *am* and increased entrenchment of *are* resulting from high-frequency *Are you . . . ?* constructions may be the causes of prolonged use of **are* within Scarlett's *yes/no*-questions.

With respect to the relationship between declaratives and interrogatives, Scarlett's data provide a direct challenge to the assumption that questions are formed through subject-auxiliary inversion. If Scarlett formed questions through movement, we would expect *am* to emerge in her interrogative constructions as soon as the form emerged in her declarative utterances (i.e., in Sample 2). The data in present study indicate that this is not the case: *am* emerges as the predominant form of 1SG.PRES BE in declarative constructions in Sample 2, while *are* overextensions are still the norm in Scarlett's *yes/no* and *wh*-questions during this period. It is only by the time of Sample 4 that *am* occurs reliably in interrogatives. The data thus indicate that interrogative constructions are formed independently of declarative constructions.

This raises the question of the nature of the relationship between declaratives and interrogatives in the developing linguistic system. We make no attempt here to provide a linguistic analysis of the relationship between the two, but merely state that we assume the two are related in the mature linguistic system. A question formed from a declarative sentence will inherit core semantic and syntactic properties from the declarative (for discussion see Van Valin 2002). The formation of this link in acquisition is an open question. Frame-based UB approaches such as those put forward by Rowland and Pine (Rowland and Pine 2000; Rowland 2006) argue that there is no initial link, since children are argued to form their first questions from lexical schemata. Such a link could be established via analogy. This process would require the child to identify the syntactic and semantic commonalities across the two forms. A key factor in this process could be children's understanding of the pragmatics of each construction type. Children clearly understand the illocutionary force associated with each, since they use both declaratives and interrogatives felicitously in discourse. From here they must identify the form- and meaning-based commonalities between constructions.[7]

Scarlett's data also suggest that unanalysed pronoun+BE units (i.e., *I'm*) may be not only unanalysed but also unmarked for BE in a child's linguistic system. The study indicates that Scarlett uses *are* with *I'm* statistically more frequently than with *I*, which suggests that she is unaware that the unit *I'm* encodes BE. As a result, Scarlett produces utterances in

which BE is encoded twice. The data therefore highlight the need to be cautious when ascribing linguistic knowledge to children on the basis of their speech. That is, given the production of an unanalysed unit we cannot ever be sure that the child is aware of its underlying components. Moreover, in the case of pronoun+BE units, constructions containing lexical items such as *I'm*, *she's*, and *it's* cannot reliably be taken as instances of BE knowledge and usage. The assumption in prior studies has been that pronoun+BE units are lexical wholes (Wilson 2003; Theakston et al. 2005) and are therefore probably unanalysed in the early stages of development (and possibly later). If this is the case, however, the question remains as to how much we can learn about the acquisition of BE from analysing these types of units. Scarlett's data indicate the importance of studying full forms as opposed to pronoun+aux units in the documentation and analysis of BE development.

In conclusion, the data indicate that, given low input frequency for a particular form, children may generalise on their existing knowledge in order to "fill a gap" in their syntactic knowledge. This strategy may lead to the production of non–adult like forms which may become entrenched in the child's linguistic system as interim solutions to the problem of sparse input. Recovery from such errors can be a lengthy process dependent on the level of entrenchment of the form and also the frequency of the form in the input. As children's production systems become less reliant on lexical formulae and better able to coordinate the online use of abstract syntactic knowledge and lexical items they will produce less errors.

Received 18 July 2006
Revision received 10 September 2006

University of Manchester

Notes

* We would like to thank Scarlett for her involvement in the study. Authors' preferred email address: ⟨t.cameron@manchester.ac.uk⟩.

1. It should be remembered that use of the full form does not preclude the possibility that the form of BE is still unanalysed. We thank an anonymous reviewer for pointing this out.
2. The break in data collection was a consequence of time pressures at home resulting from a change in routine as Scarlett started school.
3. As the token frequency for pronoun forms within constructions is low due to the overall size of the corpus, the pronoun analysis presented in this section conflates pronoun forms across construction types.
4. Although we have seen no mention of this error in the literature, the phenomena does not seem to be restricted to Scarlett. The first author has heard other children make this error, though not to such a systematic degree. Also Hudson (2000) mentions in a footnote that his daughter produced *Are I* and also *I naren't*.

5. Theakston and Lieven (2005) point out that agreement error rates vary across children and that HAVE constructions contain more agreement errors than auxiliary BE constructions.
6. Interestingly Kuczaj (1985/1986) quotes the following utterance: *I is going* (Ben 2;6); this indicates that overextension of *is* can also occur. However, the paper does not provide quantitative data and it is therefore difficult to ascertain whether this was a systematic error or a "slip of the tongue".
7. It is interesting to note that *am* emerged in Scarlett's speech after she began formal education. Studies (e.g., Perera 1984) suggest that the acquisition of particular linguistic constructions may be the result of literacy skills learned at school. It is possible that the emergence of *am* in Scarlett's speech was triggered in part through explicit lexical learning through literacy, and perhaps through sociolinguistic factors such as consistent peer input to which she would be likely to conform.

References

Arnold, J. E., T. Wasow, T. Losongco, and R. Ginstrom
2000 Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76, 28–55.

Bates, E. and B. MacWhinney
1989 Functionalism and the competition model. In MacWhinney, B. and E. Bates (eds.), *The Cross-linguistic Study of Sentence Processing*. Cambridge: Cambridge University Press.

Bock, J. K.
1986 Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 12, 575–586.
1987 An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language* 26, 119–137.

Bowerman, M.
1988 The "no negative evidence" problem: How do children avoid constructing an overgeneral grammar? In Hawkins, J. A. (ed.), *Explaining Language Universals*. Oxford: Basil Blackwell.
1996 Argument structure and learnability: Is a solution in sight? In *Proceedings of the 22nd Berkeley Linguistics Society Meeting* (BLS 22), Berkeley, CA: University of California—Berkeley, Department of Linguistics, 454–468.

Brown, R.
1973 *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Cameron-Faulkner, T., E. V. M. Lieven, and M. Tomasello
2003 A construction based analysis of child directed speech. *Cognitive Science* 27, 843–873.

Chang, F., G. S. Dell, and K. Bock
2006 Becoming syntactic. *Psychological Review* 113 (2), 234–272.

Chang, F., G. S. Dell, K. Bock, and Z. M. Griffin
2000 Structural priming as implicit learning: A comparison of models of sentence production. *Journal of Psycholinguistic Research* 29, 217–229.

Choi, S.
1988 The semantic development of negation: A cross-linguistic longitudinal study. *Journal of Child Language* 15, 517–531.

Dąbrowska, E.
2004 *Language, Mind and Brain*. Edinburgh: Edinburgh University Press.
Farrar, M. J.
1990 Discourse and the acquisition of grammatical morphemes. *Journal of Child Language* 17, 604–624.
Ferreira, V. S.
1996 Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language* 35, 724–755.
Ferreira, V. S. and G. S. Dell
2000 Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* 40, 296–340.
Gropen, J., J. Blaskovich, and G. DeDe
1996 Come it closer: Causative errors in child speech. In Stringfellow, A., D. Cahana-Armita, E. Hughes, and A. Zukowski (eds.), *Proceedings of the 20th Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press, 272–283.
Hudson, R.
2000 **I amn't*. *Language* 76, 297–323.
Hyams, N.
1992 The genesis of clausal structure. In Meisel, J. (ed.), *The Acquisition of Verb Placement: Functional Categories and V2 Phenomena in Language Acquisition*. Dordrecht: Klewer, 393–397.
Joseph, K. L., L. Serratrice, and G. Conti-Ramsden
2002 Development of copula and auxiliary BE in children with Specific Language Impairment and younger unaffected controls. *First Language* 22, 137–172.
Kidd, E.
2006 The acquisition of complement clause constructions. In Kelly, B. and E. V. Clark (eds.), *Constructions in Acquisition*. Stanford, CA: CLSI Publications, 311–331.
Kidd, E. and T. Cameron-Faulkner
in press The acquisition of the multiple senses of 'with'. To appear in *Linguistics*.
Klima, E. S. and U. Bellugi
1966 Syntactic regularities in the speech of children. In Lyons, J. and R. J. Wales (eds.), *Psycholinguistic Papers*. Edinburgh: Edinburgh University Press.
Kuczaj, S. A.
1985/1986 General developmental patterns and individual differences in the acquisition of copula and auxiliary *be* forms. *First Language* 6, 111–117.
Levelt, W. J. M.
1989 *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
Lieven, E., H. Behrens, J. Spears, and M. Tomasello
2003 Early syntactic creativity: A usage-based approach. *Journal of Child Language* 30, 333–370.
MacDonald, M. C.
1999 Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. In MacWhinney, B. (ed.), *The Emergence of Language*. Mahwah, NJ: Erlbaum, 177–196.
MacWhinney, B.
2004 A multiple process solution to the logical problem of language acquisition. *Journal of Child Language* 31, 883–914.

Marcotte, J. P.
2006 Causative alternation errors as event-driven construction paradigm completions. In Clark, E. V. and B. F. Kelly, *Constructions in Acquisition*. Stanford, CA: CSLI Publications, 205–232.
Perera, K.
1984 Children writing and reading: Analysing classroom language. Oxford: Blackwell.
Reali, F. and M. H. Christiansen
2005 Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science* 29, 1007–1028.
Rowland, C.
in press Explaining errors in children's questions. *Cognition*.
Rowland, C. F. and J. M. Pine
2000 Subject-auxiliary inversion errors and *wh*-question acquisition: "What children do know?" *Journal of Child Language* 27, 157–181.
Rowland, C. F., J. M. Pine, E. V. M. Lieven, and A. L. Theakston
2005 The incidence of error in young children's *wh*-questions. *Journal of Speech, Language, and Hearing Research* 48, 384–405.
Semel, E., E. H. Wiig, and W. Secord
2000 *Clinical Evaluation of Language Fundamentals.* 3rd UK ed. London: Harcourt.
Theakston, A. L. and E. V. M. Lieven
2005 The acquisition of auxiliaries BE and HAVE: An elicitation study. *Journal of Child Language* 32, 587–616.
Theakston, A. L., E. V. M. Lieven, J. M. Pine, and C. F. Rowland
2001 The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language* 28, 127–152.
2005 The acquisition of auxiliary syntax: BE and HAVE. *Cognitive Linguistics* 16–1, 247–277.
Tomasello, M.
2003 *Constructing a Language: A Usage-based Theory of Language Acquisition.* Cambridge, MA: Harvard University Press.
Van Valin, R. D.
2002 The development of subject-auxiliary inversion in English *wh*-questions: An alternative analysis. *Journal of Child Language* 29, 161–175.
Wells, C. G.
1981 *Learning through Interaction: The Study of Language Development.* Cambridge: Cambridge University Press.
Wilson, S.
2003 Lexically specific constructions in the acquisition of inflection in English. *Journal of Child Language* 30, 75–115.