J. Child Lang. **33** (2006), 859–877. © 2006 Cambridge University Press
doi:10.1017/S0305000906007537 Printed in the United Kingdom

NOTE

The effect of sampling on estimates of lexical specificity and error rates*

CAROLINE F. ROWLAND AND SARAH L. FLETCHER
University of Liverpool

(*Received* 14 *March* 2005. *Revised* 7 *December* 2005)

ABSTRACT

Studies based on naturalistic data are a core tool in the field of language acquisition research and have provided thorough descriptions of children's speech. However, these descriptions are inevitably confounded by differences in the relative frequency with which children use words and language structures. The purpose of the present work was to investigate the impact of sampling constraints on estimates of the productivity of children's utterances, and on the validity of error rates. Comparisons were made between five different sized samples of wh-question data produced by one child aged 2;8. First, we assessed whether sampling constraints undermined the claim (e.g. Tomasello, 2000) that the restricted nature of early child speech reflects a lack of adultlike grammatical knowledge. We demonstrated that small samples were equally likely to under- as overestimate lexical specificity in children's speech, and that the reliability of estimates varies according to sample size. We argued that reliable analyses require a comparison with a control sample, such as that from an adult speaker. Second, we investigated the validity of estimates of error rates based on small samples. The results showed that overall error rates underestimate the incidence of error in some rarely produced parts of the system and that analyses on small samples were likely to substantially over- or underestimate error rates in infrequently produced constructions. We concluded that caution must be used when basing arguments about the scope and nature of errors in children's early multi-word productions on analyses of samples of spontaneous speech.

[*] Thanks are due to Julian Pine, Javier Aguado-Orea and Ben Ambridge who provided valuable comments on a previous draft and to Lara's caregivers for collecting the data. The study was supported by the Economic and Social Research Council, Grant No. RES000220241. Address for correspondence: Caroline Rowland, School of Psychology, University of Liverpool, Bedford Street South, Liverpool, L69 7ZA, UK. tel: +44 (151) 794 1120; fax: +44 (151) 794 2945; e-mail: crowland@liverpool.ac.uk

INTRODUCTION

Naturalistic data analysis is recognised as one of the primary tools in the investigation of children's language acquisition and has played a key role in the formation and evaluation of all major theoretical frameworks (e.g. Braine, 1976; Pinker, 1984; Radford, 1990; Valian, 1986). Traditionally, samples of naturalistic data are transcripts of audio or videotaped conversations between children and their caregivers, which usually take place once or twice a month. Some studies provide cross-sectional data for a large number of children at a particular point in time (e.g. Rispoli, 1998), others follow a small number of children longitudinally through development (e.g. Brown, 1973), but both provide only a snapshot of the child's language – traditionally sampling only 1 to 2% of a child's utterances. This has implications for how naturalistic data can be used.

The impact of sampling on measures of vocabulary learning has been studied extensively in recent years. In particular, the work of Malvern and Richards has demonstrated that sample size has a significant distorting effect on measures of vocabulary diversity, which has led to the production of a programme – VOCD – that can be used to compute more accurate measures (see Malvern & Richards, 1997). However, relatively little attention has been paid to the possible limitations of the technique in investigating other areas of acquisition. In particular, sampling is likely to have a confounding effect on research in two areas: estimating the variability and range of syntactic structures that the child produces, and estimating the rate of error in children's utterances. The aim of the present paper was to investigate the effect of sampling and to provide some suggestions for the accurate use of naturalistic sampled data.

Lexical specificity in children's speech

A central tenet of much recent constructivist work is the idea that many of children's early multi-word utterances are based on lexically specific constructions (see Tomasello, 1992, 2000, 2003). On this account, the child's early knowledge of grammar is tied to individual lexical items or lexical frames (e.g. *it's a* or *where's the*). Within the theory, the child is initially attributed only with knowledge of the frame and the appropriate word types that can slot into the frame (e.g. *where can+I go/he go/you go* or *what's+he doing/he eating/she making*). Thus, the child's knowledge is initially restricted to knowledge of how lexical items behave and combine (i.e. lexically specific knowledge).

A powerful body of research on naturalistic data supports this view by demonstrating that the majority of children's early multi-word utterances consist of only a restricted range of lexical items. The studies show that most of children's early speech can be explained in terms of children applying knowledge of how a small number of individual lexically specific frames behave in their language. For example, Pine, Lieven & Rowland (1998) have demonstrated that between 67 and 90% of the different subject–verb combinations produced by 12 English-learning children in the first six months of multi-word speech could be accounted for by one of five frequently occurring lexical subject+verb patterns. Pine & Lieven (1997) have demonstrated that children's early determiner use may be concentrated around a relatively small number of lexically specific frames: 56% of children's utterances including an article were accounted for by one of three frequently occurring determiner patterns. Lieven, Pine & Baldwin (1997) have shown that 60% of the utterances produced by the children they studied could be explained in terms of a lexically-based positional analysis.

Similar effects have been found in wh-question acquisition. Fletcher (1985) has argued that the earliest wh-questions can be explained in term of the application of three formulaic patterns, and Rowland & Pine (2000) have demonstrated that even at age 3 years, the majority of one child's correct wh-questions could have been produced by the application of rote-learned semi-formulaic lexical frames (see also Dąbrowska, 2000).

These studies provide strong support for the constructivist idea that children produce utterances not by applying adultlike abstract grammatical categories and rules but by using knowledge of how individual lexical items behave. However, a crucial possible flaw with the evidence is that the lexical specificity of the data may simply be a by-product of the fact that researchers are analysing small sample sizes, combined with the effect of the frequency statistics of the language the child is speaking.

Speech, even adult speech, tends to be made up of a small number of words that occur often (e.g. the wh-words *what* and *where*, the verbs *do* and *be*) and a much larger number of words that occur far less often (e.g. *why, when, bounce, gobble*). The high frequency items are more likely to be represented in any given sample than the low frequency ones. Thus, the traditional measure of lexical specificity – demonstrating that a significant proportion of a child's utterances can be accounted for by a small number of lexically specific frames – is confounded by the fact that a small number of highly frequent utterance types are likely to account for a large amount of the data anyway. Analyses based on these samples may, then, underestimate the variety and productivity of children's speech (Naigles, 2002).

Another problem is the fact that the child can only produce utterances using vocabulary items she has already learnt, so a child with a restricted vocabulary is unlikely to produce a large range of grammatical structures. For example, a child who knows only two wh-words will appear more lexically specific than a child who knows four, even if both children have equal knowledge of how questions are formed. Thus, lexical specificity in the data could also be due to a limited vocabulary, not to limited grammatical knowledge.

An obvious solution is to collect much bigger samples (see Lieven, Behrens, Spears & Tomasello, 2003). Another is to carry out controlled experiments to test the limits of children's knowledge (e.g. Akhtar & Tomasello, 1997). However, given the large number of corpora currently available to researchers (e.g. through CHILDES, MacWhinney, 2000) a technique for testing for lexical specificity in existing samples might be of use.

One way to investigate lexical specificity in existing samples is to compare children's speech with samples of adult data matched for sample size and vocabulary (Aguado-Orea & Pine, in prep.). If children's utterances are significantly more restricted in scope than those of adults, after we have controlled for vocabulary and sample size, then we can argue that the restricted nature of the utterances cannot be attributed to sampling constraints or vocabulary size. We can then tentatively conclude that the data support the constructivist hypothesis that children's early knowledge of grammatical relations may be tied to individual lexical items and frames. If, however, the adult data patterns in much the same way as that of the child, we must conclude that any apparent lexical specificity in the child's data can be attributed only to a restricted vocabulary combined with sampling constraints. The onus will then be on explaining how much of the restricted nature of the data is due to vocabulary constraints and how much due to sampling constraints.

The first aim of the present paper was to investigate the effect of sampling constraints on the scope of one child's early wh-question productions at a particular point in development and to test the constructivist claim that early correct wh-questions are produced by the application of semi-formulaic lexical frames. A rich database of one child's wh-questions was used to investigate the effect that samples of different sizes have on the range and variation of the child's wh-question production. The samples were then compared to a matched sample of maternal speech to establish how much of the apparent lexical specificity in the data could be attributed purely to the restrictions imposed by the sampling method.

Estimating error rates

A key test for any theory of language acquisition is whether it can successfully predict the incidence, rate and pattern of grammatical errors in children's production data. Because errors cannot have been learnt imitatively from adults, they provide insights into the child's grammatical system at a given point in development. Errors of commission are particularly useful in this regard because, unlike omission errors, they are less easily attributed to memory limitations or processing difficulties.

On the basis of sampled naturalistic data, it has often been suggested that many of the types of error we might plausibly expect to see are extremely rare or even nonexistent in children's speech. This has led to the conclusion that children have early adultlike competence in language production. For example, Stromswold (1990) investigated the auxiliary use of 14 children whose data are available on the CHILDES database (MacWhinney, 2000) and concluded that the types of errors we might expect to see (e.g. ungrammatical combinations of auxiliaries) were virtually nonexistent. Valian (1986) presents similar arguments; suggesting, partly on the basis of nonexistent or extremely rare errors, that very young children have a sophisticated knowledge of a range of syntactic constructions.

There are two problems with drawing such conclusions from sampled naturalistic data. First, as Tomasello & Stahl (2004) have demonstrated in a theoretical analysis, common sampling techniques may mean that we are likely to miss periods of high error use or to underestimate error rates. In their analysis, they showed that traditional sampling densities (one hour per week and one hour every two weeks) are unlikely to capture errors that children produce with low frequency (e.g. items that are produced once a day) even if we sample over a long period of time (e.g. a year). They also argued that, even if we capture these structures within our samples, the samples will provide inaccurate estimates of the error rate, because they are likely to either over- or underestimate true frequency of use.

The second problem is that the likelihood of finding errors in small samples is reduced still further because errors are likely to occur most often on low frequency structures. For example, Pine, Lieven & Rowland (1998) found that overall low rates of pronoun case marking error disproportionately reflected children's performance with the high frequent subject pronoun *I*. Rates of case marking errors with the accusative pronoun *me* were much higher, but were not reflected in the overall figure because of the low incidence of *me*. Similarly, Rowland, Pine, Lieven & Theakston (2005) found that overall rates of error in wh-question production were influenced disproportionately by children's ability to use the highly frequent form copula *is* correctly. Error rates with rarer auxiliaries, especially forms of auxiliary DO and modal auxiliaries were much higher (see also the debate between Marcus, Pinker, Ullman, Hollander, Rosen & Xu, 1992, and Maratsos, 2000, for similar issues for past-tense over-regularization errors).

Thus, overall error rates may disproportionately reflect children's ability with high frequency lexical items, hiding any problems they may have with low frequency productions. It is possible, then, that the constraints imposed by sampling, together with the low frequency of certain types of structures, have led to an underestimation of the error rate in children's speech. The second aim of the study was to compare rates of errors in different sized samples of wh-question data from the same child in order to investigate the effect of sample size on error rates.

METHOD

Participant

The participant was Lara, the first-born monolingual English daughter of two white university graduates, who was born and brought up in Nottinghamshire, England. The data are part of a larger corpus of audio-recorded and diary data collected between the ages of 1;9 and 3;3. The data used here were taken from just under one month (23 days) between the ages of 2;8.1 and 2;8;23. Lara's MLU in morphemes was 2·82 at age 2;8.1 and 3·34 at age 2;8.23.

Sampled audio-recorded data

Procedure. Lara was taped for approximately two hours every week. A Marantz CP430 audio-recorder with an external microphone was used for the recording. During recording, Lara engaged in everyday play activities with her regular caregivers (parents and grandparents). For many of the sessions, Lara's younger sibling was present. However, this child was a pre-verbal infant who had little effect on the interaction. No additional investigator was present.

Transcription. The data were orthographically transcribed using the CHILDES system by the second author (MacWhinney, 2000). To ensure transcription accuracy, the transcriber was extensively trained and a detailed set of transcription and coding guidelines was agreed prior to the start of the study. Postcodes were used on the main line to mark utterances that were incomplete, routines, imitations or repetitions. Utterances were considered repetitions or imitations if they were partial or complete repetitions or imitations of an utterance that had occurred five or fewer speaker turns earlier, unless that had been over 10 seconds removed in time. The transcriber was also trained to recognize the types of error made by young children and to note errors with error codes. All transcripts were then checked for accuracy by the first author. The data used here consist of eight hours of audio data produced between the ages of 2;8.1 and 2;8.23. During the period, 3121 interpretable child utterances were recorded, 143 of which were object/adjunct wh-questions.

Diary data

Procedure. The diary data consisted of a written record of the wh-questions that Lara produced from age 2;8.1 to 2;8.23. Lara's caregivers (parents and grandparents) were the diary-keepers. They were provided with notebooks to record the wh-questions produced by Lara within their hearing. Diary-keepers were trained to record the exact speech of the child (e.g. to omit auxiliaries when not pronounced, to indicate contractions) and to recognize different types of wh-question. The caregivers were also given training in the different types of error that children produce and were asked to note if the utterance was an error. If diary-keepers were unclear about the exact form of the question, they were asked to mark this in the diary. All of these questions were excluded from the analysis. No notes were made when the child was at nursery (for 2 part days a week) so it is estimated that the diary contains approximately 80% of the wh-questions that were produced by Lara during this period. The diary was then supplemented by wh-questions that were recorded on the audiotapes that had been omitted from the diary.

It was intended that the diary utterances would be marked for self-repetition. However, this was problematic in practice because of the difficulty of keeping track of the time that had elapsed between two productions of an utterance. Thus, it is likely that the diary contains some material that was repetition.

Transcription. The diary data were orthographically transcribed using the CHILDES system by the first author. The transcription conventions were identical to those used for the audio data. Because of the nature of the data collection, no transcription reliability check was possible. The diary data used here consist of 613 wh-questions produced when Lara was aged 2;8.1 to 2;8.23.

Speech corpora

We extracted all spontaneous, complete, matrix object and adjunct wh-questions from the data. Partially intelligible or incomplete utterances, utterances with parts marked as unclear, quoted utterances and routines (e.g. counting, nursery rhymes and songs) were excluded. Where possible, full or partial repetitions or imitations were also excluded. Subject wh-questions, embedded wh-questions and question fragments were excluded because they were not recorded systematically in the diary data. The first author checked all wh-questions and coded all errors.

Error coding

The questions produced were coded according to the coding scheme outlined in Rowland *et al.* (2005) as follows:

CORRECT QUESTIONS

For questions requiring auxiliary *BE*, *HAVE*, *DO* or modal auxiliaries, correct questions were those in which the wh-word, auxiliary, main verb and subject were correctly chosen and positioned. Correct copula BE questions were those in which the wh-word, copula and subject were correctly chosen and positioned. Questions with omissions and errors not relevant to the grammatical rules that apply specifically to questions (e.g. determiner omission) were included.

OMISSION ERRORS

Auxiliary/copula omission. Errors in which the auxiliary/copula was omitted and tense was not overtly marked on the main verb (e.g. *where he going?*, *where he go?, where that?*).

Subject omission. Errors with omitted subjects (e.g. *where's going?*).

Subject and auxiliary omission. Questions with the auxiliary and subject omitted (e.g. *where going?*).

ERRORS OF INVERSION

Double marking errors. These errors include doubling of the auxiliary/copula (e.g. *where does he does go?*), errors in which tense and agreement were correct but were marked on both auxiliary and main verb (e.g. *where does he goes?*) and errors in which an auxiliary was present but tense and agreement were marked only on the main verb (e.g. *where do he goes?*).

Raising errors. Errors in which the auxiliary was omitted and tense and/or agreement remained on the main verb (e.g. *where he goes?*). These were coded as inversion errors as they indicate that the child has failed to raise TNS and AGR.

Non-inversion errors. Subject auxiliary/copula inversion error (e.g. *where he does go?*).

OTHER ERRORS OF COMMISSION

Agreement errors. Errors in which an auxiliary/copula was present but did not agree with the subject (e.g. *where does you go?*; *where do he go?*).

Case errors. Errors in which the subject had incorrect non-nominative case (e.g. *where's her going?*).

Unclassifiable. Errors in which it was impossible to determine what mistake had been made; for example, *why is the doctor make your tummy better?* would be coded as unclassifiable as it is unclear whether the target is a progressive (*is making*) or present tense (*does make*) construction.

Maternal data

The maternal data consisted of all the spontaneous complete object and adjunct wh-questions produced by Lara's mother during the recordings taken when Lara was 2;10.04 to 3;01.26. Maternal and child data were taken from a different time period to minimize the influence of contextual effects on the results.

RESULTS

Lexical specificity in Lara's data

The first analysis tested whether the use of small samples led to an overestimation of the lexical specificity of the child's data. All correct object and adjunct wh-questions were extracted from the diary sample and the 8-hour audio-sample. Three further smaller sample sizes were then created out of the diary data using a randomizing algorithm. These were designed to estimate to a sampling regime of four hours per month, two hours per month and one hour per month, based on the number of questions produced in the 8-hour sample. Seven samples of each sample size were created in order to provide a measure of variance across a number of samples. This allowed us to judge how accurate any one sample of a particular size was likely to be. For each sample size, each of the seven samples was composed of a different set of utterances in order to ensure that the results could not be attributed to overlap between the samples. In total there were five sample sizes:

- Diary sample – all questions produced in the presence of a caregiver and recorded in the diary plus the wh-questions recorded in the 8-hour audio-sample which had been omitted from the diary – 357 correct object/adjunct wh-questions.
- 8-hour audio-sample (approximating to 2 hours per week) – 101 correct object/adjunct wh-questions.
- Seven estimated 4-hour audio-samples (approximating to one hour a week) – each sample contained 50 wh-questions extracted from the diary sample using a randomizing algorithm.
- Seven estimated 2-hour audio-samples (approximating to 1/2 hour per week) – each sample contained 25 wh-questions extracted from the diary sample using a randomizing algorithm.
- Seven estimated one hour audio-samples (approximating to 1/2 hour every 2 weeks) – each sample contained 12 wh-questions extracted from the diary sample using a randomizing algorithm.

The traditional measure of lexical specificity is to calculate how many of the child's utterances could have been produced simply by the application of the three most frequent frames produced by the child. A frame was defined according to Rowland & Pine (2000): a wh-question frame consists of an entrenched wh-word + auxiliary unit (a pivot; e.g. *what are, where have*), which is combined with a number of lexical items (*variable*) to produce a pivot + variable pattern (e.g. *what are* + *X; where have* + *X*; see Rowland & Pine, 2000 for the rationale behind this definition).

For each sample size, the number of wh-questions that could have been produced by the application of the three most frequent frames was calculated. For example, if the three most frequent frames produced in a 2-hour sample were *what's* + *X*, *where's* + *X* and *where are* + *X*, we would

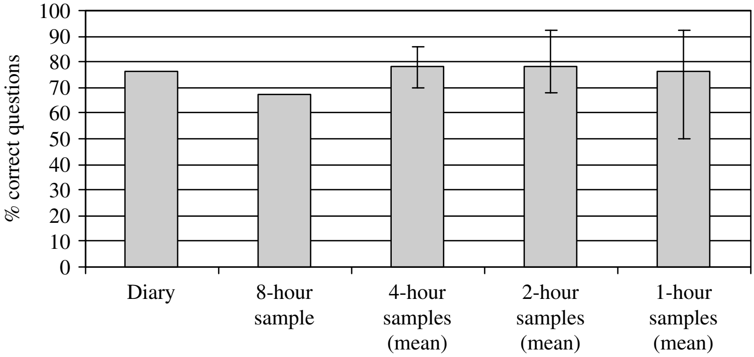

Fig. 1. Percentage of correct questions that can be produced using the three most frequent frames (error bars indicate range).

calculate the proportion of questions that could have been produced using these frames. Figure 1 indicates the results. For the diary and 8-hour sample, the figure shows the percentage of correct questions accounted for by the three most frequent frames. For the estimated 4-hour, 2-hour and one-hour sample sizes, the columns represent the mean percentage across the seven samples and the error bars represent the range across the seven samples.

Seventy-six per cent of the questions in the diary data could have been produced using just three frequent frames. If we take this diary data estimate as the most accurate approximation to the child's speech overall, the results show that even the smallest samples do not grossly distort the amount of lexical specificity in the data. All the sample sizes yielded rates of lexical specificity within 10% of the diary data estimate. More importantly, there is no clear trend of increasing lexical specificity as the sample size reduces.

However, the figures for the three smaller sample sizes illustrate the mean percentage across seven samples. The error bars, which indicate the range, show that there is quite substantial variation between estimates based on individual samples. For the one-hour samples, for example, estimates varied between 50 and 92%. Thus, if we only analysed one of these samples, we are equally likely to get a figure of 92% (a clear overestimate) or a figure of 50% (a clear underestimate) as we are of getting a more accurate estimate.

Thus, although we might not inevitably be overestimating lexical specificity with small samples, any one sample is more likely to give us an inaccurate measure if our sample size is small. One solution is to collect bigger samples. However, another is to assess whether a sample is more lexically specific than we would expect by providing a comparison measure that allows us to control for the effect of sample size.

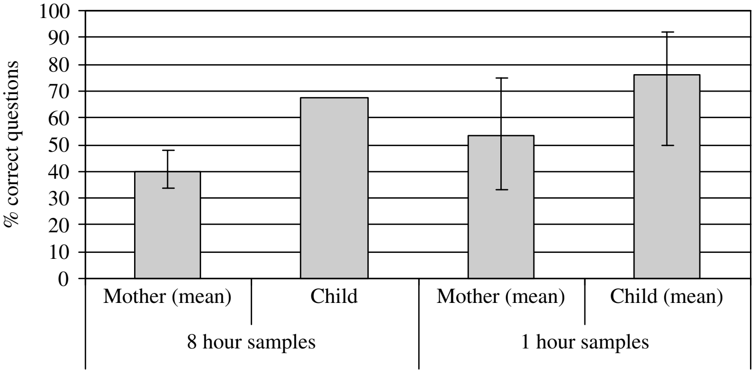

Fig. 2. Percentage of correct questions that can be produced using the three most frequent frames: data for mother and child for 8-hour and one-hour sample sizes (error bars indicate range).

Comparison with adult data

For a comparison measure, maternal data were extracted from the transcripts recorded when Lara was aged 2;10 to 3;01. To control for the fact that Lara's mother knew and used more wh-words and auxiliaries than Lara (i.e. to control for the differences in vocabulary that might be expected to confound the comparison), the maternal sample was confined to questions that included only the wh-words and auxiliaries that the child produced. These were *what, where, which, why*, auxiliary *are* and *is*, copula *are* and *is, can, did, do, does, don't, has, have, shall* and copula *was*. Lara's mother produced 922 relevant questions during this period.

Two comparison analyses were performed. First, we compared the data from Lara's 8-hour transcript with samples of maternal data matched for vocabulary and sample size. Lara produced 101 questions in her 8-hour transcript so for comparison, we analysed seven random samples of 101 questions from the mother's data. A second comparison attempted to establish the reliability of the smallest sample size (one-hour). We compared Lara's one-hour samples with samples of maternal data. We have estimated that Lara produced approximately 12 questions in one hour, so we analysed seven random samples of 12 wh-questions from the mother's data. For both comparisons, each of the seven maternal samples was composed of a different set of utterances in order to ensure that the results could not be attributed to overlap between the samples.

Figure 2 demonstrates the results. For the mother, the columns indicate the mean percentage of questions accounted for by the three most frequent frames and error bars indicate the range across the seven samples. For the child, the columns indicate the mean across seven samples for the one-hour sample size (error bars indicate range) but the raw percentage for the 8-hour sample.

Taking the data for the 8-hour sample size first, the figure shows that even when the sample size and vocabulary were controlled, the mother's data were much less lexically specific than the child's on all relevant criteria. For the child, the three most frequent frames accounted for 67% of the data. This figure was over two standard deviations above the mean across the matched maternal samples (mother's mean + 2 S.D. = 49%), it was well outside the range from the maternal samples (34–48%), and it was well above the upper bound of the 95% confidence interval (45%).

It was also the case that the child used significantly fewer frames than the mother even after we restricted the mother's data to only wh-words and auxiliaries that the child knew (child: 19 frames, mother mean = 26·14 frames). In summary, the child's speech was much more restricted in scope than the mother's, even when we controlled for vocabulary and sample size. Thus, we can conclude that the restricted nature of the child's speech cannot be attributed to sampling constraints or to a limited vocabulary of wh-words and auxiliaries.

For the one-hour samples, the three most frequent wh + aux combinations accounted for a mean of 76% of the child's wh-questions. This is much larger than the mean for the mother in the matched one-hour sample (54%). However, we cannot say that this estimate is reliably higher than the estimate from the maternal data. It is well within two standard deviations of the mother's mean (mother's mean + 2 S.D. = 85·27%) and there is substantial overlap in the ranges of mother and child (child range = 50–92%; mother range = 33–75%). Both child and mother also produce very similar numbers of frames (child mean no frames = 5·70, mother mean = 8·43). We would be very reluctant to argue on the basis of this data that the child's knowledge was based around lexically specific frames.

Analysis of error rates

The first analysis investigated whether overall error rates misrepresented error rates in low frequency parts of the system. For this analysis, all object and adjunct wh-questions – both correct and errors – were extracted from the diary data and divided according to auxiliary type: copula *BE* (e.g. is, are), auxiliary *BE* (e.g. is, are), auxiliary *HAVE* (e.g. *has, have*) and DO/ modals (e.g. *do, does, can, will*). Figure 3 shows, for each auxiliary type, the percentage of questions that were correct, omission errors, inversion errors and other commission errors in Lara's diary data (see method section for the definition of errors).

Error rates varied substantially across auxiliary type. For example, rates of omission error in questions with copula *BE* were substantially lower than

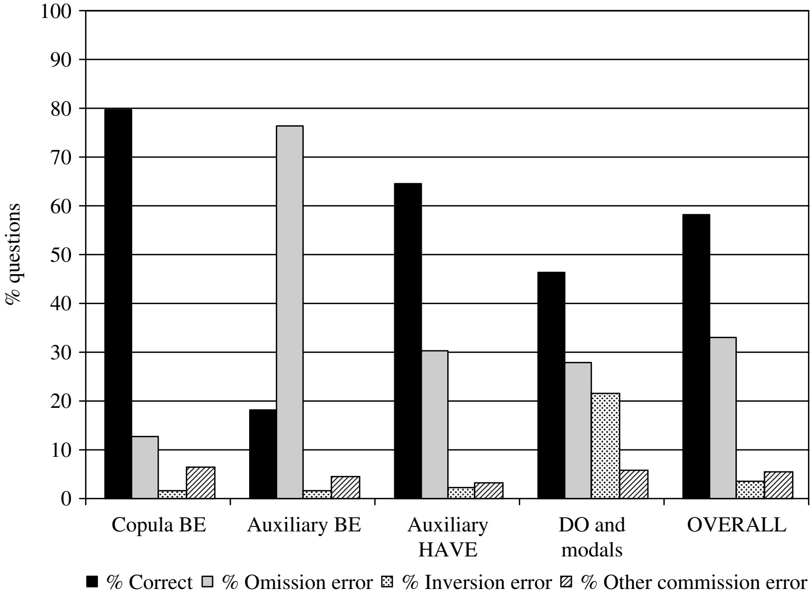

Fig. 3. Percentage of questions that were correct and errors in Lara's diary data.

the overall error rate whereas omission error rates on auxiliary *BE* were substantially higher. It is clear that overall error rates do not accurately predict error rates across the board.

The discrepancy is particularly salient for inversion errors in questions with DO/modals. Inversion error rates were low in the overall data (3·43%), which might lead us to conclude that such errors are rare. However, inversion errors were extremely frequent in questions with DO/modals (20·37% of questions: six times higher then the overall rate). Thus, the overall error rate provides a very misleading impression of the incidence of error in questions with DO/modals. This is because questions with DO/modals were relatively rare (accounting for only 9% of Lara's questions overall), so contributed very little to the overall error rate. In fact, the overall rate disproportionately reflects how good Lara was at forming questions with copula *BE* (which made up 46% of her questions).

Since overall error rates may misrepresent error rates on low frequency structures it is important to analyse these structures separately. However, estimates based on small amounts of data can provide a very inaccurate picture of the pattern of acquisition. Questions with DO/modals are of low frequency, so it is likely that small samples will provide unreliable estimates of error rates on these structures. The final analysis investigated the effectiveness of small samples at capturing reliable rates of inversion error in low frequency questions.

Three different sample sizes were created out of the diary data – 4-hour, 2-hour and one-hour sample sizes. Sample sizes were estimated based on the number of questions produced in the 8-hour sample. Seven samples of each sample size were created in order to provide a measure of variance. Each of the seven samples was composed of a different set of utterances to ensure that the results could not be attributed to overlap between the samples. In all, there were five sample sizes:

- Diary sample – all questions produced in the presence of a caregiver plus the wh-questions recorded in the 8-hour audio-sample that had been omitted from the diary – 613 object/adjunct wh-questions (357 correct, 256 errors).
- 8-hour audio-sample – 143 object/adjunct wh-questions (101 correct, 42 errors).
- Seven estimated 4-hour audio-samples – each sample contained 72 object/adjunct wh-questions extracted from the diary data using a randomizing algorithm.
- Seven estimated 2-hour audio-samples – each sample contained 36 object/adjunct wh-questions extracted from the diary data using a randomizing algorithm.
- Seven estimated one-hour audio-samples – each sample contained 18 object/adjunct wh-questions extracted from the diary data using a randomizing algorithm.

For each sample size, we calculated the percentage inversion error rate for questions with DO/modal auxiliaries (the least frequent question type) and for questions with copula *BE* (the most frequent question type). Figure 4 demonstrates the results. For the diary and 8-hour sample, the figure shows the percentage of questions that were inversion errors. For the estimated 4-hour, 2-hour and one-hour sample sizes, the columns represent the mean error rate across the seven samples and the error bars represent the range.

The diary data demonstrates that 1·43% of questions with copula *BE* and 20% of question with DO/modal auxiliaries were inversion errors, which we take as the most accurate approximation of the child's speech overall. In comparison, the smaller sample sizes estimated the error rate for copula *BE* quite accurately. However, copula *BE* questions were produced relatively frequently, accounting for about half of the questions produced. For questions with DO/modal auxiliaries, although the mean error rate calculated across seven samples was often quite accurate (means: 4-hour sample = 26%, 2-hour sample = 17%, one-hour sample = 26%) estimates from individual samples varied substantially. For both 2-hour and one-hour sample sizes, error rates varied from 0 to 100%. Four of the one-hour samples showed a 0% error

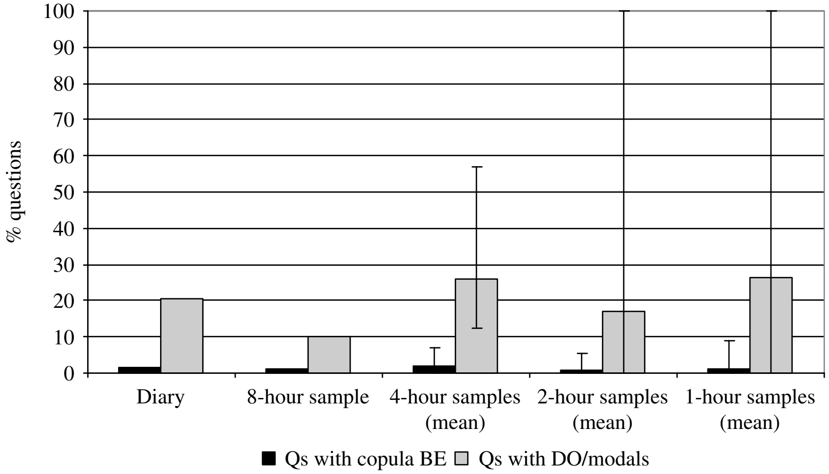

Fig. 4. Percentage of questions with copula *BE* and auxiliary DO/modal auxiliaries that were inversion errors (error bars indicate range).

rate, one showed a 33% error rate, one a 50% rate and one produced a 100% error rate. Even some of the 4-hour samples inaccurately estimated the error rate (range = 12–57%). Thus, although the error rate on questions with DO/modals is actually quite high (20% from the diary data) we are very unlikely to capture anything close to this figure in even quite large individual samples.

The variance is purely due to the small numbers of questions produced. In real terms, the only difference between the samples that showed no error rates and those that showed high error rates was the inclusion or exclusion of one or two inversion errors, but there were so few examples of these questions overall that this made a big difference to error rates. In conclusion, small samples are extremely unreliable at estimating error rates in utterance types that occur relatively infrequently but that tend to have high rates of error.

DISCUSSION

The first aim of the present study was to investigate whether small samples overestimate the amount of lexical specificity in children's speech. The results showed that samples do not inevitably overestimate lexical specificity but that the variation around the mean increases as sample size decreases. The smaller the sample size, the more likely it is that any one sample will either under- or overestimate the lexical specificity in the child's speech. This means that to reliably test whether a child's speech is lexically specific, we need to compare the data with a matched sample of data from a speaker whose grammatical knowledge is adultlike (e.g. the child's mother). One way to do this accurately and minimise the chances of either over- or underestimating the gap between adult and child is to take a number of random samples of adult data and compute confidence intervals, standard deviations, ranges and means across these samples. If the proportion of data accounted for by lexical frames in the child's sample falls comfortably outside the figures calculated for the adult sample, we can conclude that the child's data is more lexically specific than we would expect given the constraints imposed by sampling. If no substantial difference is found, it is not possible to conclude that the child's speech seems lexically specific, even if the amount of data explained by a few frequent frames is large.

When we applied controls for vocabulary and sample size and compared mother and child speech, it was clear that Lara's one-hour samples did not provide any evidence for lexical specificity in her wh-questions. However, this conclusion resulted from a restricted sample size. Once we increased the amount of data (to an 8-hour sample) we found that the child's data was significantly more restricted in scope than the mother's data, even when the samples were matched for vocabulary and sample size. Thus, with big enough samples and the correct controls, it is possible to conclude that there is evidence for lexical specificity in children's early questions.

The second aim of the present paper was to investigate the effect of sampling on error rates. First, the results demonstrated that it is important to treat claims that children's errors are rare or nonexistent with caution if they are based on low numbers of errors in sampled data. Absolute numbers of errors may be small and still correspond to high error rates. For example, Lara produced only 11 inversion errors in questions with auxiliary DO and modal auxiliaries. However, because she produced only 54 questions with DO/modals in total, error rates were high (approximately 2 out of every 10 questions produced with DO or modal auxiliaries contained an inversion error).

Second, the results demonstrated that it is not sufficient to look simply at overall error rates, because these disproportionately represent children's ability to produce the types of question they use often. In the case of wh-questions, most of children's correct questions occurred with contracted forms of copula *is*, auxiliary *is* and auxiliary *has*. Inversion error rates were much higher in questions with DO/modal auxiliaries but these were not reflected in overall error rates because the auxiliaries occurred in much smaller numbers.

Third, once we acknowledge that error rates have to be looked at more closely, it becomes clear that small samples of data are unreliable when it comes to calculating error rates on infrequently produced utterances. It is not simply the case that errors will be underestimated; small samples can also substantially overestimate the error rate. Thus, it is possible that some of the apparent contradictions in error rates reported in the literature may be due simply to sampling constraints.

For example, there are a number of suggestions in the literature that different children may show very different patterns of error use in question acquisition. Erreich (1984) found two different categories of children: children who produced both inverted and non-inverted forms in both yes–no and wh-questions and children who produced only non-inverted yes–no but both inverted and non-inverted wh-questions. Van Valin (2002) has argued for three different types of children: children who show inversion in both yes–no and wh-questions from the start (like those discussed in Ingram & Tyack, 1979), children who produce inverted yes–no questions but non-inverted wh-questions at the start (like those discussed by Labov & Labov, 1978), and children who produce inverted yes–no but both inverted and non-inverted wh-questions (like five of the children studied by Erreich, 1984). Finally, Valian, Lasser & Mandelbaum (1992) found that children inverted more consistently in wh-questions then in yes–no questions, contrary to Bellugi's (1971) conclusion that children invert more consistently in yes–no than wh-questions. Many of these differences may be due to different data collection techniques, to the inclusion of elicitation tasks in some studies or to the inclusion or exclusion of different auxiliary types. However, some of the differences, especially those reported within a particular study may arise simply as a result of the interaction of sampling constraints with small samples of data. It may be that with enough data, it will become evident that children are following very similar routes into language.

The fact that small samples of data are unreliable when assessing error rates in low frequency structures is not a new finding. In the literature on the past tense over-regularization error, Marcus *et al.* (1992) have argued that estimates of error rates based on low frequency structures in small samples are unreliable, which is the justification for their focus on overall error rates. However, as we have found (and as Maratsos, 2000, has suggested), overall error rates are also misleading, often leading to an underestimation of error rate in low frequency structures.

A solution is to use statistical techniques to determine what size samples are required to capture reliable error rates in structures of differing frequency (see Tomasello & Stahl, 2004, for some suggestions). Another solution is to collect dense databases like those currently being collected by Lieven and her colleagues (see e.g. Lieven, Behrens & Spears, 2003; Maslen, Theakston, Lieven & Tomasello, 2004). A less time-consuming solution might be to analyse data from a number of small samples. The mean error rate over a number of samples is likely to be reliable, even if the samples are relatively small. Even for our one-hour sample size, the estimate based on the mean across seven samples yielded a reliable measure of inversion error rates in questions with DO and modal auxiliaries. These samples do not necessarily need to be from the same child. Analysing data from a number of children will provide a reliable estimate of the error rate across those children. However, this technique must be used cautiously because each small sample is still likely to over- or underestimate the error rate substantially. In these cases it may be important to look at variance across children. If variance is low then individual samples can be considered reliable. However, if variance is high, then the data either reflect large individual differences or indicate that individual error rates are misleading. Either way, the individual error rates must be treated cautiously and cannot necessarily be used as the basis of effective arguments about the scope and nature of errors in children's early multi-word productions.

REFERENCES

Aguado-Orea, J. J. & Pine, J. M. (in prep.). Assessing the early productive use of Spanish verbs: implications for current theories of the acquisition of grammar.

Akhtar, N. & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology* **33**, 952–65.

Bellugi, U. (1971). Simplification in children's language. In R. Huxley & E. Ingram (eds), *Language acquisition: models and methods*. London: Academic Press.

Braine, M. D. S. (1976). Children's first word combinations. *Monographs of the Society for Research in Child Development* **41** (1, Serial No. 164).

Brown, R. (1973). *A first language: the early stages*. Cambridge, MA: Harvard University Press.

Dąbrowska, E. (2000). From formula to schema: the acquisition of English questions. *Cognitive Linguistics* **11**, 83–102.

Erreich, A. (1984). Learning how to ask: patterns of inversion in yes–no and wh-questions. *Journal of Child Language* **11**, 579–602.

Fletcher, P. (1985). *A child's learning of English*. Oxford: Blackwell.

Ingram, D. and Tyack, D. (1979). Inversion of subject NP and Aux in children's questions. *Journal of Psycholinguistic Research* **8**, 333–41.

Labov, W. & Labov, T. (1978). Learning the syntax of questions. In R. Campbell & P. Smith (eds), *Recent advances in the psychology of language*. New York: Plenum Press.

Lieven, E. V. M., Behrens, H., Speares, J. & Tomasello, M. (2003). Early syntactic creativity: a usage-based approach. *Journal of Child Language* **30**, 333–70.

Lieven, E. V. M., Pine, J. M. & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language* **24**, 187–219.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J. & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development* **57** (4, Serial No. 228).

Malvern, D. D. & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (eds), *Evolving models of language*. Clevedon: Multilingual Matters.

Maslen, R., Theakston, A. L., Lieven, E. V. M. & Tomasello, M. (2004). A dense corpus study of past tense and plural over-regularization in English. *Journal of Speech, Language and Hearing Research* **47**, 1319–33.

MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk*. Mahwah, NJ: Erlbaum.

Maratsos, M. (2000). More over-regularizations after all: new data and discussion on Marcus, Pinker, Ullman, Hollander, Rosen & Xu. *Journal of Child Language* **27**, 183–212.

Naigles, L. (2002). Form is easy, meaning is hard: resolving a paradox in early child language. *Cognition* **86**, 157–99.

Pine, J. M. & Lieven, E. V. M. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics* **18**, 123–38.

Pine, J. M., Lieven, E. V. M. & Rowland, C. F. (1998). Comparing different models of the development of the English verb category. *Linguistics* **36**, 807–30.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Radford, A. (1990). *Syntactic theory and the acquisition of English syntax: the nature of early child grammars of English*. Oxford: Blackwell.

Rispoli, M. (1998). Patterns of pronoun case error. *Journal of Child Language* **25**, 533–54.

Rowland, C. F. & Pine, J. M. (2000). Subject–auxiliary inversion errors and wh-question acquisition: what children do know? *Journal of Child Language* **27**, 157–81.

Rowland, C. F., Pine, J. M., Lieven, E. V. M. & Theakston, A. L. (2005). The incidence of error in young children's wh-questions. *Journal of Speech, Language and Hearing Research* **48**, 384–404.

Stromswold, K. (1990). *Learnability and the acquisition of auxiliaries*. Dissertation, MIT (Distributed by MIT Working Papers in Linguistics).

Tomasello, M. (1992). *First verbs: a case study of early grammatical development*. Cambridge: C.U.P.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition* **74**, 209–53.

Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, M. & Stahl, D. (2004). Sampling children's spontaneous speech: how much is enough? *Journal of Child Language* **31**, 101–21.

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology* **22**, 562–79.

Valian, V., Lasser, I. & Mandelbaum, D. (1992). *Children's early questions*. Manuscript, Hunter College, New York.

Van Valin, R. D. (2002). The development of subject–auxiliary inversion in English wh-questions: An alternative analysis. *Journal of Child Language* **29**, 161–75.