*J. Child Lang.* **31** (2004), 101–121. 
DOI: 10.1017/S0305000903005944 Printed in the United Kingdom

# Sampling children's spontaneous speech: how much is enough?*

MICHAEL TOMASELLO AND DANIEL STAHL
*Max Planck Institute for Evolutionary Anthropology*

(*Received* 4 *March* 2003. *Revised* 19 *June* 2003)

ABSTRACT

There has been relatively little discussion in the field of child language acquisition about how best to sample from children's spontaneous speech, particularly with regard to quantitative issues. Here we provide quantitative information designed to help researchers make decisions about how best to sample children's speech for particular research questions (and/or how confident to be in existing analyses). We report theoretical analyses in which the major parameters are: (1) the frequency with which a phenomenon occurs in the real world, and (2) the temporal density with which a researcher samples the child's speech. We look at the influence of these two parameters in using spontaneous speech samples to estimate such things as: (a) the percentage of the real phenomenon actually captured, (b) the probability of capturing at least one target in any given sample, (c) the confidence we can have in estimating the frequency of occurrence of a target from a given sample, and (d) the estimated age of emergence of a target structure. In addition, we also report two empirical analyses of relatively infrequent child language phenomena, in which we sample in different ways from a relatively dense corpus (two children aged 2;0 to 3;0) and compare the different results obtained. Implications of these results for various issues in the study of child language acquisition are discussed.

## INTRODUCTION

A primary research methodology in the study of child language acquisition is naturalistic observation. In the classic method, parents keep a diary of their child's language production using one of several different sampling techniques. This yields a very broad and rich picture of one child's

[*] For their helpful comments we would like to thank the following people: Elena Lieven, Julian Pine, Gina Conti-Ramsden, Anna Theakston, Heike Behrens, and Caroline Rowland. Address for correspondence: Michael Tomasello, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany. tel: +49 341 3550 400. e-mail: tomas@eva.mpg.de

language, but of course the diary method is by its very nature highly selective (see, e.g. the discussions of Braunwald & Brislin, 1979; Mervis, Mervis, Johnson & Bertrand, 1992). For this and other reasons, the main form of naturalistic observation in the modern study of child language acquisition is audio and/or video recordings of children's spontaneous linguistic interactions with a parent or other interlocutor. This is a much more systematic method of observation than diary keeping, but now different issues of sampling come to the fore. For example, there has always been some concern that children and parents do not talk as they normally do when researchers are present with their recorders turned on – typically in one room with toys for a half-hour or hour. For that reason there have been several major child language projects in which children's speech has been sampled in a wider variety of naturalistic settings (e.g. Hall, Nagy & Linn, 1984; Wells, 1985).

But, perhaps surprisingly, there has been very little discussion in the field of the quantitative aspects of child language sampling, that is, how much to sample and at what intervals and for how long and for how many children. This is in contrast to other scientific fields in which naturalistic observation is especially important. For example, in the study of animal behaviour, much attention is paid to the issues introduced by Altman (1974; see also Martin & Bateson, 1986), who systematically weighs the advantages and disadvantages of such things as focal animal sampling, scan sampling, *ad libitum* sampling, various schemes of time sampling, and so forth and so on. In the field of child language acquisition, the vast majority of samples of children's spontaneous speech, in many different languages, have been collected following the lead of Roger Brown and colleagues, and many of theses are on file in the CHILDES database (MacWhinney & Snow, 1985). Typically, several children are observed one hour every one to two weeks for a year or more. In terms of quantity, assuming that a child is awake and talking roughly 10 hours/day, this represents something like 1–1·5% of the language a given child hears and produces during the sampling period. Is this enough?

The answer to this methodological question obviously depends on the research question. For high-frequency phenomena, for instance, children's use of copulas or pronouns in English, the typical samples used in the study of child language are no doubt adequate – at least for some kinds of analyses. But recently there have been prominent discussions of some phenomena that occur with relatively low frequency, and for these cases such sparse sampling is almost certainly not adequate. For example, in the Marcus, Pinker, Ullman, Hollander, Rosen & Xu (1992) study of English-speaking children's past tense overgeneralization errors, issues of frequency and sampling were crucial. Just to give one example, Marcus *et al.* decided, for perfectly good reasons, not to include in their main analyses past tense overgeneralizations that occurred very rarely in their samples (i.e. they excluded all verbs that occurred in the past tense less than 10 times for a given child). Since the lower frequency verbs were the ones that were overgeneralized most often, this procedure almost certainly led to an underestimation of error rate (Maratsos, 2000). Maratsos (2000) also points out more generally that, given the 1–2% samples, each error observed by Marcus *et al.* presumably represents something that the child does more than 50 times in the real world. The low numbers also led Marcus *et al.* in some cases to sum observed errors across many months, potentially obscuring developmental effects.

Another phenomenon for which this same issue has arisen with special urgency is so-called optional infinitives, especially in English. The problem is that some researchers have based significant theoretical claims on the relative rarity of child errors with such things as the third person *-s* agreement marker. For example, Rice, Wexler, Marquis & Hershberger (2000) argue that the very few errors they observed were so infrequent that they could be disregarded as noise in the data. The problem is that children do not have occasion to use the third person *-s* agreement marker very often, especially not with lexical verbs, and so the few observed errors actually represent a fairly high percentage, in some cases, of the opportunities the child had to make the error (Pine, Rowland, Lieven & Theakston, 2001). The general lesson here, then, is simply that the combination of an infrequent phenomenon and sparse sampling means that frequency estimates of all kinds must perforce be highly unreliable.

Another place where issues of sampling are especially important is in estimating such things as vocabulary size or the age of emergence of some linguistic item or structure. For example, for the question of whether children first learn nouns or verbs, it has been pointed out that children use each of their verbs more frequently than they use each of their nouns (Gentner, 1982; Tardiff, Gelman & Xu, 1999). This means that spontaneous speech samples of 1–2% are more likely to capture each verb than each noun, and so the two estimates are not really comparable (and so some have argued for maternal report as a fairer measure; Caselli, Bates, Casadio, Fenson, Fenson, Sanderl & Weir, 1995; Caselli, Casadio & Bates, 1999). Similarly, quite often researchers want to compare the age of emergence of two related structures – for example, ditransitive datives and prepositional datives – that occur with different frequencies. But it takes only a moment's reflection to see that with periodic sampling a frequently occurring construction will, on average, be detected at a time point closer to its 'real' first emergence than will a less frequently occurring construction. Therefore, the age of emergence of two linguistic structures can only be compared using periodic sampling if they occur with close to the same frequencies in the real world – and of course the same issue arises if we compare the age of emergence of a given structure for different children who use that structure with different frequencies.

Obviously, everyone knows that more is better, and there are very good practical reasons for not sampling too often. The limiting factor – as all linguists and psycholinguists know all too well – is transcription time, estimated by most researchers to represent between 10 and 20 hours per hour of speech sampled (if we are not especially concerned with phonetic accuracy). The Brown-type method represents an excellent compromise for, for example, establishing the first corpora of child speech in a previously undocumented language. It allows the researcher to sample several children over a several year period and still be able to report results in a timely manner. But the field has progressed to a point where we should perhaps begin thinking more systematically about different sampling techniques for specific problems. Thus, such things as tense and agreement errors in English occur most frequently for most children during a somewhat limited period, say one year. This means that for the same amount of transcription time a researcher could sample several children at a much denser rate for a shorter time – or even one child for a short time with even denser sampling intervals – and obtain a much better developmental picture of this phenomenon. Of course one practical issue is that the alternative provided by the CHILDES databases is zero transcription time, since those samples have already been transcribed, and so one may address a question immediately rather than several years down the line. Nevertheless, we would argue that for some low frequency phenomena the majority of CHILDES-like samples are not dense enough to support valid and reliable analyses.

Our goal in the current paper is to provide quantitative information that might help researchers make decisions about how to sample children's speech for particular research questions. We report several theoretical analyses in which the major parameters are: (1) the frequency with which a phenomenon occurs in the real world, and (2) the temporal density with which a researcher samples the child's speech. We look at the influence of these two parameters in using spontaneous speech samples to estimate such things as: (a) the percentage of the real phenomenon (targets) actually captured, (b) the probability of capturing at least one target in any given sample (hit rate or power), (c) the confidence we can have in estimating the frequency of occurrence of a target from a given sample, and (d) the estimated age of emergence of a target structure. In addition, we also report two empirical analyses of relatively infrequent phenomena (English past tense overgeneralization errors and German passives in the 2;0 to 3;0 period), in which we sample in different ways from a relatively dense corpus and compare the different results obtained. Attempting to be fairly practical, in all cases we are aiming to help researchers with two related questions, depending on whether they do or do not already have a sample:

(1) Given my question (and resources), how should I sample?
(2) Given my sample, how confident should I be in my results?

## THEORETICAL ANALYSES

In the analyses that follow we assume that a normal child is awake and talking 10 hours/day (70 hours/week), and that a given language sample is representative of the language used by the child during non-sampled times. We assume further that any given target structure of interest occurs at random intervals in the child's speech, with each occurrence independent of the others. This latter assumption is clearly not wholly valid, as children may produce particular linguistic structures in clumps in discourse in ways that are dependent on one another. But because we have no information on exactly how this interdependence manifests itself in children's production of target structures, we assume independence and randomness – in part to make statistical treatment more straightforward. These assumptions should not affect the substance of any of our conclusions, and indeed we will provide a small empirical test below.

Because the following analyses are intended to be used as illustrations only, we have chosen the following values for our two most important parameters. First, we investigate target structures that might hypothetically occur at the following rates in the real world:

- 7 occurrences/week (1 occurrence/day)
- 14 occurrences/week (2 occurrences/day)
- 35 occurrences/week (5 occurrences/day)
- 70 occurrences/week (10 occurrences/day)

Second, in terms of sample densities, we have chosen four: the two most frequently used in child language research (0·5 and 1 hour/week) and in addition two others that we have used in some of our own recent research (e.g. Lieven, Behrens, Speares & Tomasello, in press). These are:

- 0·5 hour/week (i.e. one hour biweekly)
- 1 hour/week
- 5 hours/week
- 10 hours/week

Most of the analyses below use these values, and in some cases a few additional ones, to assess the quality of various sampling procedures.

### *Number of targets captured*

The first and most straightforward analysis uses simple arithmetic to estimate the number and/or proportion of targets we might capture using

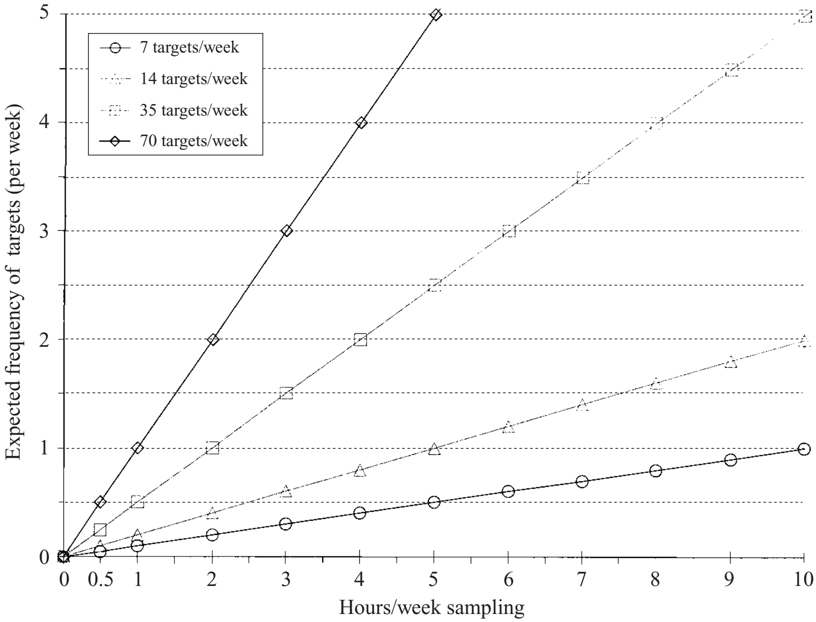


Fig. 1. Expected number of targets captured per week as a function of rate of occurrence and sample density.

various sample densities – for targets that occur in the real world at various different rates. Figure 1 presents this analysis. As an example, if we sample one hour/week and the target occurs 70 times/week, then we expect to capture, on average, one target in that weekly hour. This represents, obviously, 1/70th (1·4%) of all targets occurring in the real world during the one-week sampling period. To be fairly certain to capture one instance of a less frequently occurring target, for instance, one that the child produces only 7 times/week, we need to sample much more frequently – approximately 10 hours/week (as impractical as that might be). If we focus on the two sample densities most often used in modern research – 0·5 and 1 hour/week – we can see in Figure 1 that in every case these yield very low estimated weekly capture rates (0·5 targets or less) for targets the child produces 35 or fewer times per week.

Another approach to the question of capture rate is to simulate the number of targets one would capture on a weekly basis over an entire year using different sampling schemes (and for different rates of occurrence). Our procedure was as follows. First we generated random numbers with an underlying Poisson distribution, with these random numbers representing the day and hour when a production might occur – then the sampling was

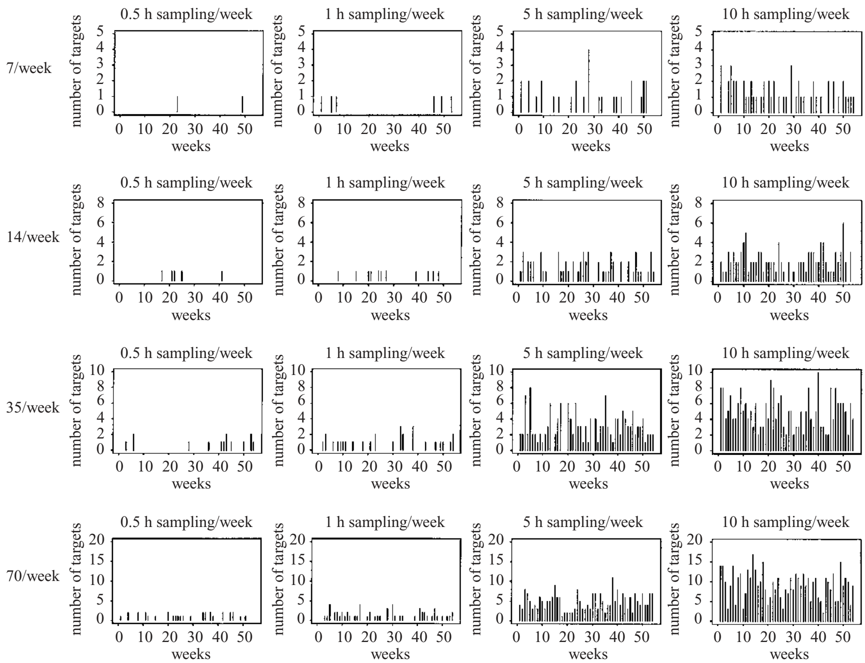


Fig. 2. Number of targets captured in a random number simulation as a function of rate of occurrence and sample density.

done randomly as well, at the rate specified. The Poisson distribution is a discrete distribution used to model the number of events occurring in some unit of time (or space), and it is mainly used if the occurrence of events is rare. It assumes that each event occurs independently of the others and at random. The Poisson distribution is characterized entirely by one parameter $\lambda$, the mean (as mean and variances are equal). In the current case, $\lambda$ was calculated by:

$$\lambda = \frac{\text{sample density (hours/week)}}{\text{hours talking per week (i.e. 70)}} \times \text{number of targets/week}$$

Using a specified rate, we simulated the number of targets one would capture each week for a one-year period under the different sampling schemes and rates of occurrence. Figure 2 illustrates the outcome of these simulations (just one per diagram – other simulations with the specified parameters would yield different outcomes for each diagram, of course). The problem with capturing low frequency targets may be illustrated most dramatically by focusing on the lowest rate of occurrence: 7 times/week (top row in Figure 2). We see that for the 0·5 hour/week sampling scheme we do not pick up the first target until halfway through the year, and only pick up 2 for the entire year. Note that this is not due to some quirk of the simulation procedure, as simple arithmetic tells us that of the 364 targets occurring during the year (7/week for 52 weeks) a 0·007 sample (0·5 hours out of 70/week) predicts that about 2·5 targets should be picked up. The 1 hour/week sampling scheme is of course twice as effective, picking up 6 targets (expected = 5·1), but simply due to chance 3 targets are picked up in the first 10 weeks and 3 in the last 10 weeks, with none being picked up in the 8 months during the middle of the year. Obviously, in a real study this would lead to some erroneous inferences about child skills. The 5 hours/week sampling scheme is clearly much better. However, one can still see some fairly major inconsistencies, for example, in some weeks as many as 4 of the 7 targets are captured whereas in other weeks none are captured. Indeed, in the majority of weeks no targets are captured. Finally, the 10 hours/week sampling scheme begins to look pretty consistent, with only about one-third of the weeks yielding no captures and no week yielding more than 3 captures.

Examining the 14/week, 35/week, and 70/week rates of occurrence (second, third, and fourth rows of Figure 2) also demonstrates the limitations of the 0·5 and 1 hour/week sampling regimes. For example, let us focus on the very best case of all of these, the 70 target occurrences per week, and let us do this when sampling is at the commonly used rate of one hour/week (second graph in bottom row). Using the formula for the Poisson distribution from above we calculated the expected probabilities for 0, 1, 2, 3, 4 and 5+ targets. Figure 3 is a histogram of the expected probabilities of capturing these numbers of targets under the specified conditions. In this figure we see that with one hour/week sampling and a target that occurs 70 times/week, we can expect to capture during our weekly sample 0 targets 37% of the time, 1 target 37% of the time, 2 targets 18% of the time, 3 targets 5% of the time, 4 targets 2% of the time, and 5 or more of the 70 weekly targets less than 1% of the time. And so even when the child is producing something each and every hour of the day, seven days a week, a one hour/week sample will miss all of them more than a third of the weeks and will virtually never catch more than 2–3% of them in any given week.

### *Estimating frequency*

There are many different ways to estimate the frequency with which a target occurs. As just one illustration, we look at the process of frequency estimation if we wish to know how many times during a one-week period a given target occurs – assuming a constant weekly rate throughout an observation period, e.g. 4 weeks. To estimate the weekly frequency during this 4 week study period for a given sampling scheme and target occurrence

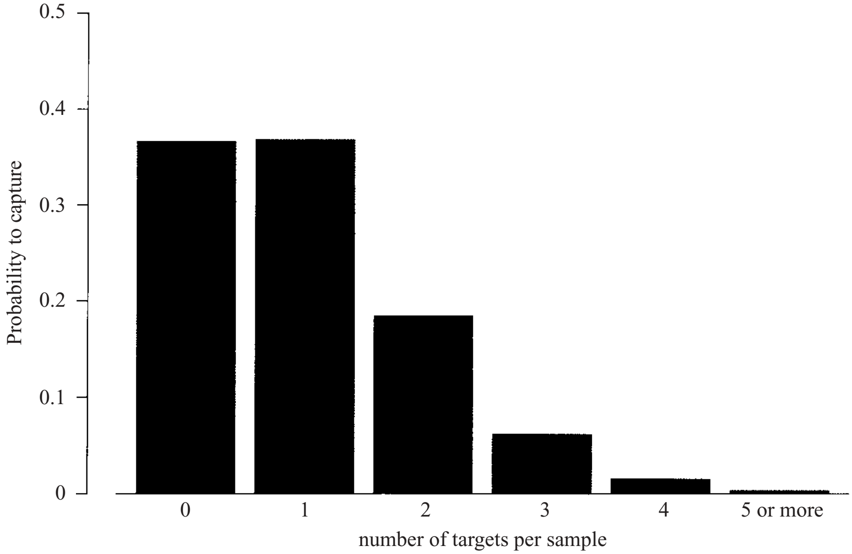


Fig. 3. Probability density of poisson distribution with a mean of 1 (1 hour sampling 70 targets/week).

rate we again simulated the number of targets for each week. The average of the four samplings was then used as an estimate for the weekly frequency.

We used Monte Carlo simulation methods to calculate medians (middle value, with equal numbers higher and lower) and confidence intervals of the estimated weekly frequencies (Manly, 1998). By assuming an underlying Poisson distribution with a $\lambda$ of $x$ (mean frequency of targets/week) we generated 1000 random samples and determined the median and 95% confidence intervals for the different sampling schemes and rates of target occurrence. Because we estimate weekly frequency from a 4 week observation period, we sampled in each simulation 4 times, then added up the 4 samples to obtain one sample for each of the 1000 simulations, and each sample was then divided by 4 to obtain an estimate for a mean weekly rate of occurrence.

The 95% confidence interval of a Monte Carlo sample is given by the value that falls below 2·5% of the sorted (ordered) simulated data and the value that exceeds 97·5% of these data (percentile confidence interval method of Efron, 1979). The analysis of all 1000 samples for each frequency/density combination is shown in Figure 4. This figure presents the median values and 95% confidence intervals for estimating one-week frequency using this Monte Carlo technique – under different sampling schemes and rates of target occurrence. Thus, we can see that for a rate of occurrence of 7 targets/week, the two least dense sampling techniques yield median

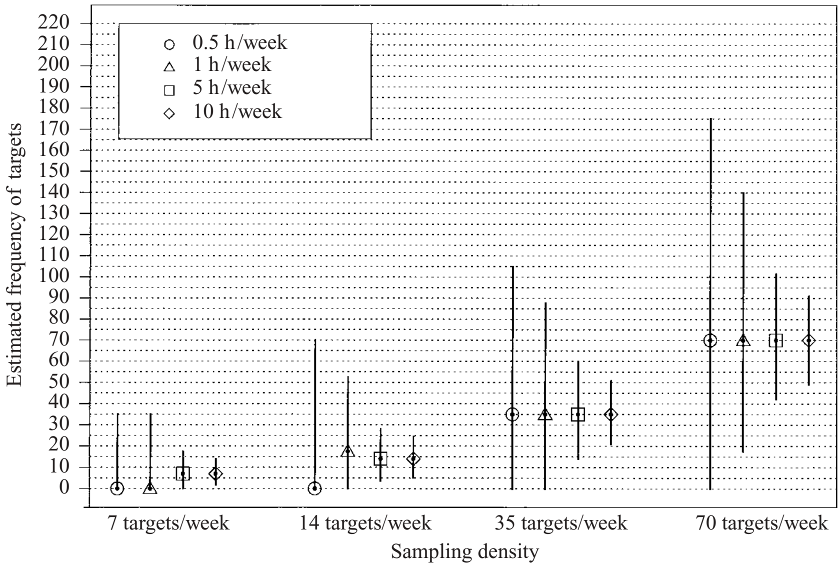


Fig. 4. Estimated weekly frequency of target (median and 95% confidence intervals) as a function of rate of occurrence and sample density.

estimates of zero; the two denser sampling techniques yield reasonable predictions with reasonable confidence intervals. For a rate of occurrence of 14 targets/week, the least dense sampling technique (0·5 hour/week) again yields a median estimate of zero, whereas the other sampling techniques yield reasonable predictions (but with very large confidence intervals in the case of 1 hour/week sampling). The two highest rates of occurrence (35 and 70 times/week) yield fairly stable median estimates under all sampling techniques, although the confidence intervals are quite large for the two least dense samples.

As noted previously, the assumption of independent and random occurrences is perhaps not realistic in the case of child language, as children may produce target structures in a nonrandom manner temporally. But it is theoretically not the case that changing these assumptions – for example, assuming that children produce targets in specific kinds of temporal clumps – would improve the picture if an identical sampling scheme were used. Figure 5 presents the same analysis presented above, but when the occurrences of the targets are clumped together in time. To simulate this clumping we took a Poisson distribution but with half of the mean as expected according to the sampling density and target frequency (e.g. 0·5 instead of 1 target for 70 targets using 1 hour/week sampling). Each time a target was captured it was multiplied by two. Therefore, targets always

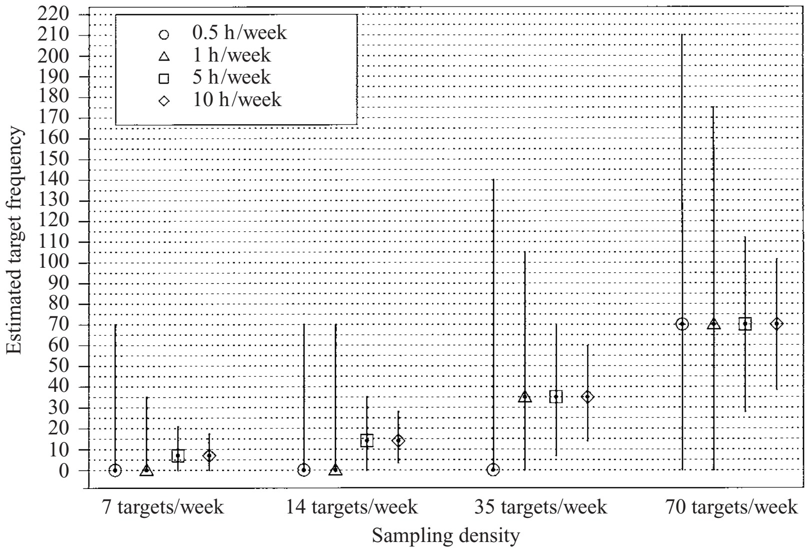


Fig. 5. Estimated weekly frequency of target (median and 95% confidence intervals) as a function of rate of occurrence and sample density (clumped).

occurred in pairs and the expected frequency is the same as for a normal Poisson distribution. It can be seen that the results are uniformly worse than when the targets are distributed in time randomly and independently (with 5 median values of 0) – which means that basically all of the theoretical analyses in this paper probably present a slightly optimistic picture.

*Power analysis*

A particularly revealing way to compare the different sampling schemes is using hit rate or hit probability. Hit rate is defined as the probability to detect at least one Poisson distributed target event during a sampling time period, for example, one week. The hit probability can therefore be seen as the power of the sampling scheme to detect at least one target. It is calculated by:

$$\text{Hit Rate} = 1 - [P(k=0)]$$

where $[P(k=0)]$ is the probability that no target will be captured.

Figure 6 presents the hit probabilities for various rates of occurrence and various sample densities. Using as an arbitrary criterion a hit probability of 0·5 – indicating that during a given week one is as likely to detect a target as not – we can see the following patterns. Sampling at 0·5 hour/week none of the depicted rates of occurrence – up to 70 targets/week – yields a hit

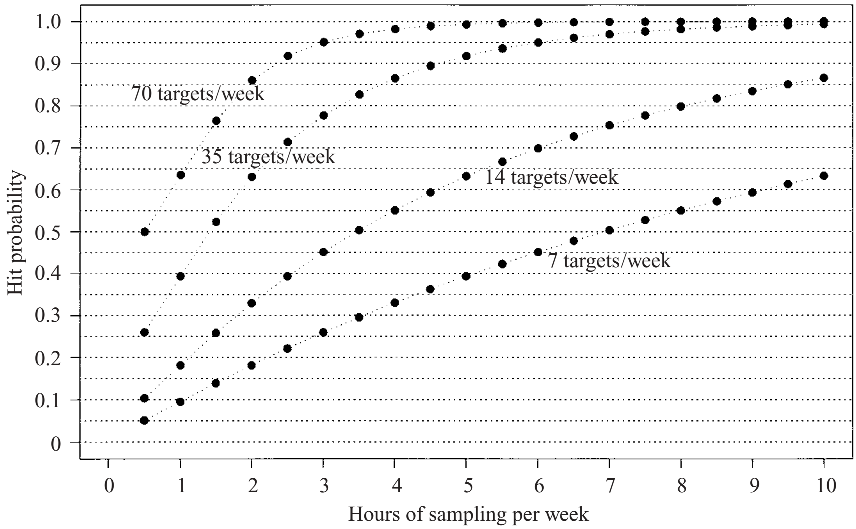


Fig. 6. Probability of capturing at least one target during a one week period.

probability greater than 0·5; the most likely occurrence is that we detect none. (Note that the values in this table are slightly lower than the expected frequencies depicted in Figure 1 because in that case sometimes more than one target is captured per sample.) Sampling at 1 hour/week yields a hit probability greater than 0·5 only for targets that occur approximately 50 times/week or more. Sampling at 5 hours/week yields values over 0·5 for all rates of occurrence except 7 occurrences/week, and sampling at 10 hours/week yields values over 0·5 for all rates of occurrence. Looking from the other direction, for targets that occur less frequently (e.g. 7 or 14 times/week) 4 to 8 hours of sampling per week are required to yield a hit probability greater than 0·5, whereas for more frequently occurring targets (35 or 70 occurrences/week) only one to two hours/week sampling is required.

One interesting feature of this analysis is that we can see an asymptote with the more frequently occurring targets. That is, for targets that occur 5–10 times/day (35–70 times/week), anything more than 3 or 4 hours of sampling per week yields very little additional power to detect at least one target per week – although of course a greater number of targets will continue to be captured with these denser samples (so the asymptote applies to hit rate only, not frequency estimations and the like).

### *Age of emergence*

Many analyses of child language attempt to estimate the age at which a particular target structure emerges in the child's linguistic competence.

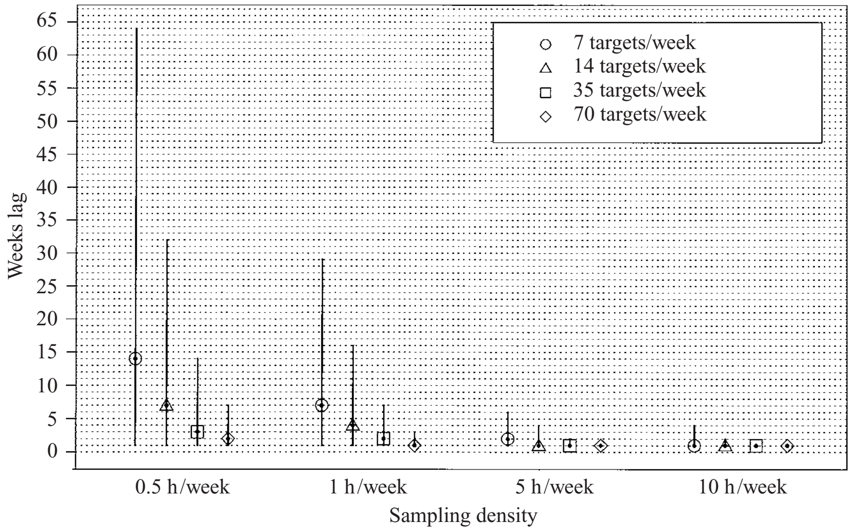


Fig. 7. Lag (delay) in estimated age of emergence (median and 95% confidence intervals) as a function of rate of occurrence and sample density.

Again, the accuracy with which this may be done will vary as a function of rate of target occurrence and sampling density. We thus used Monte Carlo methods to simulate for different Poisson distributed target frequencies the lag between the time a target was produced for the first time in 'reality' and the time that that target was detected in a sample for the first time. Again different sampling densities and different target frequencies were used as variables. For each target frequency condition we simulated random numbers with an underlying Poisson distribution until at least one target occurred. The number of simulations until one (or more) target occurred was used as the time in weeks needed to detect the target after it first occurred in the real world. This procedure was repeated 1000 times and the median and 95% confidence intervals were calculated (see above).

Figure 7 depicts the delay in picking up a target structure under different sampling schemes and rates of occurrence. What we see is that for sampling densities of 5 and 10 hours/week, the delays are quite small, about 1 to 3 weeks. For the most frequently used sampling techniques in the study of child language acquisition – 0·5 and 1 hour/week – the delays are relatively small for the most frequently occurring targets (a few weeks), but they are fairly large with very large confidence intervals for the two least frequently occurring targets (a few months).

Note that as a practical matter one could use these Monte Carlo simulations to statistically compare age of emergence for different targets taking into account target frequency (or different children's use of the same target). For example, we might wish to compare the age of emergence of the ditransitive dative and the prepositional dative for a given child. Let us assume, just for illustration, that the ditransitive dative occurred at an estimated rate of 70 times/week and emerged at 2;2;1, whereas the prepositional dative occurred at an estimated rate of 14 times/week and emerged 9 weeks later at 2;4;2. Given that we are working with a child sampled at 1 hour/week, based on Figure 7 we may use the confidence intervals at these two frequencies of occurrence to state that we are 95% confident that the ditransitive estimate we have is late by no more than 3 weeks (top of confidence interval), and we are similarly confident that our prepositional dative estimate is late by no more than 16 weeks (top of confidence interval). The distributions thus overlap, and so we cannot say with statistical confidence that the ditransitive dative emerged before the prepositional dative for this child. If the difference had been more like 20 weeks (instead of 9), we could have confidently established order of emergence for these two constructions, even taking into account their different frequencies of occurrence in the real world. Conversely, even a 9-week difference would have been enough if the prepositional dative had occurred at a rate of 35 (instead of 14) times/week in the real world. Of course systematic tables for a much wider variety of rates of occurrence and sample densities could be generated for use in making similar comparisons with all kinds of data.

## DATA-BASED ANALYSES

As a supplement to these theoretical analyses, we also conducted two simple empirical analyses. Both concerned well-known and important linguistic structures that occur relatively infrequently in child language. The first is English past tense overregularization errors, and the second is German passives. These were chosen simply because (i) there were existing analyses available to the authors that made counting frequencies relatively easy (see Abbot-Smith & Behrens, 2002; Maslen, Theakston, Lieven & Tomasello, 2003), and (ii) both were conducted using relatively dense sampling techniques (5–7 hours/week) over a one-year period. The basic strategy in both cases was to compare the full data sampled (5–7 hours/week) to subsets of the full data based on 0·5 and 1 hour/week (randomly sampled).

### *English past tense overregularizations*

Maslen *et al.* (2003) investigated the past tense overregularization errors of one English-speaking boy over a one-year period. They used a relatively dense corpus consisting of one hour per day five days per week from age

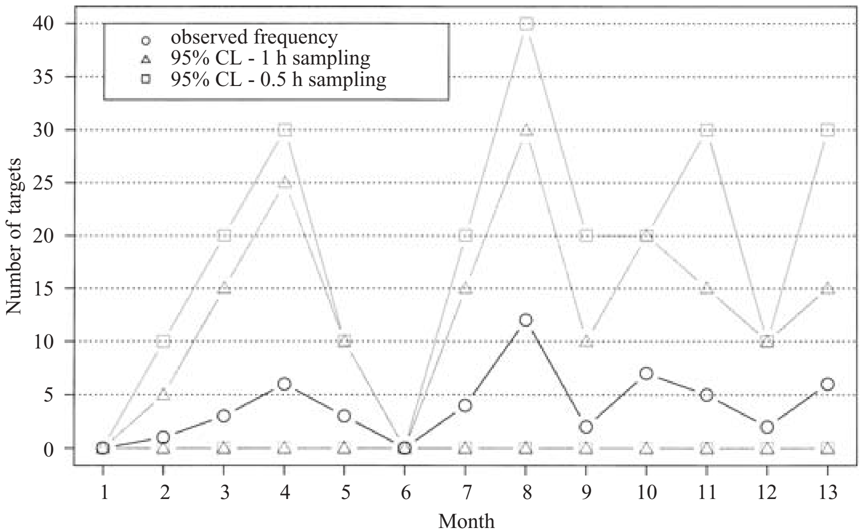


Fig. 8. Observed and estimated target frequencies of English past tense overgeneralization as a function of sampling density.

2;0 to 3;0. Based on that analysis, we graphed the number of errors observed in each month-long period, as shown in Figure 8.

We then wanted to see what would happen if we pretended that we had only a 1 hour/week sampling scheme, or a 0·5 hour/week sampling scheme. We therefore derived monthly estimates for these two sampling schemes in the following way. We randomly selected one day per week (in the case of the one hour/week scheme) or one day every two weeks (in the case of the 0·5 hour/week scheme) – and then added together the four weeks or two weeks sampled to get a monthly estimate. To estimate frequency in the same way as the observed figures, we multiplied by the appropriate amount (5 or 10). To provide stable estimates we did this 1000 times for each of the sampling schemes, and we present in Figure 8 the 95% confidence intervals (upper and lower) for those 1000 samplings.

The main thing to notice in Figure 8 is simply the great amount of variability in the estimates based on the sparser samples. For example, at month 4 during the third year of this boy's life, we observed in the 5 hours/week sampling 6 errors. Estimates based on the 1 hour/week sampling ranged from 0 to 25; estimates based on the 0·5 hour/week sampling range from 0 to 30. At month 9, the observed frequency is 2, and the estimates based on sparser samples range from 0 to 10 and 22, respectively. Since the lower bound estimate was in all cases 0, we can compute the difference in the upper bound estimates only. In general, the confidence intervals for

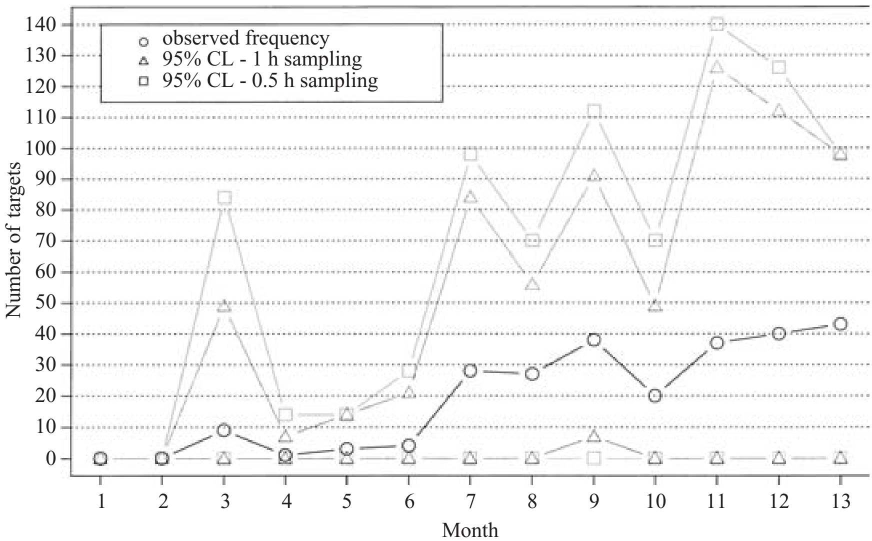


Fig. 9. Observed and estimated target frequencies of German passive as a function of sampling density.

0·5 hour/week estimate were about 50% larger (i.e. about 50% worse) than the one hour/week estimate – as would be expected (excluding the one month with a 0 observed frequency).

*German passives*

Abbot-Smith & Behrens (2002) investigated the passive utterances produced by one German-speaking boy over a one-year period from age 2;0 to 3;0. They used a relatively dense corpus consisting of one hour per day five days per week audio recordings, but also diary notes from the mother on the other two days of the week. These diary notes will be treated here as one hour of recording per day for those 2 days – so we have a total of 7 hours/week recording. Based on that analysis, we graphed the number of passives (*werden* passives only, as this is the main form in German) observed in each month-long period, as shown in Figure 9.

Following the lead of the previous analysis, we selected days based on a 1 hour/week sampling scheme or a 0·5 hour/week sampling scheme and did the appropriate mathematics (including the 1000 times samplings). We present in Figure 10 the 95% confidence intervals (upper and lower) for the 1000 samplings for both sampling schemes.

Again a salient feature of Figure 10 is the great amount of variability in the estimates based on the sparser samples. For example, at month 12 during the third year of this boy's life, we observed in the 5 hours/week

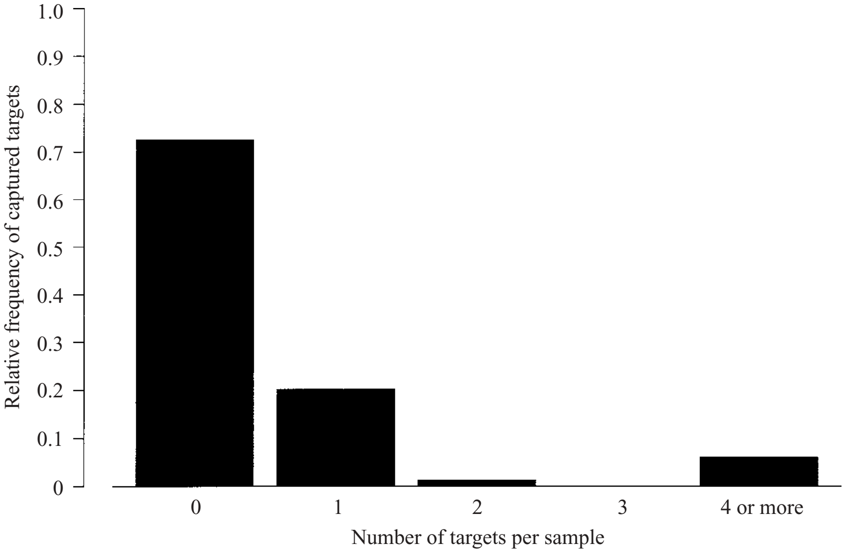


Fig. 10. Frequency distribution of captured targets after 1000 simulation of a 0·5 hour sampling of month 3 of German passive.

sampling 40 passives. Estimates based on the 1 hour/week sampling ranged from 0 to 112; estimates based on the 0·5 hour/week sampling range from 0 to 125. At month 7, the observed frequency is 28, and the estimates based on sparser samples range from 0 to 83 and 98, respectively. Since the lower bound estimate was again in (almost) all cases 0, we can compute the difference in the upper bound estimates only. In general, the confidence intervals for 0·5 hour/week estimate were about 40% larger (i.e. about 40% worse) than the one hour/week estimate – a bit better than would be expected by straight arithmetic (excluding the months with a 0 observed frequency).

As a way of seeing some of what is depicted in Figure 9 in a bit more detail, in Figure 10, we choose one month and look at a histogram of the sample values using the 0·5 hours/week sampling scheme. Specifically, at the third month, 9 examples were actually observed in the dense 5-hour sample, but in the 0·5 hour sample we captured 0 of these on 72·5% of the 1000 samplings, one of them on 20·2% of the samplings, and more than one in 7·3% samplings. Thus, the previous analysis showed what a wide range of estimates occurred, and this analysis shows that, even so, very many of these are 0.

## DISCUSSION

Scientific observation, as opposed to casual observation, is ever cognizant of possible limitations and biases built into the observational process.

One important dimension that always needs attention is the amount of sampling required for obtaining an accurate picture of the phenomenon of interest. Linguists in general, and child language researchers in particular, have not worried about this as much as they should have. In the fast-disappearing era in which we were simply concerned with which linguistic structures children produced at approximately which ages, this oversight was perhaps not so damaging. But as researchers become more and more concerned with issues of usage and processing and learning, things such as the frequency with which certain structures are produced and the precise timing of ontogenetic emergence become crucially important. If recordings of children's spontaneous speech are to play an important role in this new focus on process, we simply must get our methodological act in order.

The current paper represents only a very modest first step. Indeed, our major message for the moment is more negative and cautionary than positive and prescriptive. The main cautionary point is that the majority of existing child speech samples that have already been transcribed (e.g. in the CHILDES database) represent only a very small proportion of all the language the child produces and hears – on average around 1%. For some research questions this may be good enough. In particular, if we are only interested in the linguistic structures children produce and the approximate ages at which they produce them – and we are only interested in linguistic structures that occur with a fair amount of frequency – then we are on relatively safe ground. But as soon as we become interested in linguistic items and structures that a child produces only rarely (one or a few times per day), or we become interested in the relative frequency of particular linguistic structures in child speech, or we need to know the precise age of emergence of different structures that occur with different frequencies, we simply must attend carefully to issues of sampling – and in some cases 1% sampling is not adequate to answer the question at hand.

Being practical, we cannot simply ignore the immensely useful data already collected by many dedicated researchers, and, to repeat, the existing data are invaluable for answering many basic questions. But what we must do is to become more self-critical about the sampling process. For example, researchers should always take into account frequency when making age of emergence estimates, especially when comparing structures that occur at different frequencies (or children that use a given item or structure with different frequencies). And structures that are observed with very low frequency in our samples must simply be labelled as not analysable. More generally, the lesson is that we should not assume that the same sampling procedures are adequate for all questions. It is not a matter of one-size-fits-all, but rather we must sample children's speech in a manner appropriate for the question at hand.

Returning to our two practical questions from the introduction, we may say the following. Given that one has a question and wishes to design a sampling procedure, the following course of action might be recommended. The most important constraint is transcription, which determines a certain amount of both time and money. A researcher might begin by fixing the available amounts of time and money. There are then three major variables that affect transcription time:

- the number of children to be observed
- the length of time (ages) for which they are to be observed
- the density of the sampling during that observation time

With some simple mathematics, these three variables may be adjusted to fit within the resources available. A major consideration in this process, as we hope we have demonstrated in this paper, is the frequency with which the phenomenon of interest occurs. Quite simply: rarer phenomena need denser samples. How to estimate the rate of occurrence of a target in the real world – so that appropriate sampling techniques may be chosen – is a difficult question. But as a first approximation one may simply scale up from a sample using simple arithmetic (if one has a 2% sample, one multiplies everything by 50).

And we should also attempt to be creative in designing alternative kinds of sampling. For example, in some of our recent data collections we have sampled relatively densely (i.e. 5 or 10 hours/week) but only one week per month over a one-year period. This means that we have relatively large temporal gaps between samples, but at each sampling period we should be able to deal with all kinds of structures, including ones that occur with low frequency – all for the same amount of transcription time as if we had sampled uniformly across the year one hour/week (see also Bloom, 1970). This method thus has some advantages, as well as some disadvantages, relative to traditional sampling methods.

On the other hand, if a researcher does not have the time or resources to collect a new sample, then the issue is simply the confidence they can have in their analyses of existing corpora. There is of course no simple answer to this question, but in a sense it is the kind of question for which statistics are created, and we have attempted to make a modest contribution towards this end here. Some of the things hinted at in the current paper are: ways of comparing ages of emergence taking into account frequencies of occurrence and sample density, ways of assigning a kind of power quotient to different sampling techniques, and ways of assigning probabilities to frequency estimates (e.g. using Monte Carlo methods and confidence intervals). There are many more things that need to be done, and there are also other scientific fields from which methods could be borrowed to good effect (e.g. Borchers, Buckland & Zucchini, 2002).

The coming decades in linguistics in general will almost certainly be dominated by the analysis of corpora. Corpus analyses are already beginning to play an important role in most fields of linguistics, including even the writing of basic grammars (e.g. Biber, Johansson, Leech, Conrad & Finegan, 1999). Much of that work is done on written texts, since they can be scanned and so require no transcription time, but linguists interested in the most basic processes of language use and conversation focus as much as possible on the analysis of spoken language – where of course the corpora are much smaller (see e.g. the Santa Barbara corpus; Du Bois, 2000.). In the study of child language acquisition we have only transcripts of spontaneous spoken speech, which is a great advantage. We should exploit and develop that resource as much as possible. One part of doing this should be to develop the analytic tools that will enable researchers to make valid and reliable inferences from the transcriptions already available, and also to collect new corpora appropriately designed to fit specific questions.

## REFERENCES

Abbot-Smith, K. & Behrens, H. (2002). *One child's acquisition of the German passive*. Paper presented at IASCL/SRLCD Madison, Wisconsin.

Altman, J. (1974). Observational study of behavior: sampling methods. *Behaviour* **49**, 227–67.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson.

Bloom, L. (1970). *Language development: form and function in emerging grammars*. Cambridge, MA: MIT Press.

Borchers, D. L., Buckland, S. T. & Zucchini, W. (2002). *Estimating animal abundance. Closed populations*. Berlin: Springer.

Braunwald, S. R. & Brislin, R. W. (1979). The diary method updated. In E. Ochs & B. B. Schieffelin (eds), *Developmental pragmatics*. New York: Academic Press.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L. & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development* **10**, 159–200.

Caselli, M. C., Casadio, P. & Bates, E. (1999). A comparison of the transition from first words to grammar in English and Italian. *Journal of Child Language* **26**, 69–111.

Du Bois, J. W. (2000). *Santa Barbara corpus of spoken American English. Parts 1–3*. CD-ROM, Copyright by University of California, Santa Barbara.

Efron, B. (1979). Bootstrap methods: another look at the jacknife. *Annals of Statistics* **7**, 1–26.

Gentner, D. (1982). Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In S. Kuczaj (ed.), *Language development, Volume 2*. Hillsdale, NJ: Erlbaum.

Hall, W. S., Nagy, W. E. & Linn, R. (1984). *Spoken words: effects of situation and social group on oral word usage and frequency*. Hillsdale, NJ: Erlbaum.

Lieven, E., Behrens, H., Speares, J. & Tomasello, M. (in press). Early syntactic creativity: a usage-based approach. *Journal of Child Language*.

MacWhinney, B. & Snow, C. (1985). The child language data exchange system. *Journal of Child Language* **12**, 271–96.

Manly, B. F. J. (1998). *Randomization, bootstrap and Monte Carlo methods in biology*. 2nd edition. London: Chapman & Hall.

Maratsos, M. (2000). More overregularizations after all. *Journal of Child Language* **28**, 32–54.

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**, 34–69.
Martin, P. & Bateson, P. (1986). *Measuring behaviour: an introductory guide*. Cambridge: CUP.
Maslen, R., Theakston, A., Lieven, E. & Tomasello, M. (2003). *Past tense and plural overregularisations*. Paper presented at Child Language Seminar, Newcastle.
Mervis, C. B., Mervis, C. A., Johnson, K. E. & Bertrand, J. (1992). Studying early lexical development: the value of the systematic diary method. In C. Rovee-Collier & L. P. Lipsitt (eds), *Advances in Infancy Research*. Norwood, NJ: Ablex.
Pine, J. M., Rowland, C. F., Lieven, E. V. M. & Theakston, A. L. (2001). *Testing the Agreement/Tense Omission Model: why the data on children's use of non-nominative third person singular subjects count against the ATOM*. Poster presented at the Conference on Generative Approaches to Language Acquisition, Palmela, Portugal.
Rice, M. L., Wexler, R., Marquis, J. & Hershberger, S. (2000). Acquisition of irregular past tense by children with specific language impairment. *Journal of Speech, Language, and Hearing Research* **43**, 1126–45.
Tardif, T., Gelman, S. & Xu, F. (1999). Putting the noun bias in context: a comparison of English and Mandarin. *Child Development* **70**, 620–35.
Wells, G. (1985). Language development in the pre-school years. Cambridge: CUP.