Canadian Journal of Linguistics/Revue canadienne de linguistique, 63(4): 667–685, 2018
doi: 10.1017/cnj.2018.19

What is a representative language sample for word and sound acquisition?

NAOMI YAMAGUCHI
Laboratoire de Phonétique et Phonologie, CNRS and Université Sorbonne Nouvelle Paris 3
naomi.yamaguchi@univ-paris3.fr

Abstract

Naturalistic data are a useful source for language acquisition research. Recently, the emphasis has been placed on using denser corpora, in order to capture a more accurate picture of child language development. However, working with large amounts of data raises resource issues, since it is time-consuming to record and to transcribe. In this article, I focus on what would be the *ideal* duration of a naturalistic recording for it to be considered a representative enough sample of children's linguistic behaviours to observe the acquisition of words and sounds. Some of the results suggest that 30 minutes of recording may be enough to capture these specific developments, but these results are discussed in the perspective of what an *ideal* session could be.

Keywords: methodology, language acquisition, naturalistic data, lexical development, phonological development

Résumé

Les données naturalistes sont une ressource utile pour la recherche en acquisition du langage. Récemment, l'emphase a été mise sur l'importance de corpus plus denses afin de rendre compte de façon précise du développement du langage chez l'enfant. Cependant, travailler avec de grandes quantités de données soulève des problèmes de ressources, étant donné le coût en temps de l'enregistrement et de la transcription. Dans cet article, je traite de la durée *idéale* d'un enregistrement naturaliste, pour que ce dernier soit considéré comme un échantillon représentatif des comportements linguistiques des enfants quand on veut observer l'acquisition des mots et des sons. Certains des résultats suggèrent que 30 minutes d'enregistrement peuvent suffire pour rendre compte de ces développements spécifiques, mais ces résultats sont discutés par rapport à ce que pourrait être une session *idéale*.

Mots-clés: méthodologie, acquisition du langage, données naturalistes, développement lexical, développement phonologique

1. Introduction

In language acquisition research, naturalistic data have constituted a preferred way to observe children's speech. Naturalistic data are gathered by collecting children's speech in their natural environment – their home, their daycare centre or other familiar location – spontaneously interacting with those close to them, and with no specific instructions given by the researcher. Historically, the collection of these data was first carried out using parental diaries (e.g., in French, Grégoire 1937), as direct transcriptions of children's utterances. With the advancement of technology, naturalistic data are now digitally audio and/or video-recorded and can be transcribed later on.

Naturalistic data are a useful source for language acquisition research, because they have a "high ecological validity as the recording situation closely approximates the real-life situation under investigation" (Eisenbess 2010:12). In a naturalistic setting, the linguistic behaviour of the child is less likely to change from her usual behaviour than it would in experimental conditions. Moreover, except for recording equipment, the collection of naturalistic data does not require specific conditions, nor the establishment of an experimental protocol, and it is accessible to any speaker.

Naturalistic data can be collected in two different ways: longitudinally and cross-sectionally. Longitudinal collection captures the continuous language development of one child, on the premise that this individual development might be generalized to the global language development of children who speak this particular language. Cross-sectional collection captures stages of language development in children of different ages, on the premise that these different stages might represent a continuous temporal development.

In both cases, the purpose of collecting spontaneous child speech is to open a window on the child's language development, in terms of stages. For this purpose, one has to decide how frequent or how long the recordings must be. For instance, in phonological development studies, naturalistic longitudinal recordings occur from every week (Fikkert 1994) or two weeks (e.g., de Boysson-Bardies and Vihman 1991, Demuth et al. 2006, Rose 2000) to every month (e.g., Freitas 2003, Yamaguchi 2012, Wauquier and Yamaguchi 2013). The length of a single recording session varies from 30 minutes to one hour.

But even at this frequency, recording sessions are still only a sample of the child's actual productions. Sampling data may have effects on findings on language development. Such misleading results can be illustrated with the example of overregularizations in child productions. Marcus et al. (1992) found that overregularizations of regular past tense to irregular verbs represented a small proportion of their data. In the reexamination of the same data, Maratsos (2000) found that a different sampling would yield a different conclusion; another study by Maslen et al. (2004) showed substantial overregularizations, based on dense corpora. This example shows the importance of an adequate data sampling for a linguistic study. As Maratsos (2000) suggested, "fine-grained analyses" may be missed, because "these periods pass relatively quickly in time, or may be very sparsely sampled".

More generally, Rowland et al. (2008) examined the effects of data sampling on results, and they concluded that an inadequate data sampling would potentially lead

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

to two types of misleading results. The first type is a miscalculation of errors, be they infrequent or occurring in infrequent structures. The second type is a misestimation of linguistic productivity, the chance that frequent structures are overrepresented in smaller samples.

In order to avoid data sampling issues, many studies have recently stressed the importance of denser corpora to capture an accurate picture of child language development. *Denser* is understood as more frequent sessions of short duration (Tomasello and Stahl 2004, Rowland and Fletcher 2006, Lieven and Behrens 2012), for a total of two to 10 hours per week, for example, or sessions of longer duration at specific points in the child's development (Gilkerson and Richards 2008, Chabanal et al. 2015), as a continuous five to 12 hours of recording every six months, for example.

However, working with large amounts of dense corpora raises resource issues, since they are time-consuming to record and transcribe. For instance, an orthographic and phonetic transcription of a one-hour session may take up to 30 hours of work. At this point, one might ask whether denser corpora fit the purpose of the study.

All studies cited above dealt with syntactic or morphological analyses of children's production. In a half hour of speech, different morphological or syntactic events, such as the use of different tenses, different syntactic frames, or different morphological categories may occur very rarely, even in adult speech. But in the same amount of time, adult speech displays many exemplars of sounds (about 18,000, Rouas et al. 2004), syllables (from about 7000 to 12 000, depending on the speaking rate, Fougeron and Jun 1998), and consequently as many stress patterns, and words (about 5200, Grosjean and Deschamps 1975). The level of linguistic investigation is important in the sampling of data: while one needs more corpora in order to observe morphological or syntactic events, a phonological or lexical investigation can be performed on a smaller data sample.

The question of data sampling has not been as well documented for phonological or lexical development in child productions, as Demuth (2008) and Edwards and Beckman (2008) stressed. Lexical development is often analyzed through the evolution of vocabulary size, the composition of the lexicon, and the variability of the different words used (e.g., Bates et al. 1994, Bassano et al. 2005, Kern 2007). For the purpose of this study, two lexical variables produced by children were selected: word types and word tokens. Counting word types measures the diversity of the lexicon (that is, how many different words the child produces), while counting word tokens quantifies the frequency of occurrence of words. Phonological development concerns the acquisition of sounds and phonological structures, such as syllables, feet, stress, tones, etc. The analysis of these phenomena is linked to lexical development: the more different words a child produces, the more different phonological contexts there are. Phonological development may be analyzed through lexical production, but also through sound production (e.g., Demuth 1995, Rose 2000, Beckman et al. 2003, Demuth and Kehoe 2006, Fikkert 2007, dos Santos 2007, Yamaguchi 2012). I focused here on sound development, by selecting three different variables: produced sound types, produced sound tokens and target sound types. Produced sound types indicate how many different sounds a child produced, and produced sound tokens measure the frequency of each sound type. Target

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

sound types indicate the children's selectivity with respect to the targeted sound system.

This article tackles the issue of data sampling in terms of duration in the context of studying the development of words and sounds. The goal is to identify the ideal duration of a naturalistic recorded session for it to be considered a representative sample of children's linguistic behaviours, for phonological and/or lexical questions. In this sense, *ideal* should be understood as long enough to reflect as faithfully as possible the child's productions, but short enough to be transcribed in a reasonable amount of time.

The identification of the perfect session duration is done using two perspectives. Currently, if a researcher wants to analyze naturalistic child productions, two options are open: either using available corpora, or recording a new corpus. With the growth of available databases in the language acquisition research community, such as as CHILDES (MacWhinney 2000) or PhonBank (Rose and MacWhinney 2014), the first option has become a valid alternative. This is the first perspective: if we have access to already-recorded data, what do we need to transcribe? If, for example, recorded sessions are one hour long, is it possible to transcribe only part of them? The second option – recording a brand new corpus – takes more time, but may be necessary in order to study rare languages, for instance. In this case, I tried to identify what the adequate recording duration was in order to study the acquisition of words and sounds.

With these two perspectives in mind, I first present the method used, detailing the corpora used and the linguistic variables analyzed: word types, word tokens, sound types and sound tokens produced, and sound types targeted. Comparisons of child productions in different recorded sessions are then exposed, and balanced with parental input. Finally, I discuss all these results and suggest what an *ideal* recording may be for the study of word and sound development.

2. Method

The data used in this article come from two distinct corpora: the PREMS corpus and the PSPT corpus. Both consist of longitudinal recordings of naturalistic interactions between children and their parents, all monolingual French-speakers. In what follows, I first give details about the specific participants, the collection and transcription of each set of data, and then introduce the different variables and predictions.

2.1 The PREMS corpus

This corpus was collected and transcribed within the research project PREMS, supported by French National Agency for Research[1]. For the present study, the productions of four children from this corpus, three boys and one girl, were studied. They were recorded every two weeks at home, from the age of one to two years old.

[1]Grant reference: ANRBlanc_SHS2_2011: PREMS; Principal investigator: Dr Sophie Kern.

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

Sessions were recorded by an experimenter using a video camera and a digital audio recorder. This corpus is available on-line, as part of the CHILDES database[2] (MacWhinney 2000).

In this corpus, children's utterances were transcribed orthographically and phonetically using Logical International Phonetic Programs (LIPP). The transcriptions were then converted to the CLAN format (MacWhinney 2000) and then to the PHON format (Rose et al. 2006, Rose and MacWhinney 2014). Parents' utterances were orthographically transcribed directly using PHON, but not all transcriptions included parental productions. Parental phonetic transcription was automatically generated with PHON. All transcriptions were made by trained Linguistics students. All phonetic transcriptions were checked, and corrected if necessary, by the author.

2.2 The PSPT corpus

This corpus was collected and transcribed within the research project "Psychological Significance of Production Templates in Phonological and Lexical Advance: A cross-linguistic study", supported by the United Kingdom Economic and Social Research Council[3] (Wauquier and Yamaguchi 2013). For the present study, the productions of all seven children (4 boys and 3 girls) from this corpus were analyzed. Sessions were video-recorded using a camera and audio-recorded using a wireless microphone worn by the child. The children were recorded over a one-year period. The first session was recorded when they produced 20 different words on the basis of a parental questionnaire, namely the French adaptation (Kern and Gayraud 2010) of the MacArthur-Bates Communicative Development Inventory (Bates et al. 1988, Fenson et al. 2007). The ages of the children at the first recording session ranged from 17 to 23 months.

The corpus was transcribed directly using PHON (Rose et al. 2006, Rose and MacWhinney 2014). Parental productions were transcribed orthographically, and children's productions were transcribed orthographically and phonetically. All transcriptions were done by the author.

The data examined in this article is summarized in Table 1.

2.3 Comparing corpora

The aim of this article is to give researchers in language acquisition methodological tools to exploit longitudinal sessions without prior knowledge of the child's language development or of her communicative behaviour. The main factor is duration, and the comparison landmark between the children is age. In order to test the development of words and sounds, five variables were used: word types, word tokens, sound types, sound tokens and target sound types. Predictions about the influence of factors on these variables are presented.

[2] http://childes.psy.cmu.edu/media/Romance/French/Kern/

[3] Grant reference: RES-062-23-1889; Principal investigator: Pr Marilyn Vihman.

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

Corpus	Children	Age range	Recording frequency	# Sessions per child
PREMS	1 girl 3 boys	1;0–2;0	bi-weekly	from 17 to 28
PSPT	3 girls 4 boys	1;5–2;8	monthly	from 5 to 12

Table 1: Summary of the data analyzed in this article.

2.3.1 Duration

In language acquisition studies, bi-weekly or monthly recordings vary from 30 minutes to one hour. In the data used here, the recordings from the PREMS corpus were 50 to 60 minutes long (mean duration = 54 minutes); these sessions are henceforth termed *long sessions*. The recordings from the PSPT corpus were 30 minutes long, and are henceforth termed *short sessions*. Long and short sessions were compared in terms of language production.

In order to compare the language productions of the same children, each long session was also divided into two equal parts, based on duration only, regardless of the number of utterances produced. Then the language production was compared in each half with that of the entire long session.

2.3.2 Age

As shown previously, (e.g., Bates et al. 1995), individual children of the same age do not obligatorily share the same language development stage. Nevertheless, age can be a predictor of linguistic productivity in relation to certain age ranges. For example, it has been shown that there is a correlation between age and mean length of utterances (MLU) produced by children (see Conant 1987 for a review of studies about correlations between age and MLU).

Regarding lexicon development, one way to assess it is to use parental reports such as the MacArthur-Bates Communicative Development Inventory (Fenson et al. 1993). This questionnaire has been standardized and used for many languages. Studies have shown a correlation between chronological age and lexical growth in production for English (Fenson et al. 1994) as well as French (Kern 2003, 2007).

Even if there is individual variability between children, chronological age might give hints about children's linguistic development. Moreover, to avoid circularity, it is important to have an external, non-linguistic factor of the child's global development, in order to test the predictions about linguistic development.

2.3.3 Lexical and phonological variables

Five dependent variables were used in order to test the development of sounds and words according to the above factors.

As detailed above, two lexical variables were chosen: word types and word tokens produced by the children. The word type count measures the diversity of

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

the lexicon, that is, how many different words a child produces, and the word token count quantifies the frequency of occurrence of words.

Three different variables were used for the evaluation of sound development: produced sound types, produced sound tokens and target sound types. Targeted sound types are to be understood as including the phonemes of the language, that is the 36 French phonemes that compose French words and that the children need to acquire. Produced sound tokens and produced sound types are to be understood as any sound produced by the children, even if it is not a phoneme of the French language. Phonetic transcriptions were done perceptually, but the transcribers were encouraged to use diacritics if needed. Thus, produced sound types can be a clue to the phonetic variability of the children and produced sound tokens indicate the frequency of each produced sound.

These different linguistic variables could be influenced by the duration of the recorded session, as stated in the predictions below. In these predictions, we collapse short sessions and halves of long sessions as *30-minute sessions*, since we do not expect differences between the short sessions and the halves of long sessions.

(1) There would be more word types in a long session than in a 30-minute session, since the children are engaged in more and potentially more diverse activities.

(2) There would be more word tokens in a long session than in a 30-minute session, since the children have the time to produce more utterances.

(3) There would be no difference in the number of target sound types between long and 30-minute sessions, since thousands of instances of sounds may occur in 30 minutes, so every phoneme of the language has chances to be produced. The same applies for produced sound types, since the children have the chance to produce many instances of every sound they make.

(4) There would be more produced sound tokens in a long session than in a 30-minute session, since the children have the time to produce more utterances.

3. Results

In this section are presented the results relative to the predictions about the five linguistic variables in long, half and short sessions.

3.1 Focus on long sessions

In this analysis, I tried to determine whether it is necessary to transcribe a whole one-hour session in order to achieve the previously mentioned goals. Since many exemplars of words and sounds are produced in a half-hour, half a session may be sufficient. In this perspective, I tried to determine whether one half of a session is representative of the whole hour; and second, I tried to determine which half best represents the whole session.

Firstly, first and second halves are compared, to check whether one or the other was better in terms of linguistic productivity, using a Wilcoxon test with R, on the PREMS corpus, from 12 to 25 months old, for all four children. The means of

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

each linguistic variable on the overall sessions, standard deviations, and the results of the Wilcoxon test are presented in Table 2.

As shown in Table 2, even if it seems that the second half of each long session is more productive in terms of word types and tokens as well as sound types and tokens than the first half, the differences found are not statistically significant for all five variables. There seems to be no effect of tiredness or habituation on the children's linguistic productivity. It is worth noting that standard deviations are extremely high, showing great variability in the data.

I then compared second halves with whole long sessions, using a Wilcoxon test with R, on the PREMS corpus, from 12 to 25 months old, for all four children. The means and standard deviations of each linguistic variable, and the results of the tests on the comparison between second halves and overall sessions, are presented in Table 3.

As shown in Table 3, all linguistic variables are greater in the whole long sessions than in their second half. These differences are highly significant; standard deviations are also high, showing variability in the data.

These results confirm predictions 1, 2 and 4. In the one hour sessions, children have more time to produce more utterances. Prediction 3 is invalidated by these results: there are more sound types, produced or targeted, in a whole session than in half of it. These results suggest that, with a one-hour recorded session, it is better to transcribe the whole session.

3.2 Comparing long and short sessions: what should I record?

This second comparison is different from the last one. In long sessions, nearly one hour of parent-child interactions was recorded. The question in the preceding section was about the efficiency of transcribing the whole session. In the following comparison, interactions were recorded during 30 minutes only. The parents were told from the beginning of the recording period that the sessions would be 30 minutes long. The question here is whether the duration of recording is correlated to the linguistic productivity of the child.

	Mean Half 1	St. Dev. Half 1	Mean Half 2	St. Dev. Half 2	p-value
Word types	27.46	32.79	27.49	35.29	0.631
Word tokens	100.4	116.9	110.3	143.04	0.493
Sound types	28.83	4.45	28.91	3.94	0.901
Sound tokens	583.9	341.57	617.8	402.0	0.661
Target sounds	18.61	11.12	18.06	11.55	0.553

Table 2: Comparison of the mean number of occurrences for the linguistic variables in each of the two halves of the long sessions.

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

	Whole session	St. Dev (whole session)	Half 2	p-value
Word types	45.25	73.29	27.49	<.001
Word tokens	210.75	305.03	110.3	<.001
Sound types	33.62	5.36	28.91	<.001
Sound tokens	1232.17	746.27	617.8	<.001
Target sounds	21.44	11.65	18.06	<.001

Table 3: Comparison of the mean number of occurrences for the linguistic variables in half #2 and in the whole sessions.

I compared children from the PREMS and the PSPT corpora, by selecting data within the same age range (17–24 months old). If we follow the results of the first set of comparisons, then we should expect all linguistic variables to be greater in the long sessions than in the short sessions. Word types, word tokens, sound types, sound tokens are presented longitudinally according to the duration of the recordings sessions. Results are then compared to the parental input.

3.2.1 Children's productions

The comparison of the mean number of word types between long and short sessions is presented in Figure 1, along with standard deviation bars. As displayed in this figure, the number of word types is comparable in long and short sessions. In short sessions, the range of word types goes from 13 (at 18 months old) to 213 (at 24 months old). In long sessions, the range of word types goes from 5 (at 17 months old) to 284 (at 24 months old). Contrary to the prediction in 1, there is no significant difference between the mean number of word types in long sessions (67.42) and the mean number of

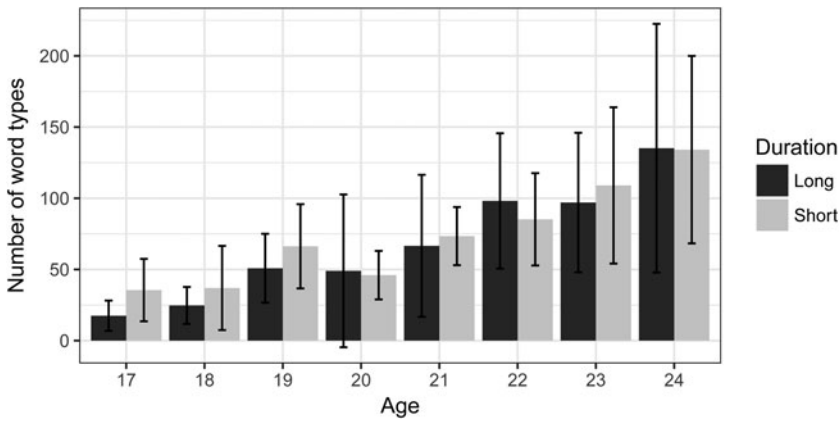

Figure 1: Number of word types in long and short sessions

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

word types in short sessions (84.41), as confirmed by a Mann-Whitney test, with U = 822 and p = 0.060.

However, the standard deviation bars on figure 1 indicate great variability among children, with data overlapping at each age point.

The comparison of the mean number of word tokens in long and short sessions is presented in Figure 2 along with standard deviation bars. In short sessions, the range of word tokens goes from 49 (at 18 months old) to 831 (at 24 months old). In long sessions, the range of word tokens is wider, and goes from 11 (at 17 months old) to 1357 (at 24 months old). Surprisingly, there is no significant difference between the mean number of word tokens in long sessions (309.3) and the mean number of word tokens in short sessions (297), as confirmed by a Mann-Whitney test, with U = 976.5 and p = 0.491. This result means that, even if the child has twice the time to produce words, she does not produce more words in a 54-minute recording session than in a 30-minute recording session.

But, as displayed in Figure 2, this result conceals a great variability depending on age. Until the age of 20 months, there are slightly more word tokens in short sessions than in long ones. But from the age of 21 months, word tokens seem to be fewer in short sessions than in long ones, and this difference increases until the age of 24 months. It seems that the prediction in 2 is invalidated until the age of 20 months, but is validated from the age of 21 months.

Moreover, as for word types, word tokens show great individual variability. The extended standard deviation bars indicate that the individual productions of the children overlap regardless of the duration of the session.

The comparison of the mean number of produced sound types in long and short sessions is presented in Figure 3 along with standard deviation bars. In short sessions, the range of produced sound types goes from 20 (at 18 months old) to 46 (at 23 months old). In long sessions, the range of produced sound types and goes from 25 (at 17 months old) to 38 (at 19 months old). As displayed in this figure, there

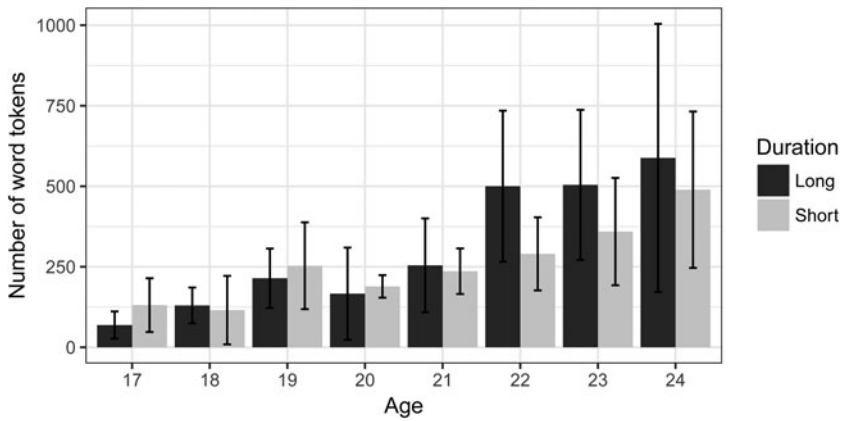

Figure 2: Number of produced word tokens in long and short sessions

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

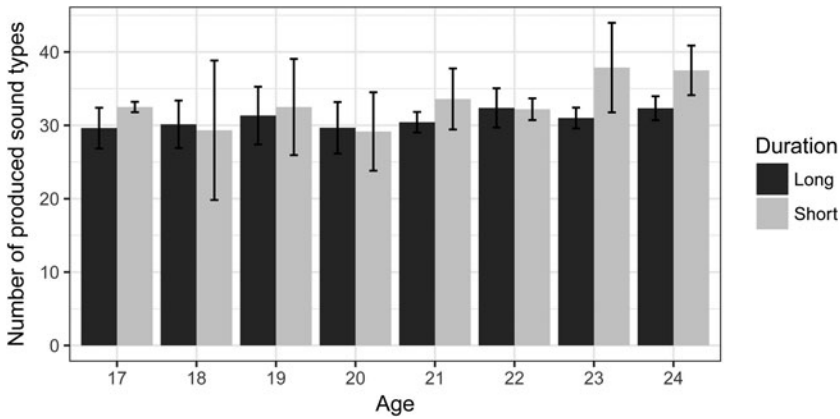

Figure 3: Number of produced sound types in long and short sessions

are more sound types in short sessions than in long sessions. The mean number of sound types is 33.9 in short sessions and 30.92 in long sessions. This difference is significant, as confirmed by a Mann-Whitney test, with $U = 622$ and $p = 0.001$. The prediction in 3 stated that there would be no difference in the number of sound types in long and short sessions, so this result – the children producing more varied sounds in a shorter session – is surprising. However, this result is to be taken with caution, since there is great individual variability exhibited by the extended standard deviation bars in figure 3, especially for children in short sessions. Moreover, recall that produced sound types do not necessarily correspond to phonemes of the target language, but to phones that the children produced. Since the transcribers were different for short and long sessions, it could be that the transcribers of the short sessions were more specific in the phonetic transcriptions than the transcribers of the long sessions. To support this hypothesis, the total number of phones used by the transcribers was counted, and it was found that indeed, the transcribers of the short sessions used 129 phones (including diacriticized phones), compared with 73 total phones used by the transcribers of the long sessions.

The comparison of the mean number of target sound types in long and short sessions is presented in Figure 4 along with standard deviation bars. In short sessions, the range of target sound types goes from 14 (at 18 months old) to 35 (at 24 months old). In long sessions, the range of target sound types and goes from 10 (at 17 months old) to 35 (at 24 months old). As displayed in this figure, the number of target sound types is similar in long and in short sessions. There is no significant difference between the mean number of sound types in long sessions (27.61) and the mean number of sound types in short sessions (30.44), as confirmed by a Mann-Whitney test, with $U = 825.5$ and $p = 0.083$. This result confirms the prediction in 3. As with the previous results, there is a great deal of individual variability, reflected in the nearly overlapping standard deviation bars for each session type.

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

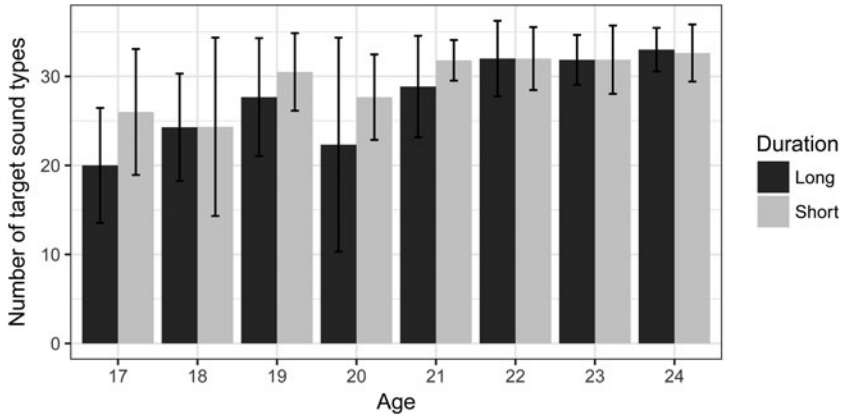

Figure 4: Number of target sound types in long and short sessions

The comparison of the mean number of produced sound tokens in long and short sessions is presented in Figure 5 along with standard deviation bars. In short sessions, the range of produced sound tokens goes from 269 (at 18 months old) to 2249 (at 24 months old). In long sessions, the range of produced sound tokens goes from 445 (at 17 months old) to 3697 (at 24 months old). As displayed in this figure, there are more sound tokens in long sessions than in short sessions. The mean number of sound tokens is 1507.5 in long sessions and 1058.2 in short sessions. This difference is significant, as confirmed by a Mann-Whitney test, with W = 1468, p-value = 0.002. As expected in prediction 4, there are more sound tokens in a 54-minute session than in a 30-minute session. Nevertheless, as was shown in figures 1 and 2, 54-minute sessions do not display more word types and word tokens globally. This fact, like the

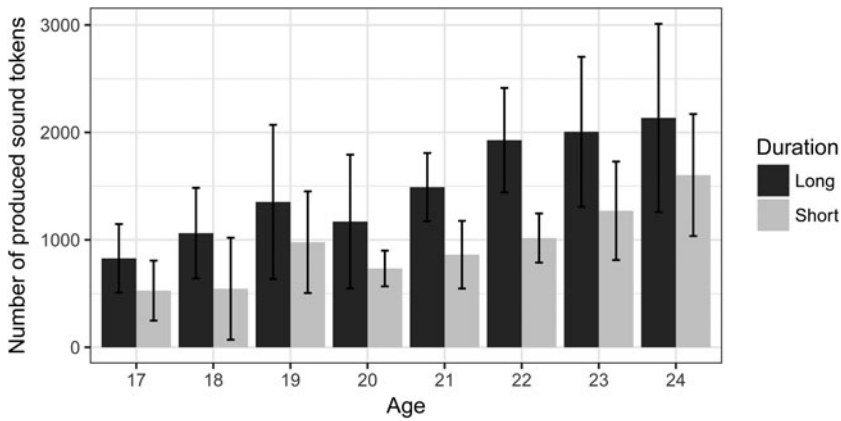

Figure 5: Number of sound tokens in long and short sessions

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

preceding result, suggests that the number of sound tokens may not be related to the number of word types or tokens.

3.2.2 Parents' productions

In order to explain these different results, an analysis of parental input was performed. This was done on fewer sessions, since not all parental utterances were transcribed in the PREMS corpus. In this corpus, only 23 sessions out of 52 were transcribed for parental input. The variables studied are word types and word tokens, since the phonetic transcription is missing for almost all sessions in both corpora.

The comparison of the mean number of parental word types and word tokens in long and short sessions is presented in Figures 6 and 7. As displayed in these figures,

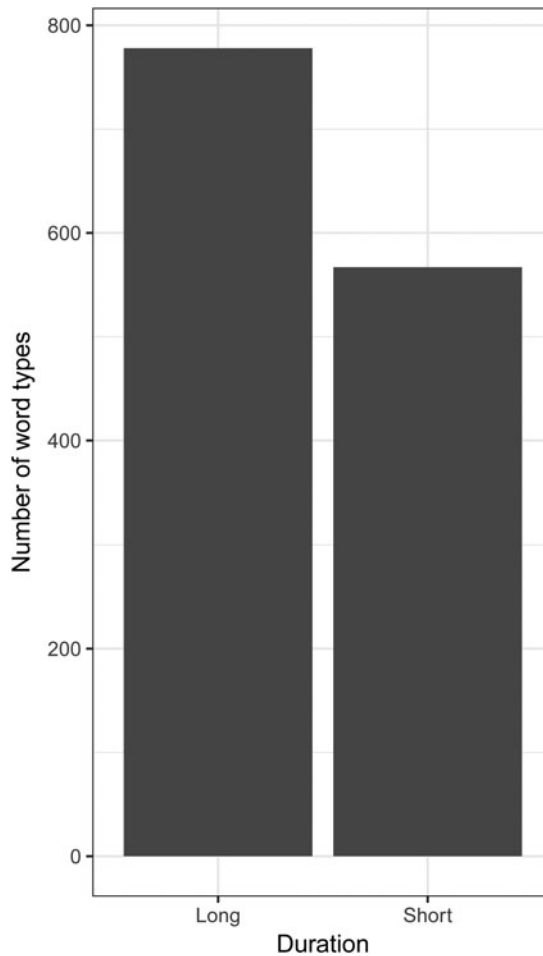

Figure 6: Number of word types in parental productions, long & short sessions

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

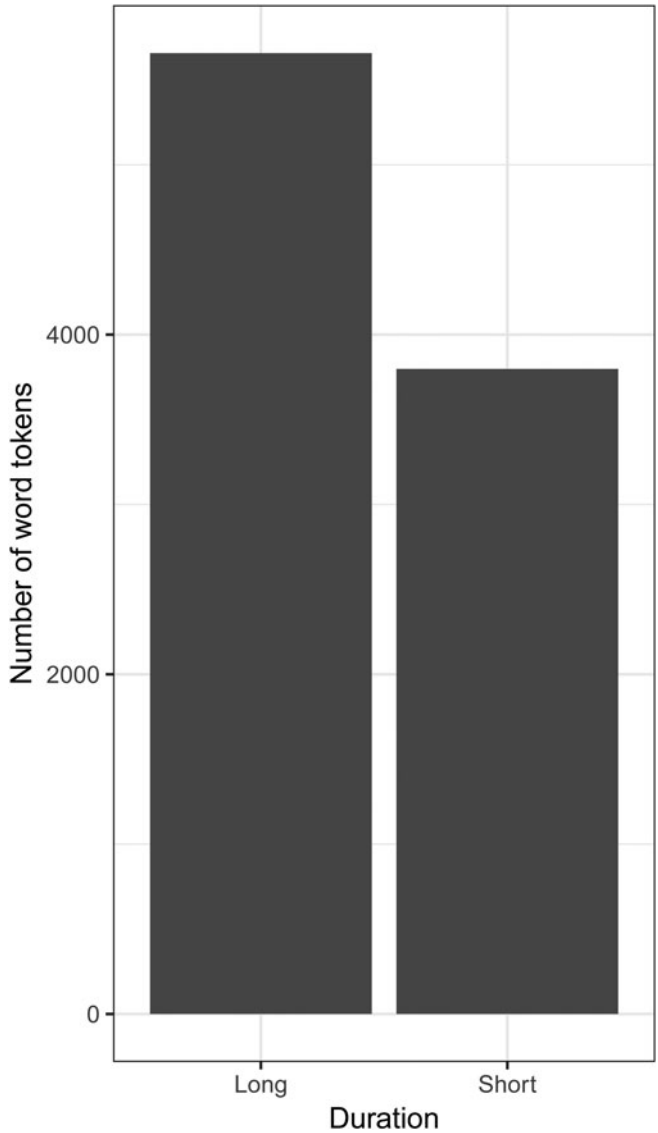

Figure 7: Number of word tokens in parental productions, long & short sessions

there are more word types and word tokens in long sessions than in short sessions. The mean number of word types is 563.43 in long sessions and 420.34 in short sessions. This difference is significant, as confirmed by a Mann-Whitney test, with W = 798, p-value > 0.001. The mean number of word tokens is 3855.52 in long sessions, and 2684.85 in short sessions. This difference is significant, as confirmed by a Mann-Whitney test, with W = 740, p-value > 0.001.

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

This result seems logical: parents produce more words in a 54-minute recording session than in a 30-minute recording session. Nevertheless, it should be noticed that parents do not produce twice as many words in a long session as in a short session.

4. Discussion and conclusion

The aim of this article was to identify the ideal duration of naturalistic parent-child interactions in order to have insights about children's acquisition of sounds and words.

The first question was about the efficiency of transcribing a whole one-hour session (if these sessions are already recorded). The first set of results suggested that, as expected, transcribing the whole session would give more data on word types, word tokens, produced sound types, produced sound tokens, and target sound types.

The second question was upstream of the question of transcription. The second set of results showed first that the development of the studied linguistic variables follows the same pattern for short and long sessions. As for quantitative results, the global results seemed to show that, as expected, the number of produced sound tokens is greater in long sessions than in short ones. As for the number of word types, word tokens, and target sound types, there was no difference between long and short sessions, and there were more produced sound types in short sessions than in long ones. Nonetheless, these surprising results need to be viewed with caution. Several hypotheses are offered to explain these results.

Age. Children of the same age may be at different levels of language development (for instance, in word productions, see Kern 2007). This is supported by the fact that there is a great deal of variability in the data examined, as shown by the standard deviation bars, which overlap at each point. As for word tokens, the results seem to indicate that from the age of 21 months there is a difference in favour of long sessions. This suggests that age should be taken into account when deciding on the ideal duration for a recording session. Before 20 months, the difference between a 30-minute and a 60-minute session may not be relevant, but it could be significant at a later stage.

Transcription. The results for sound types are interesting, because they suggest that they are the same or greater in short sessions as in long ones. As we have seen, these results may be due to a transcription bias, since many more phonetic symbols were used in the transcription of the short sessions. This suggests that the comparison of data should be done using inter-transcriber reliability and agreement (Vihman et al. 1985).

Context. The great variability in the results may also be explained by the variability of the situations in the recordings. One hypothesis is that parents may feel more involved in shorter sessions than in longer ones. It has been shown that the global involvement of parents favours children's linguistic skills (Tamis-LeMonda et al. 2004). This involvement may be reflected in the type of activities proposed during the recording session. While the children may be left alone for some time

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

in a one-hour session, this almost never occurs in a 30-minute session. This difference may affect the linguistic production of the children. Glas and Kern (2015) have shown that child language use is favoured in maintenance (health care, eating time) and social activities, compared to solitary activities. Since in a one-hour session, this last type of activity is more likely to occur than in a 30-minute session, it may explain the unexpected results for word types in short versus long sessions.

Finally, it should be noted that this study focused on the question of the quantity of data needed to study the development of sounds and words. Perhaps there is a need to investigate the question of the quality of the data, in the sense of diverse kinds of production. Previous studies have shown that children's productions are different in terms of speech acts (Leaper and Gleason 1996), lexicon (Gleason et al. 2009), or referential expressions (Salazar Orvig et al., in press) depending on the types of activity they are engaged in. In this perspective, recording different activities may help analyze how, how often and when children use the different linguistic resources available to them.

At first glance, some of these results seem to go against the generalization of dense corpora in language acquisition. Actually, if dense corpora are used in the perspective of recording multiple activities and situations, the chances of recording rare events, such as rare phonemes, rare combinations of phonemes, and rare words are multiplied, which could help to provide a fuller picture of child language development.

References

Bassano, Dominique, Pascale-Elsa Eme, and Christian Champaud. 2005. A naturalistic study of early lexical development: General processes and inter-individual variations in French children. *First Language* 25(1): 67–101.

Bates, Elizabeth, Inge Bretherton, and Lynn Snyder. 1988. *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge: Cambridge University Press.

Bates, Elizabeth, Philip S. Dale, and Donna Thal. 1995. Individual differences and their implications for theories of language development. In *The handbook of child language*, ed. Paul Fletcher and Brian MacWhinney, 96–151. Oxford: Blackwell.

Bates, Elizabeth, Virginia A. Marchman, Donna J. Thal, Larry Fenson, Philip Dale, J. Steven Reznicka, Judy Reilly, and Jeff Hartung. 1994. Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language* 21(1): 85–123.

Beckman, Mary E., Kiyoko Yoneyama, and Jan Edwards. 2003. Language-specific and language-universal aspects of lingual obstruent productions in Japanese-acquiring children. *Journal of the Phonetic Society of Japan* 7(2): 18–28.

de Boysson-Bardies, Bénédicte and Marilyn M. Vihman. 1991. Adaptation to language: Evidence from babbling and first words in four languages. *Language* 67(2): 297–319.

Chabanal, Damien, Loic Liegeois, and Thierry Chanier. 2015. Acquisition de la variation phonologique et recueil de corpus d'interactions naturelles parents-enfants : nouvelle méthode, nouveaux enjeux. *Lidil* 51: 65–88.

https://doi.org/10.1017/cnj.2018.19 Published online by Cambridge University Press

Conant, Susan. 1987. The relationship between age and MLU in young children: A second look at Klee and Fitzgerald's data. *Journal of Child Language* 14(1): 169–173.

Demuth, Katherine. 1995. Problems in the acquisition of tonal systems. In *The acquisition of non-linear phonology*, ed. John Archibald, 111–134. Hillsdale, NJ: Lawrence Erlbaum Associates.

Demuth, Katherine. 2008. Exploiting corpora for language acquisition research. In *Corpora in language acquisition research: History, methods, perspectives*, ed. Heike Behrens, 199–205. Amsterdam: John Benjamins.

Demuth, Katherine, Jennifer Culbertson, and Jennifer Alter. 2006.Word-minimality, epenthesis, and coda licensing in the early acquisition of English. *Language and Speech* 49 (2): 137–174.

Demuth, Katherine and Margaret Kehoe. 2006. The acquisition of word-final clusters in French. *Catalan Journal of Linguistics* 5: 59–81.

Edwards, Jan and Mary E. Beckman. 2008. Methodological questions in studying consonant acquisition. *Clinical Linguistics and Phonetics* 22(12): 937–956.

Eisenbess, Sonia. 2010. Production methods in language acquisition research. In *Experimental methods in language acquisition research*, ed. Elma Blom and Sharon Unsworth, 11–34. Amsterdam: John Benjamins.

Fenson, Larry, Philip S. Dale, J. Steven Reznick, Elizabeth Bates, Donna J. Thal, Stephen J. Pethick, Michael Tomasello, Carolyn B. Mervis, and Joan Stiles. 1994. *Variability in early communicative development*. Chicago: Society for Research in Child Development. URL http://www.jstor.org/stable/1166093.

Fenson, Larry, Philip S. Dale, J. Steven Reznick, Donna J. Thal, Elizabeth Bates, J.P. Hartung, S. Pethick, and J.S. Reilly. 1993. *The MacArthur Communicative Development Inventories: User's guide and technical manual*. Baltimore: Paul H. Brookes.

Fenson, Larry, Virginia A. Marchman, Donna J. Thal, Philip S. Dale, J. Steven Reznick, and Elizabeth Bates. 2007. *MacArthur-Bates communicative development inventories*. Baltimore: Paul H. Brookes, second ed.

Fikkert, Paula. 1994. On the acquisition of prosodic structure. Doctoral dissertation, Rijksuniversiteit Leiden. URL http://hdl.handle.net/2066/32125.

Fikkert, Paula. 2007. Acquiring phonology. In *The Cambridge handbook of phonology*, ed. Paul de Lacy, 537–554. Cambridge: Cambridge University Press.

Fougeron, Cécile and Sun-Ah Jun. 1998. Rate effects on French intonation: Prosodic organization and phonetic realization. *Journal of Phonetics* 26(1): 45–69.

Freitas, Maria João. 2003. The acquisition of onset clusters in European Portuguese. *Probus* 15 (1): 23–46.

Gilkerson, Jill and Jeffrey A. Richards. 2008. The LENA natural language study. Technical report, LENA Foundation, Boulder CO.

Glas, Ludivine and Sophie Kern. 2015. Early vocabulary development in french monolingual children and activity types. Presented at the Workshop on Infant Language Development, Stockholm.

Gleason, Jean Berko, R Ely, B Phillips, and Elena Zaretsky. 2009. Alligators all around: The acquisition of animal terms in English and Russian. In *Crosslinguistic approaches to the psychology of language: Research in the tradition of Dan Isaac Slobin*, ed. Jiansheng Guo, Elena Lieven, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura, and Seyda Ozcaliskan, 17–26. New York: Psychology Press.

Grégoire, Antoine. 1937. *L'apprentissage du langage: les deux premières années*. Paris: Alcan.

Grosjean, François and Alain Deschamps. 1975. Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica* 31(3–4): 144–184.

Kern, Sophie. 2003. Le compte-rendu parental au service de l'évaluation de la production lexicale des enfants français entre 16 et 30 mois. *Glossa* 85: 48–62.

Kern, Sophie. 2007. Lexicon development in French-speaking infants. *First Language* 37(3): 227–250.

Kern, Sophie and Frédérique Gayraud. 2010. *Inventaire français du développement communicatif*. Grenoble: Editions La Cigale.

Leaper, Campbell and Jean Berko Gleason. 1996. The relationship of play activity and gender to parent and child sex-typed communication. *International Journal of Behavioral Development* 19(4): 689–703.

Lieven, Elena and Heike Behrens. 2012. Dense sampling. In *Research methods in child language: A practical guide*, ed. Erika Hoff, 226–239. Oxford: Wiley-Blackwell.

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates, third ed.

Maratsos, Michael. 2000. More overregularizations after all : New data and discussion on Marcus, Pinker, Ullman, Hollander, Rosen and Xu. *Journal of Child Language* 27(1): 183–212.

Marcus, Gary F., Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, Fei Xu, and Harald Clahsen. 1992. *Overregularization in language acquisition*. Chicago: Society for Research in Child Development.

Maslen, Robert J. C., Anna L. Theakston, Elena V. M. Lieven, and Michael Tomasello. 2004. A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language, and Hearing Research* 47(6): 1319–1333.

Rose, Yvan. 2000. Headedness and prosodic licencing in the L1 acquisition of phonology. Doctoral dissertation, McGill University.

Rose, Yvan and Brian MacWhinney. 2014. The PhonBank project: Data and software-assisted methods for the study of phonology and phonological development. In *The Oxford handbook of corpus phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 380–401. Oxford: Oxford University Press.

Rose, Yvan, Brian MacWhinney, Rodrigue Byrne, Gregory Hedlund, Keith Maddocks, Philip O'Brien, and Todd Wareham. 2006. Introducing Phon: A software solution for the study of phonological acquisition. In *Proceedings of the 30th annual Boston University Conference on Language Development*, ed. David Bamman, Tatiana Magnitskaia, and Coleen Zaller, 489–500. Somerville, MA: Cascadilla Press.

Rouas, Jean-Luc, Jérôme Farinas, and François Pellegrino. 2004. Évaluation automatique du débit de la parole sur des donnés multilingues spontanées. In *XXVe journées d'étude sur la parole*, 437–440. Fez, Morocco.

Rowland, Caroline F. and Sarah L. Fletcher. 2006. The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language* 33(4): 859–877.

Rowland, Caroline F., Sarah L. Fletcher, and Daniel Freudenthal. 2008. How big is big enough? Assessing the reliability of data from naturalistic samples. In *Corpora in language acquisition research: History, methods, perspectives*, ed. Heike Behrens, 1–24. Amsterdam: John Benjamins.

Salazar Orvig, Anne, Haydée Marcos, Julien Heurdier, and Christine Da Silva. in press. Referential features, speech genres and activity types. In *Sources of variation in first language acquisition: Languages, contexts, and learners*, ed. Maya Hickmann, Edy

Veneziano, and Harriet Jisa, Trends in Language Acquisition Research. Amsterdam: John Benjamins.

dos Santos, Christophe. 2007. Développement phonologique en français langue maternelle: Une étude de cas. Doctoral dissertation, Université Lumière Lyon 2.

Tamis-LeMonda, Catherine S., Jacqueline D. Shannon, Natasha J. Cabrera, and Michael E. Lamb. 2004. Fathers and mothers at play with their 2- and 3-year-olds: Contributions to language and cognitive development. *Child Development* 75(6): 1806–1820.

Tomasello, Michael and Daniel Stahl. 2004. Sampling children's spontaneous speech: How much is enough? *Journal of Child Language* 31(1): 101–121.

Vihman, Marilyn May, Marlys. A. Macken, Ruth Miller, Hazel Simmons, and Jim Miller. 1985. From babbling to speech: A re-assessment of the continuity issue. *Language* 61 (2): 397–445.

Wauquier, Sophie and Naomi Yamaguchi. 2013. Templates in French. In *The emergence of phonology: Whole word approaches and cross-linguistic evidence*, ed. Marilyn Vihman and Tamar Keren-Portnoy, 317–342. Cambridge: Cambridge University Press.

Yamaguchi, Naomi. 2012. Parcours d'acquisition des sons du langage chez deux enfants francophones. Doctoral dissertation, Université Sorbonne Nouvelle Paris 3.