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

Naomi Yamaguchi

What is a representative language sample for words and sounds acquisition?

Naomi YAMAGUCHI
Laboratoire de Phonétique et Phonologie, CNRS & U. Sorbonne Nouvelle Paris 3

naomi.yamaguchi@univ-paris3.fr

Abstract

Naturalistic data are a useful source for language acquisition research. Recently, the importance of denser corpora has been emphasized in order to capture an accurate picture of child language development. However, working with large amounts of data raises resources issues, since it is time-consuming to record and to transcribe. In this article, we focus on the *ideal* duration of a naturalistic recording for it to be considered a representative sample of children's linguistic behaviors to observe the acquisition of words and sounds. Some of our results may 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 : method, language acquisition, naturalistic data, lexical development, phonological development

Abstract

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, nous traitons de la durée *idéale* d'un enregistrement naturaliste, pour que ce dernier soit considéré comme un échantillon représentatif des comportements linguistiques des enfants afin d'observer l'acquisition des mots et des sons. Certains de nos 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*.

Keywords : méthodologie, acquisition du langage, données naturalistes, développement lexical, développement phonologique

1 Introduction

In language acquisition research, naturalistic data have constituted a preferential way to observe children's speech. Naturalistic data consist in collecting children's speech in their natural environment, that is their home, their nursery, a familiar location, spontaneously interacting with

their close relatives, with no specific instructions given by the researcher. Historically, the collection of these data was first carried out by the means of 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). The linguistic behavior of the child is less likely to change from her usual behavior in a naturalistic setting than 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 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, and the premise is 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, and the premise is that these different stages might represent a continuous temporal development.

In both ways, 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 each recording must be. For instance, in phonological development studies, naturalistic longitudinal recordings occur from every week (Fikkert 1994) or two weeks (for instance de Boysson-Bardies and Vihman 1991; Demuth et al. 2006; Rose 2000) to every month (for instance Freitas 2003; Yamaguchi 2012; Wauquier and Yamaguchi 2013). The length of recording sessions varies from 30 minutes to one hour long.

But even at this frequency rate, recording sessions are still 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 to a different conclusion; another study of Maslen et al. (2004) showed substantial overregularizations in their study 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 to two types of misleading results. The first type of misleading results is a miscalculation of errors, may they be infrequent or occurring in infrequent structures. The second type of misleading results is a misestimation of linguistic productivity, the chance of having frequent structures being increased 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 small durations (Tomasello and Stahl, 2004; Rowland and Fletcher, 2006; Lieven and Behrens, 2012), for example for a total of 2 hours to 10 hours per week, or sessions of longer durations at specific points in the child's development (Gilkerson and Richards, 2008; Chabanal et al., 2015), for example a continuous 5 to 12 hours recording every 6 months.

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

All studies cited above concerned syntactic or morphological analyses of children's productions. In half an 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 duration 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 decisive in the sampling of data: if one needs more corpora in order to observe morphological or syntactic events, a phonological or lexical investigation could be performed on a smaller data sample.

The question of data sampling has not been as well documented concerning 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 (for instance Bates et al., 1994; Bassano et al., 2005; Kern, 2007). We chose to select two lexical variables produced by children: word types and word tokens. Word types count measures the diversity of the lexicon, that is how many different words children produce, and word tokens count quantifies the frequency of occurrences of words. Phonological development concerns the acquisition of sounds and phonological structures, such as syllables, feet, stress, tones... 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 (for instance Demuth, 1995; Rose, 2000; Beckman et al., 2003; Demuth and Kehoe, 2006; Fikkert, 2007; dos Santos, 2007; Yamaguchi, 2012). We 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 sound types indicate the children's selectivity concerning the targeted sound system.

This article tackles the issue of data sampling to study the development of words and sounds, in terms of time duration. Our goal is to identify the ideal duration of a naturalistic recorded session for it to be considered a representative sample of children's linguistic behaviors, 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 possible: either using available corpora, or recording a new corpus. With the growth of available databases in the language acquisition research community, such as CHILDES (MacWhinney, 2000) or PhonBank (Rose and MacWhinney, 2014), the first option is a valid alternative. This is our 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 it ? 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, we tried to identify the adequate recording duration in order to study the acquisition of words and sounds.

With these two perspectives in mind, we first present our method, detailing the corpora used and the linguistic variables analyzed in this study: 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 we discuss all these results and suggest what an *ideal* recording may be for the study of words and sounds development.

2 Method

The data used in this article gather two distinct corpora: the Prems corpus and the PSPT corpus. Both consist in longitudinal recordings of naturalistic interactions between children and their parents, all being monolingual French-speakers. In what follows, we first give details about the specific participants, the collection and transcription of each set of data, and then introduce our 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 were studied: three boys and one girl. They were recorded every two weeks at home, from the age of one to two years-old. 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).

Children utterances in this corpus 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 sessions were transcribed for parental productions. Parental phonetic transcription was automatically generated with PHON. All transcriptions were made by trained students in Linguistics. 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 UK's Economic and Social Research Council[3] (Wauquier and Yamaguchi, 2013). For the present study, the productions of all the seven (4 boys and 3 girls) children 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 during one year from the first session. The first session was recorded when they produced 20 different words on a basis of a parental questionnaire, namely the French adaptation (Kern and Gayraud, 2010) of the Mac Arthur Bates Communicative Development Inventory (Bates et al., 1988; Fenson et al., 2007). The ages of the first recording session varied from 17 to 23 months old.

This corpus was transcribed directly using PHON (Rose et al., 2006; Rose and MacWhinney, 2014). Parental productions were transcribed orthographically and children productions were transcribed orthographically and phonetically. All transcriptions were made 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

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

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

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

Corpus	Nb of children	Age range	Frequency of recording	Nb of sessions / 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.

her communicative behavior. Our main factor is time duration, and our comparison landmark between the children is age. In order to test the development of words and sounds, we used five variables: word types, word tokens, sound types, sound tokens and target sound types. Predictions about the influence of factors on these variables are presented.

2.3.1 Time duration

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

In order to compare the language productions of the same children, we also divided each long session into 2 halves, based on time duration only, regardless of the number of utterances produced. We compared the language production in one half with the whole long session.

2.3.2 Age

As shown previously in the literature (e.g. Bates et al., 1995), individual children of the same age do not obligatory 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).

Concerning 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 that there is 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, it is important to have an external, non-linguistic factor of the child's global development in order to test our predictions about linguistic development, so that circularity can be avoided.

2.3.3 Lexical and phonological variables

We used five dependent variables 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. Word types count measures the diversity of the lexicon, that is how many

different words children produce, and word tokens count quantifies the frequency of occurrence of words.

We selected three different variables for the evaluation of sound development: produced sound types, produced sound tokens and target sound types. Targeted sound types are to be understood as the phonemes of the language, that is the 36 French phonemes that the children need to acquire and that compose French words. 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 of the phonetic variability of the children and produced sound tokens indicate the frequency of each produced sound.

We predict that these different linguistic variables would be influenced by the time duration of the recorded session, as the below predictions state. In these predictions, we collapse short sessions and halves of long sessions as *30 minutes sessions*, since we do not expect differences between halves of long sessions and short sessions.

1. We predict more word types in a long session than in a 30 minutes session, since the children are engaged in more and potentially more diverse activities.
2. We predict more word tokens in a long session than in a 30 minutes session, since the children have the possibility to produce more utterances.
3. We predict no difference in the number of target sound types between long and 30 minutes session, 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. We expect more produced sound tokens in a long session than in a 30 minutes session, since the children have the possibility to produce more utterances.

3 Results

In this section we present the results concerning the predictions about the five linguistic variables in long, half and short sessions.

3.1 Focus on long sessions

In this analysis, we tried to determine if it is necessary for our sub-mentioned goals to transcribe a whole one-hour session. Since many exemplars of words and sounds are produced in half an hour, half a session may be sufficient. In this perspective, we tried to determine if one half of a session is representative of the whole hour; and secondly, we tried to specify which half is the best representative of the whole session.

First, we compared first and second halves in order to check if one 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 each linguistic variable on the overall sessions, standard deviations and 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 5 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 variability in the data.

	Half 1	St. Dev Half 1	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	0.661
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.

	Whole session	St. Dev	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.

We 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 more important 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 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.

We 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 more important 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.

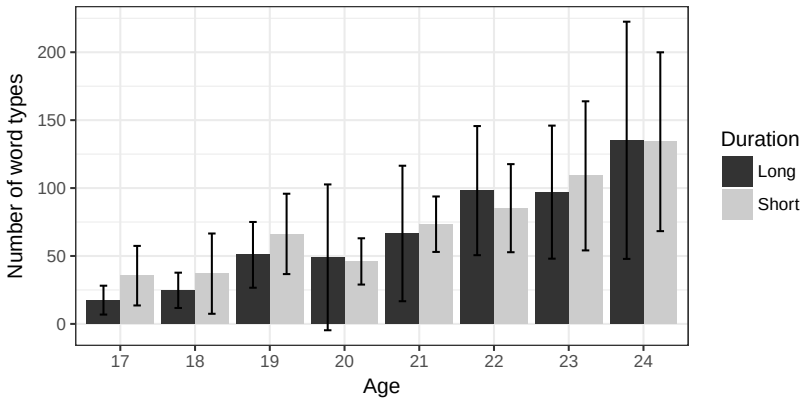

Figure 1: Number of word types in long and short sessions

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 (24 months old). In long sessions, the range of word types goes from 5 (at 17 months old) to 284 (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 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 a 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 (18 months old) to 831 (24 months old). In long sessions, the range of word tokens is wider, and goes from 11 (17 months old) to 1357 (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 minutes recording session than in a 30 minutes recording session.

But, as displayed in Figure 2, this result hides a great variability depending on age. Until the age of 20 months old, there are slightly more word tokens in short sessions than in long ones. But from the age of 21 months old, word tokens seem to be fewer in short sessions than in long ones, and this difference increases until the age of 24 months old. 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, like for word types, word tokens show a 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

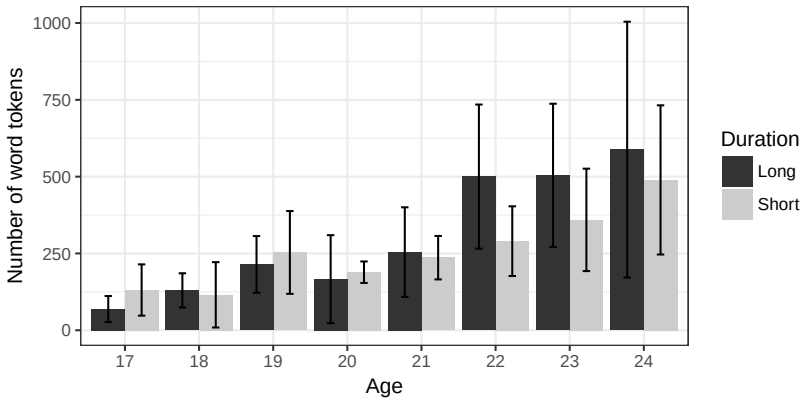

Figure 2: Number of produced word tokens in long and short sessions

produced sound types goes from 20 (18 months old) to 46 (23 months old). In long sessions, the range of produced sound types and goes from 25 (17 months old) to 38 (19 months old). As displayed in this figure, there are more sound types in short sessions than in long sessions. The mean number of sound types is equal to 33.9 in short sessions to 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, that is the children producing more varied sounds in a shorter session, is surprising. However, this result is to be taken with caution, since there is a great individual variability exhibited by the extended standard deviation bars in figure 3, especially with the children in short sessions. Moreover, recall that produced sound types do not obligatory 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 the case 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, we counted the total number of phones used by the transcribers; and we found that indeed, the transcribers of the short sessions used much more phones (including diacriticized phones), with a number of 129, compared to a number of 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 (18 months old) to 35 (24 months old). In long sessions, the range of target sound types and goes from 10 (17 months old) to 35 (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 for the previous results, there is a great individual variability, reflected in the nearly overlapping standard deviation bars for each session type.

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 (18 months old) to 2249 (24 months old). In long sessions, the range of produced sound tokens goes from 445 (17 months old) to 3697 (24 months old). As displayed

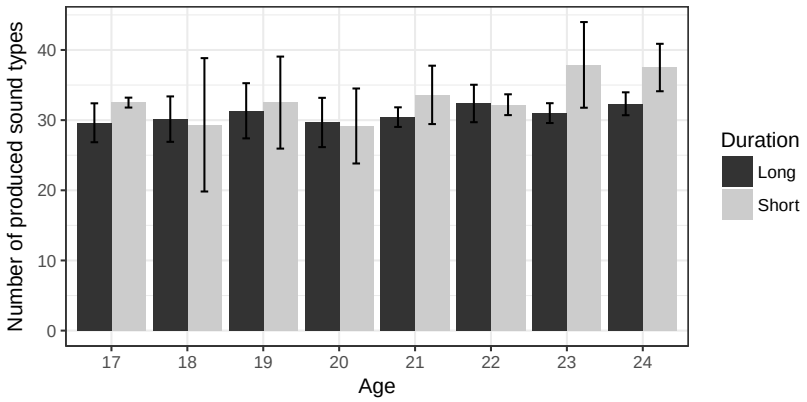

Figure 3: Number of produced sound types in long and short sessions

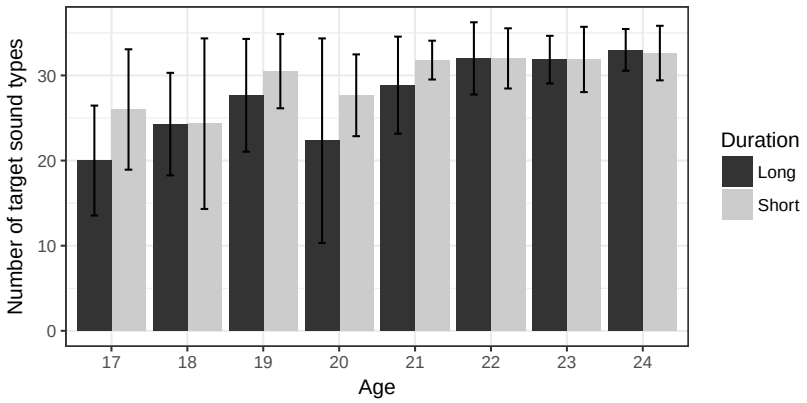

Figure 4: Number of target sound types in long and short sessions

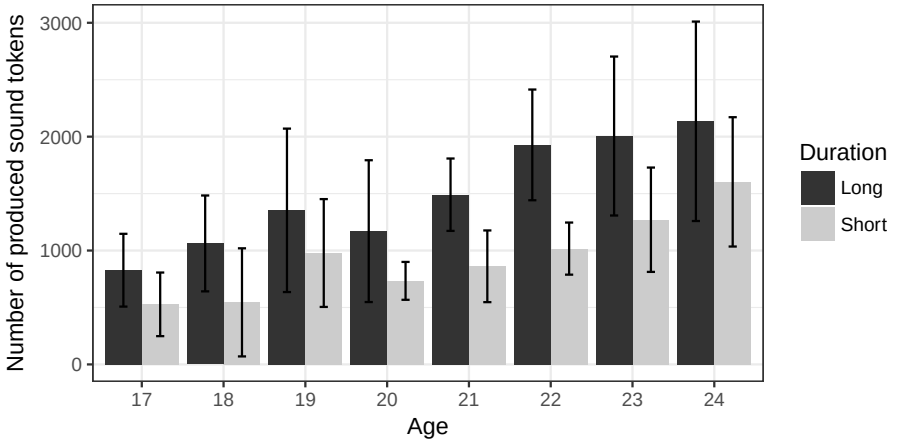

Figure 5: Number of sound tokens in long and short sessions

in this figure, there are more sound tokens in long sessions than in short sessions. The mean number of sound tokens is equal to 1507.5 in long sessions and to 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 shorter recording session of 30 minutes. Nevertheless, recalling the results shown in figures 1 and 2, 54 minutes sessions do not display more word types and word tokens globally. This fact, as the 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, we propose an analysis of parental input. It 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 the parental input. The studied variables 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, there are more word types and word tokens in long sessions than in short sessions. The mean number of word types is equal to 563.43 in long sessions and to 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 equal to 3855.52 in long sessions and to 2684.85 in short sessions. This difference is significant, as confirmed by a Mann-Whitney test, with W = 740, p-value > 0.001.

This result seems logical: parents produce more words in a 54 minutes recording session than in a 30 minutes recording session. Nevertheless, it should be noticed that parents do not produce twice the amount of words in a long session compared to a short session.

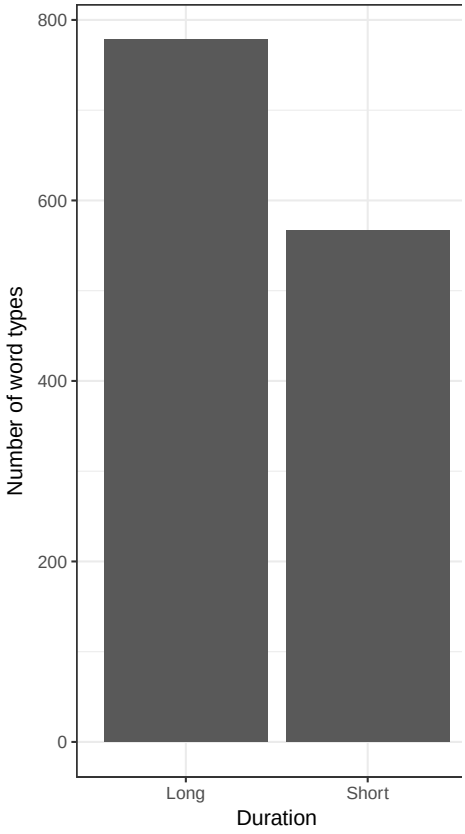

Figure 6: Number of word types in parental productions, long & short sessions

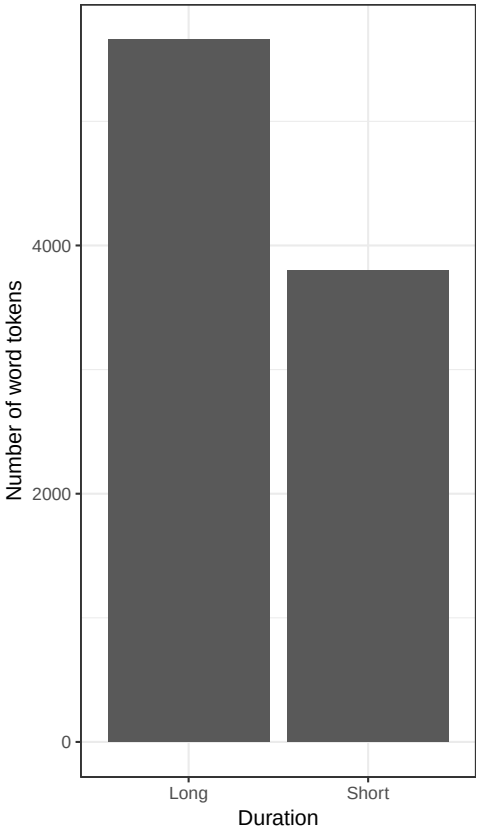

Figure 7: Number of word tokens in parental productions, long & short sessions

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.

Our 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 the question of transcription. The second set of results first showed 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, 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 taken with caution. We offer several hypotheses 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 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 favor of long sessions. This suggests that age should be taken into account when deciding of an ideal time duration for a recording session. Before 20 months, the difference between a 30-minutes and a 60-minutes session may not be relevant, but it could be significant afterwards.

Transcription. The results concerning sound types are interesting because they suggest that they are the same or greater in short sessions than in long ones. As we have seen, these results may be due to a transcription bias, since a lot 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 our 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 session than in longer ones. It has been shown that the global involvement of parents favors children's linguistic skills (Tamis-LeMonda et al., 2004). This involvement may be reflected in the type of activities proposed during the recording session. In a one-hour session, the children may be left alone for some time. It almost never occurs in a 30-minutes session. This difference may have consequences on the linguistic productions of children. Glas and Kern (2015) have shown that child language use is favored in maintenance (health care, eating time) and social activities, in comparison to solitary activities. Since in a one-hour session, this last type of activity is more likely to occur than in a 30 minutes session, it may explain the unexpected results concerning the word types differences in short and long sessions.

Finally, it has to be noticed that our study focused on the question of the quantity of data needed to study the development of sounds and words. Perhaps we need to investigate the question of the quality of the data, in the sense of diverse kinds of productions. 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., Under press) according to 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 a first sight, some of our 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 to record rare events, such as rare phonemes, rare combinations of phonemes, or rare words are multiplied, which could help obtain a fuller picture of child language development.

References

Bassano, Dominique, Eme, Pascale-Elsa, and Champaud, Christian (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, Bretherton, Inge, and Snyder, Lynn (1988), *From first words to grammar: Individual differences and dissociable mechanisms.* Cambridge: Cambridge University Press.

Bates, Elizabeth, Dale, Philip S., and Thal, Donna (1995), Individual differences and their implications for theories of language development. In *The Handbook of Child Language*, eds. Paul Fletcher and Brian MacWhinney, Oxford: Basil Blackwell, pages 96–151.

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

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

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

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

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* 14:169–173.

Demuth, Katherine (1995), Problems in the acquisition of tonal systems. In *The Acquisition of Non-Linear Phonology*, ed. J. Archibald, Hillsdale, N.J.: Lawrence Erlbaum Associates, pages 111–134.

Demuth, Katherine (2008), Exploiting corpora for language acquisition research. In *Corpora in Language Acquisition Research: History, methods, perspectives*, ed. Heike Behrens, Amsterdam, Philadelphia: John Benjamins Publishing Company, pages 199–205.

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

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

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

Eisenbess, Sonia (2010), Production methods in language acquisition research. In *Experimental Methods in Language Acquisition Research*, eds. Elma Blom and Sharon Unsworth, Amsterdam, Philadelphia: John Benjamins Publishing Company, pages 11–34.

Fenson, Larry, Dale, Philip S., Reznick, J. Steven, Bates, Elizabeth, Thal, Donna J., Pethick, Stephen J., Tomasello, Michael, Mervis, Carolyn B., and Stiles, Joan (1994), Variability in early communicative development. *Monographs of the Society for Research in Child Development* 59(5):i–185.

Fenson, Larry, Dale, Philip S., Reznick, J. Steven, Thal, Donna J., Bates, Elizabeth, Hartung, J.P., Pethick, S., and Reilly, J.S. (1993), *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual.* Baltimore : Paul H. Brokes Publishing Co.

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

Fikkert, Paula (1994), *On the acquisition of prosodic structure*. Ph.D. thesis, La Hague: HAG, The Netherlands.

Fikkert, Paula (2007), Acquiring phonology. In *The Cambridge Handbook of Phonology*, ed. Paul de Lacy, Cambridge : Cambridge University Press, pages 537 – 554.

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

Freitas, Maria Joao (2003), The acquisition of onset clusters in european portuguese. *Probus - International Journal of Latin and Romance Linguistics* 15 (1):23–46.

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

Glas, Ludivine and Kern, Sophie (2015), Early vocabulary development in french monolingual children and activity types. In *Workshop on Infant Language Development*, Stockholm.

Gleason, Jean Berko, Ely, R, Phillips, B, and Zaretsky, Elena (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*, eds. Jiansheng Guo, Elena Lieven, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura, and Seyda Ozcaliskan, New York: Psychology Press, pages 17–26.

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

Grosjean, François and Deschamps, Alain (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 Gayraud, Frédérique (2010), *Inventaire français du développement communicatif*. Grenoble: Editions La Cigale.

Leaper, Campbell and Gleason, Jean Berko (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 Behrens, Heike (2012), Dense sampling. In *Research Methods in Child Language. A Practical Guide*, ed. Erika Hoff, Oxford: Wiley–Blackwell, pages 226–239.

MacWhinney, Brian (2000), *The CHILDES Project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates, third edition.

Maratsos, Michael (2000), More overregularizations after all : new data and discussion on marcus, pinker, ullman, hollander, rosen & xu. *Journal of Child Language* 27:183–212.

Marcus, Gary F., Pinker, Steven, Ullman, Michael, Hollander, Michelle, Rosen, T. John, Xu, Fei, and Clahsen, Harald (1992), Overregularization in language acquisition. *Monographs of the Society for Research in Child Development* 57(Serial no. 228).

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

Rose, Yvan (2000), *Headedness and Prosodic Licencing in the L1 Acquisition of Phonology*. Ph.D. thesis, McGill University, Montreal.

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

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

Rouas, Jean-Luc, Farinas, Jérôme, and Pellegrino, François (2004), Evaluation automatique du débit de la parole sur des données multilingues spontanées. *XXVe Journees d'Etude sur la Parole (JEP 2004), Fes, Maroc* 437:440.

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

Rowland, Caroline F., Fletcher, Sarah L., and Freudenthal, Daniel (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, Amsterdam, Philadelphia : John Benjamins Publishing Company, pages 1–24.

Salazar Orvig, Anne, Marcos, Haydée, Heurdier, Julien, and Da Silva, Christine (Under press), Referential features, speech genres and activity types. In *Sources of variation in first language acquisition: Languages, contexts, and learners*, eds. Maya Hickmann, Edy Veneziano, and Harriet Jisa, Trends in Language Acquisition Research, Amsterdam : John Benjamins Publishing Company.

dos Santos, Christophe (2007), *Développement phonologique en français langue maternelle : Une étude de cas*. Ph.D. thesis, Université Lumière Lyon 2.

Tamis-LeMonda, Catherine S., Shannon, Jacqueline D., Cabrera, Natasha J., and Lamb, Michael E. (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 Stahl, Daniel (2004), Sampling children's spontaneous speech: how much is enough? *Journal of Child Language* 31:101–121.

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

Wauquier, Sophie and Yamaguchi, Naomi (2013), Templates in french. In *The Emergence of Phonology: Whole word Approaches and Cross-linguistic Evidence*, eds. Marilyn Vihman and Tamar Keren-Portnoy, Cambridge : Cambridge University Press, pages 317–342.

Yamaguchi, Naomi (2012), *Parcours d'acquisition des sons du langage chez deux enfants francophones*. Ph.D. thesis, Université Sorbonne Nouvelle Paris 3.