



CogniTExtes

Revue de l'Association française de linguistique cognitive

Volume 19 | 2019

Corpora and Representativeness

How large should a dense corpus be for reliable studies in early language acquisition ?

Christophe Parisse



Electronic version

URL: <http://journals.openedition.org/cognitextes/1483>

DOI: 10.4000/cognitextes.1483

ISSN: 1958-5322

Publisher

Association française de linguistique cognitive

Electronic reference

Christophe Parisse, « How large should a dense corpus be for reliable studies in early language acquisition ? », *CogniTExtes* [Online], Volume 19 | 2019, Online since 17 June 2019, connection on 01 July 2019. URL : <http://journals.openedition.org/cognitextes/1483> ; DOI : 10.4000/cognitextes.1483

This text was automatically generated on 1 July 2019.



CogniTExtes est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

How large should a dense corpus be for reliable studies in early language acquisition ?

Christophe Parisse

1. Introduction

- 1 When studying the development of child language from a usage-based perspective, the underlying rationale is that the data about usage (child and adult) should be as thorough as possible to ensure that demonstrations about usage-based development are valid. The organization of data collection for spontaneous speech studies needs to be carefully done (see Demuth 1996; Ambridge & Rowland 2013). A major issue is whether a corpus contains enough data to represent the information available to a child (Demuth 1996). Tomasello & Stahl (2004) tackled this issue for phenomena that occur with a low frequency. They suggested that it is necessary to have as much data as possible, otherwise the representation of a phenomenon in the data might be too sparse and therefore erroneous. Thus, they presented the concept of “dense corpus”, i.e. corpora which contain enough data so as to represent even rarely occurring events.
- 2 The concept of dense corpora for language acquisition has been applied among others by Behrens (2006), Lieven, Salomo & Tomasello (2009), Maslen, Theakston, Lieven & Tomasello (2004). The work presented in these papers is based on remarkable corpora, and two of them (Behrens 2006; Lieven et al. 2009) are available in the CHILDES database (MacWhinney 2000). Downloading and working with these corpora makes one realize the huge amount of work that underlies all this data. They have indeed provided the community with magnificent data to work on.
- 3 Unfortunately, not everybody works in an institute or a laboratory with the means to achieve such a project. And even for those who were able to fund such projects, they managed to do it for only one child. The funds available for transcription are a clear limitation on the creation of dense corpora. For a very large project such as the Human

Speechhome project (Roy et al. 2006), it was not possible to transcribe all the data available and a large amount of work had to be dedicated to automatic data processing. Future work that takes advantage of full text automatic speech recognition might be able to transcribe large amounts of data, but such tools are not yet available, especially for speech in noisy contexts and with speaker overlap. Moreover, even when these tools become available, they will still have shortcomings. In particular, they might not be able to tackle rare and unusual data. Moreover, for a more fine-grained analysis based for example on gesture or semantics, automatic processing will remain difficult to apply. Manual processing may well remain necessary, and limiting the amount of work necessary without diminishing the quality of the results remains an interesting goal.

1.1. Goal of the paper

- 4 One of the major purposes of gathering large dense corpora is to study language acquisition. Specifically, the goal is to be able to explain how a child might construct her production based on the data that she has heard or produced herself previously. This was for example the aim of the work by Lieven et al. (2009). They applied “a usage-based method called ‘traceback’ to the multi-word utterances of four two-year-olds to see how closely related these utterances are to their previous utterances” (p. 481). In other words, they “attempted to match each novel multiword utterance in a two-hour corpus to lexical strings and schemas that the child had said before” (p. 481). Their corpus was a dense corpus such as the Thomas corpus (Lieven et al. 2009) used in this paper (see below). This is a very good reason for the existence of dense corpora, as this kind of experimentation would have been difficult or even impossible to achieve with classical non-dense corpora.
- 5 The issue that the current paper aims to address is whether it is possible to limit the amount of data that a dense corpus needs to contain while allowing research such as that of Lieven et al. (2009) presented above. If this were possible, with the same workforce, it would be possible to study a corpus containing data of several children instead of only one. Or it would be possible to study a single child in a language other than English or German within a reasonable amount of time.
- 6 For the current paper, I shall try to assume as little knowledge as possible about language structure and child language. For example, I will not take into account structural information about sentences or utterances, nor will I base my analysis on classical grammatical categories or on a grammatical analysis. I will use only the raw text as basic material for my analysis. This is the approach used for unsupervised grammar induction (Chater & Manning 2006) or in more complex grammatical induction (Hsu et al. 2013). In these studies, the goal is to demonstrate which mathematical principles can explain how a grammar can be inferred using only positive evidence and without innate knowledge. The way data is handled is quite similar to what Lieven et al. (2009) did, but the target goal was to explain grammars rather than to explain the production of a child. This process of induction is applied in construction grammars (Croft & Cruse 2004) – exemplars, the most basic occurring items, are gathered into constructions, which can be included into larger constructions and generate at the end a grammatical network.
- 7 In the present work, I will not try to induce a full grammar. I will restrict myself to looking only at basic information provided in a corpus. My assumption is that looking at very low-level information is always necessary before looking at more general or high-level information. If there is some information which cannot be deduced from basic

simple elements, then this lack has to be accounted for and will have consequences at a higher level (for example in the generativist literature, this would correspond to basic grammatical properties of human languages).

- 8 Two types of basic elements could be used for this work: letters or words. Although letters are the most basic level, I chose to work at word level. Not only is this easier to implement than working at letter level, but this also makes it possible to study both lexical knowledge (using the word-metric as described below) and syntactic knowledge (using the bi-word metric as described below).
- 9 In order to determine the optimal size of a corpus, it is necessary:
 - A. To have a metric to measure the information available;
 - B. To use this metric to measure whether there is less information missing from a very large corpus than from a smaller corpus.

1.2. Responses to A: Which metrics should be used to evaluate the optimal size of dense corpora in order to study language acquisition?

- 10 I will use two simple metrics, the *word-metric* based on words, and the *bigram-metric* based on word bigrams (two consecutive words). In both cases the raw orthographic transcription is used and words are considered in their full form, not using lemmas, so that for example “word” and “words” are two different forms. All special symbols used in the transcription (in the current case it is the CHAT format, edited with the CLAN software, from the CHILDES database: MacWhinney 2000) were removed but the orthographic transcription was not modified. If the transcription used special symbols to express what was said (for example family words or names), it is assumed that the same symbols are used for the whole corpus, so this will not change the results.
- 11 Both metrics use the same basic method which is to measure, for a certain recording in the corpus, how many words or bigrams produced by the child are attested in the production of other/adult speakers in previous recordings. This corresponds to the idea that the child is capable of building her production based on material previously heard.
- 12 The two metrics are probably not optimal because they do not correspond to the actual data available to children. Crucially, children hear and memorize chunks of phonological data (Peters 1983). The constituents of the chunks—phonemes, syllables, or phonological words—will at some point become available to the child. At the same time, children will compose and generalize chunks, and later, parts of chunks. However, the data available in the corpus are not usually at the phonological level, which is certainly useful to the researcher on child language acquisition.
- 13 For the purpose of the current paper, I assume that information about words is related to information at phoneme level, and that information about bigrams is related to information at chunk level. This is not a perfect representation of child language development, but has the advantage of being theory free, reproducible, and easy to obtain.
- 14 Two parameters can be applied to both metrics:
 - (1) SIZE: the number of previous recordings taken into account;
 - (2) TYPE: the fact that the recording containing the child’s production under scrutiny is taken into account or not.

- 15 The rationale for the SIZE parameter is what the present work wants to test. If it is possible to predict the use of a word or a bigram with small values of SIZE, then developmental studies about language acquisition do not require large datasets to be compiled. Also, SIZE corresponds to a parameter that implements a limit to the child's memory. The child might not be able to (or does not need to) remember more than a certain amount of data. This has to be accounted for when explaining how a word is produced for the first time. The value of SIZE corresponds to the number of previous recordings that will be taken into account when computing the two metrics (words and bi-grams). So, the larger this amount, the larger the model for the child's memory.
- 16 The rationale for the TYPE parameter is the following. This parameter can take two values: 'previous sessions only' and 'including current session'. When the first value is used, the language heard by the child in the current recording is not considered; in this case the simulation is that of the child using only the knowledge available in long-term memory, which means that there was at least one period of sleep between the current session and the previous sessions (Peigneux et al. 2001). This presupposes that the data (or at least a large part of it) of all the previous sessions is available in long-term memory, which is probably too strong an assumption. When the second value is used, the data that the child hears in the recording under scrutiny are also taken into account. The order of production between the child and the adult is not important because both orders correspond to natural situations. If the child speaks after the adult, this means that the working memory of the child is taken into account. If the child speaks before the adult, this means that the adult production is a confirmation of the child's production, and probably corresponds to language shared by the child and the adult.

1.3. Responses to B: Using the metrics to optimize the size of dense corpora

- 17 The goal of the experiments in the present paper is to find out to what extent variation in the SIZE parameter affects the word-metric and the bigram-metric. The smaller the parameter can be for both metrics to produce the highest possible results, the easier it will be to conduct developmental studies. To do this, I will compute the word-metric and bigram-metric for existing developmental corpora for values of the SIZE parameter going from *one* (only one previous session is considered) to the highest possible number (this depends on the actual number of recordings in the developmental corpus).
- 18 This will be done for both values of the TYPE parameter, so as to check the stability of the results according to two different values of TYPE.

2. Corpus material and methodology

- 19 The current study used three English corpora available in the CHILDES database (MacWhinney 2000) that are longitudinal corpora with a large number of sessions, one dense and two not dense.
- 20 The corpora tested are:
 The Thomas¹ corpus (Lieven et al. 2009): 379 one-hour sessions (dense corpus).
 The Sarah² corpus (from the Brown corpus: Brown 1973): 139 30-minute sessions (corpus with many sessions).

The Lily³ corpus (from the Providence corpus: Song et al. 2009): 80 one-hour sessions (corpus with many sessions).

The Sarah and the Lily corpora are not dense corpora, but they are very large corpora. This makes it possible to compare dense and non-dense corpora and also makes it possible to test our proposal on more than one corpus only.

All the corpora tested here were recorded during spontaneous interactions between the child and a parent or a caretaker. Details about the protocols used are available on the CHILDES website in the download section for each corpus.

The Thomas corpus is the only truly dense English corpus available in the CHILDES database. It covers the production of Thomas from age 2;00 to age 4;11, but the densest part covers age 2;00.12 to age 3;02.12 and corresponds to 279 recordings. All the recordings come with audio material. This corpus was created following the dense corpus methodology proposed by Tomasello & Stahl (2004).

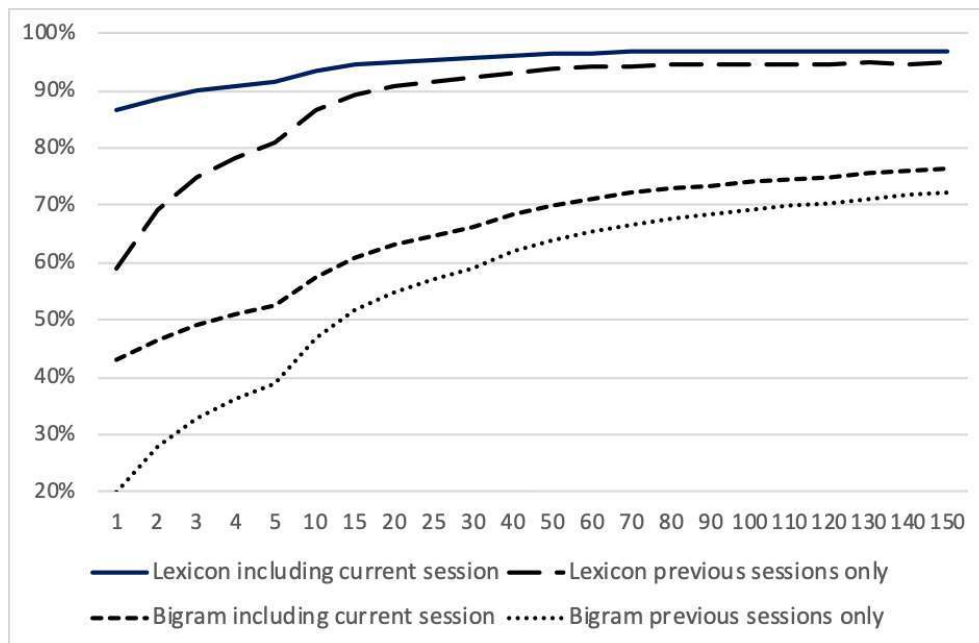
The Sarah corpus is one of the earliest full transcriptions of child oral language in a natural context. It is also one of the longest longitudinal corpora going from age 2;03 to age 5;01. Although not created using the dense corpus methodology, it offers a large amount of data.

The Lily corpus is a corpus going from age 1;01 to age 4;02 and belongs to the Providence corpus. All the recordings come with video material.

- 21 For each session of each of the three corpora, the percentage of words (or bigrams) present in the recording and found in the previous recordings were computed. This led to *detailed results* (percentages for each session) and *overall results* (percentage averages for all sessions). The detailed and overall results were systematically computed according to the two parameters described above: SIZE and TYPE.
- 22 First, the overall results will be presented for all three corpora and for all possible values of both parameters. Second, the detailed results of the most interesting parameters will be presented.

3. Overall results

Figure 1: Percentage of data covered for the lexicon (word level) and the bigrams according to the number of sessions memorized for the Thomas corpus



Note: Values of the SIZE parameter go from 1 to 150 (x-axis). The TYPE parameter represents the variants between curves including current session and curves not including current session.

- 23 Figure 1 presents the global results for both parameters. The values of the SIZE parameter range from 1 to 150. For the word-metric and for both values of the TYPE parameter, the results reach an asymptote. Values are higher when including the current session than when including the previous sessions only. This is especially true when the child is young, probably because in this case the interaction within a session has an important impact or because the role of working memory is more important. However, with both values of the TYPE parameter, a maximum of about 96% (SD 1.9%) is reached when including the current session and 94% (SD 2.5%) when excluding it. The asymptote is reached earlier when including the current session: at about value 20 for the SIZE parameter against value 60 otherwise. In both cases, a value of 100% is never achieved, which indicates that some information is missing when trying to understand word production by looking at previous productions only.
- 24 The results for the bigram-metric do not reach such high values. The maximum values attained are 75% (SD 3.8%) when including the current session and 71% (SD 4.4%) when excluding it. The asymptote is reached much later than for the word-metric, at value 120 including the current session and 130 otherwise.

Figure 2: Percentage of data covered for the lexicon and the bigrams according to the number of sessions memorized for the Sarah corpus

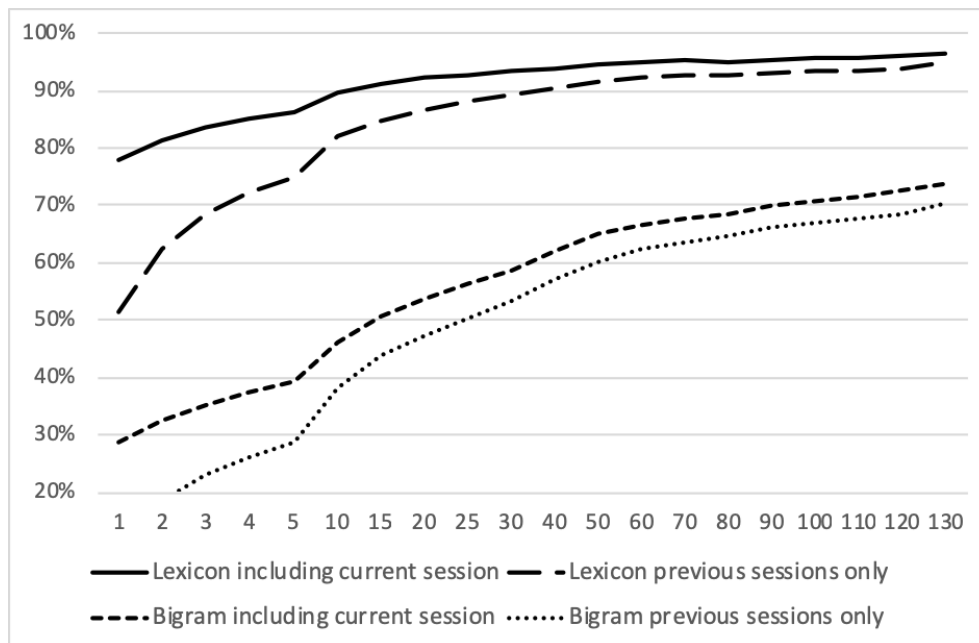


Figure 3: Percentage of data covered for the lexicon and the bigrams according to the number of sessions memorized for the Lily corpus

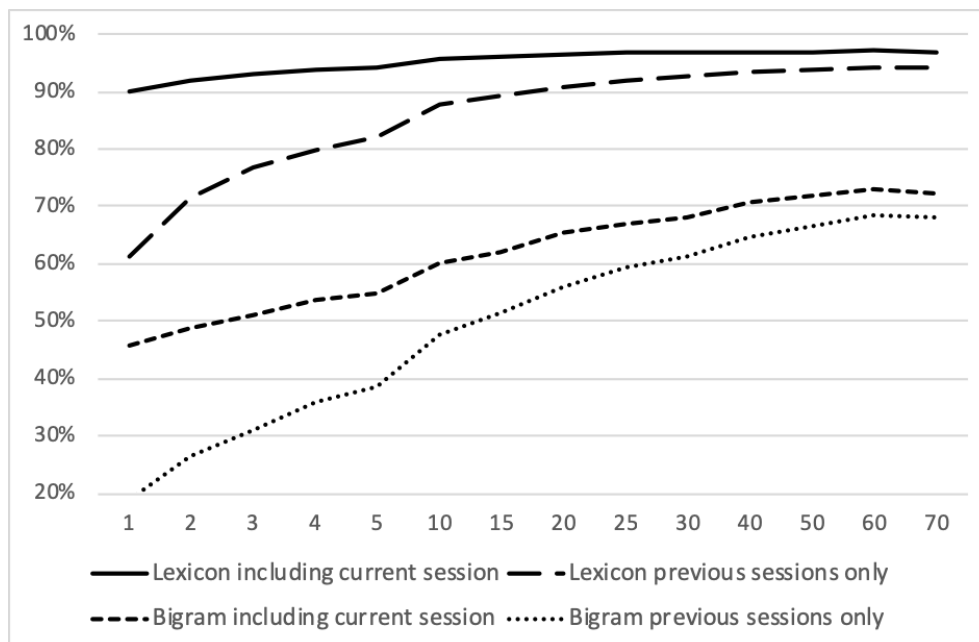


Figure 2 and 3 provide the same results for the Sarah corpus and the Lily corpus. The general trends of the results are very similar to those obtained for the Thomas corpus. The Sarah corpus leads to slightly less satisfactory results than the other two corpora. When comparing for a value of 40 for SIZE, the results for the Sarah corpus are 93% (SD 2.1%) for the word-metric instead of 96% (SD 1.8%) for the other two corpora; for the bigram-metric, a value of 62% (SD 4.2%) is obtained for the Sarah corpus, instead of 68% (SD 4.3%) for the Thomas corpus and 70% (SD 3.1%) for the Lily corpus.

4. Detailed results

- 25 The overall results indicate the mean values for the two metrics over the three corpora. This can be used to choose the optimal value for the SIZE parameter. An optimal value is small enough to be easy to implement, but large enough to provide good results. From the results above, 40 one-hour recordings seems to be close to the optimal size of a dense corpus.
- 26 I thus set the SIZE parameter to 40 to compute detailed results (other values close to 40 could have been used and would yield similar trends in the results).
- 27 Detailed results consist in computing the two metrics for each point in time in the developmental corpus, instead of giving the means for all points as for the global results. This makes it possible to find out whether the use of a dense corpus to predict child language output on the basis of previous input is easier at some age (at some point in the developmental corpus), for a given SIZE value.

Figure 4: Percentage of data covered for the lexicon and the bigrams according to the session number for the Lily corpus

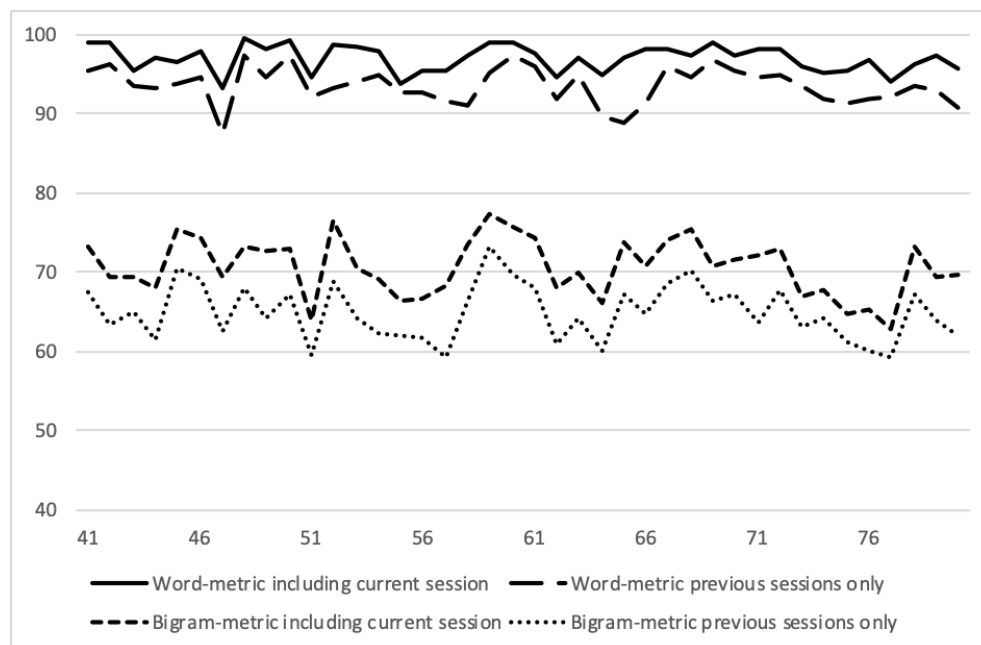


Figure 5: Percentage of data covered for the lexicon and the bigrams according to the session number for the Thomas corpus

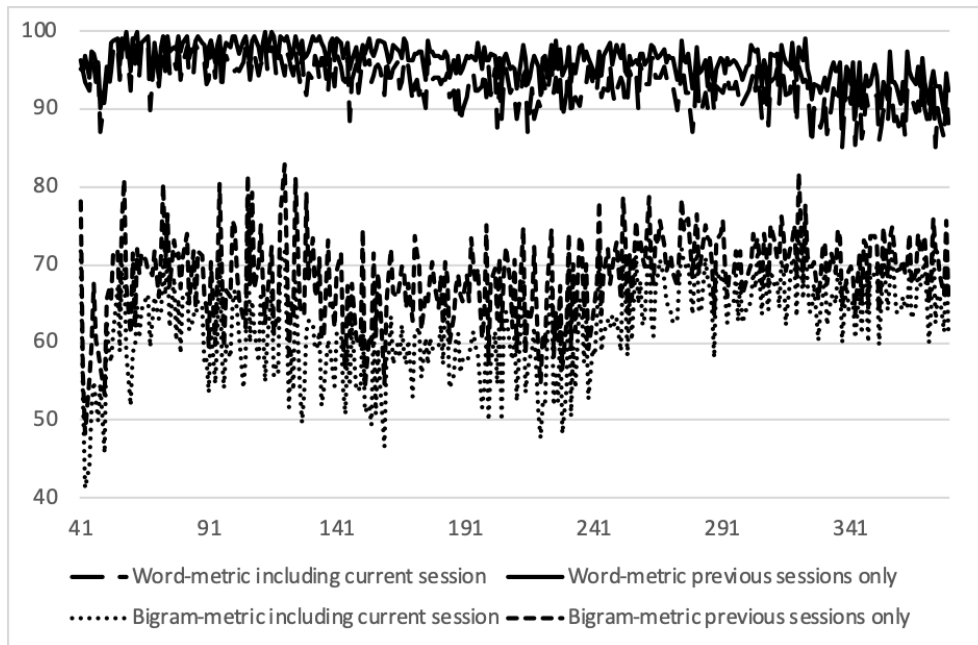


Figure 6: Percentage of data covered for the lexicon and the bigrams according to the session number for the first 80 sessions of Thomas corpus (sub-part of Figure 5)

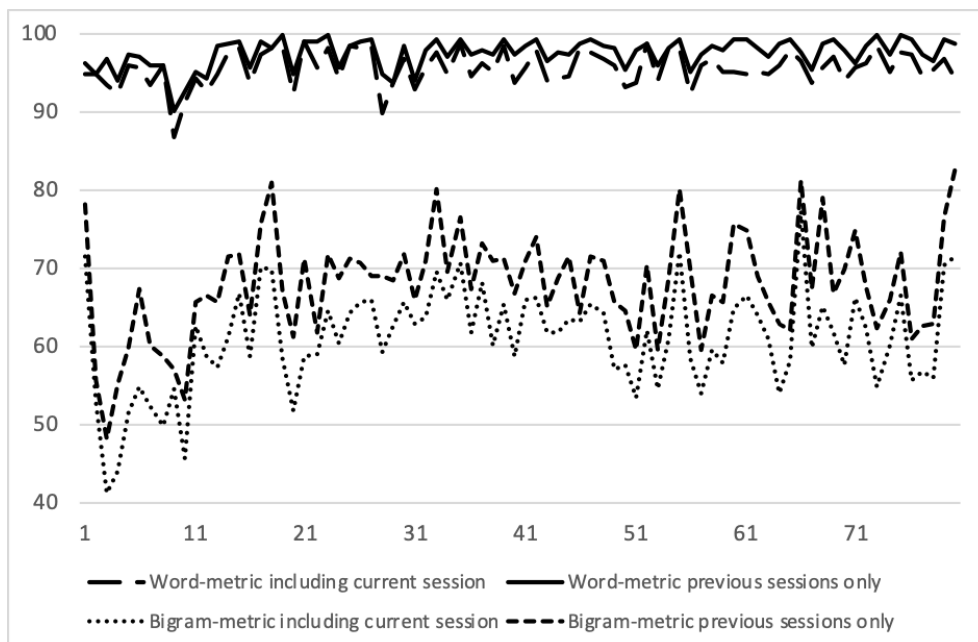
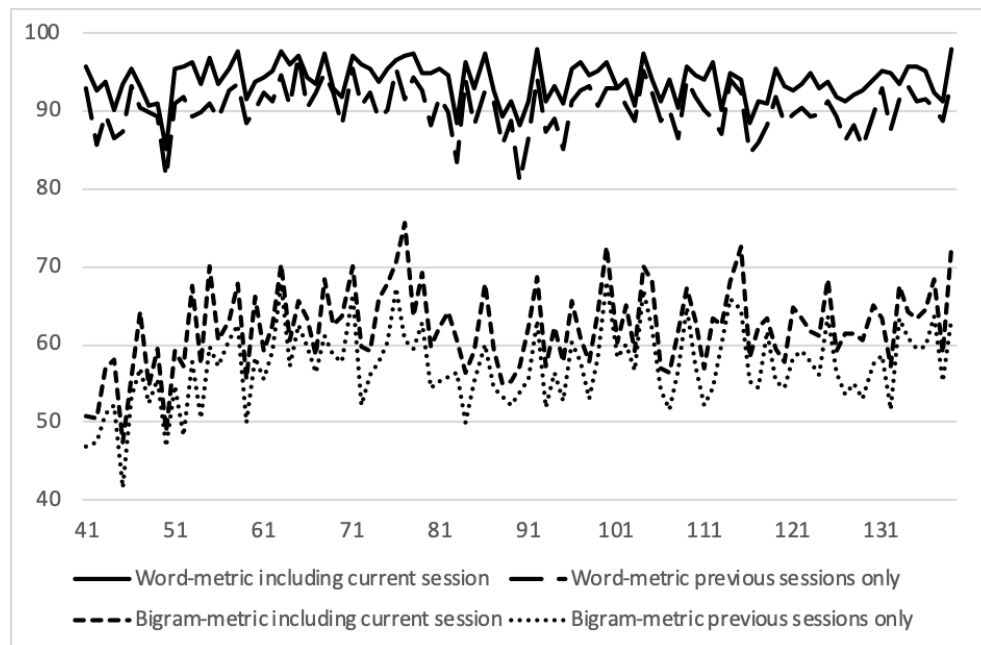


Figure 7: Percentage of data covered for the lexicon and the bigrams according to the session number for the Sarah corpus Figure 4 gives the evolution of the results for the Lily corpus.



- 28 The first result in the figure corresponds to session 41, as the SIZE parameter is 40. The figures for the Thomas corpus (see figure 5 and figure 6) and the Sarah corpus (see figure 7) have the same overall shape for the same SIZE parameter. There is variation from one session to another, but the general trend remains very stable. For Thomas, but not for Sarah or Lily, there is a significant difference in the percentages reached in the first half and the second half of the corpus. For words, percentages are lower in the second half: $M=94.7\%$ ($SD=2.37\%$) vs $M=91.7\%$ ($SD=2.64\%$), $t(332)=10.877$, $p < 0.00001$. For bigrams, percentages are higher in the second half: $M=59.4\%$ ($SD=5.64\%$) vs $M=64.5\%$ ($SD=5.33\%$), $t(336)=8.514$, $p < 0.00001$.

Table 1: results and variations for each of the corpora for the two metrics, for both values of TYPE.

	Word-metric including current session		Word-metric previous sessions only		Bigram-metric including current session		Bigram-metric previous sessions only	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Thomas	96.2	1.75	93.2	2.30	68.5	4.35	61.9	4.82
Sarah	93.9	1.96	90.3	2.34	62.2	4.23	57.1	3.85
Lily	96.9	1.38	93.4	1.83	70.6	3.06	64.8	3.00

All the results were computed for a parameter SIZE 40.

5. Discussion

- 29 The current paper tried to evaluate the coverage of dense corpora when they are used to infer the development of lexicon and grammar in a child. The results showed that a very large corpus was not necessary to reach a maximum level of coverage. Using a word-metric, an asymptote value was reached before finding a 100% coverage between the lexicon already heard (the adults from the previous sessions) and the lexicon that is under scrutiny (the child in the current session). This means that having more data available would not change the nature of the work to be done. There would not be any additional lexical information available with more data, which means that children are able to produce words that they did not hear before, and it is up to the researcher to understand how the child does this. A detailed analysis of creativity in the use of the lexicon is beyond the purpose of the current paper, but at first glance it appears that there was indeed lexical creativity of about 5%, which included producing non-existing words, new names, and of course correct lexical or grammatical derived words (correct generalisations). Finally, some of the missing words might be heard in other settings (most recordings were always produced in the same setting throughout all the corpora).
- 30 The case is less clear for the production of bigrams. The results did not exactly show an asymptote, but rather a slow increment of the coverage between bigrams already heard (the adults from the previous sessions) and the bigrams under scrutiny (the child in the current session). This is a major argument for very large dense corpora, as there is actually some development in the correspondence between child input and child output beyond the threshold of 40 sessions. However, this development is slow and the coverage quality obtained in about 40 or 50 sessions is already high. Is the upgrade from 70% of correspondence to 75% worth the cost? In some cases, the answer would be yes, as far as we have examples of large dense corpora available to test them.
- 31 A second result from the bigram coverage is that the percentages obtained were far from 100% and this discrepancy between what was heard and what was said will probably not be explained by having only more dense data. This means that the importance of creativity is much larger than for the lexicon, or that there are parts of utterance construction where the syntactic constraints are much laxer than others (for example, the frontier between two phrases, or two sentences). Also, the quality of the coverage did not change much with the age of the child (see figures 4, 5, and 6). So the mechanism for syntactic creativity, whatever it is, is already present at the beginning of our corpora (about age 2) and maybe even before. Children have the ability to be creative from the very start of language development. This shows that 1) it is necessary to study how language generalization can be made even with a small quantity of data; 2) having a larger amount of data (such as the Human Speechome project) might not solve all development issues. Explaining frequency effects on the large scale (cf. Ambridge et al. 2015) and explaining generalizations on the small scale are both necessary.
- 32 Finally, no major difference was found between a truly dense corpus such as the Thomas corpus, a semi-dense corpus such as the Sarah corpus, and a non-dense corpus such as the Lily corpus. This was the result obtained when trying to predict the general development of the lexicon and the grammar, but it might not be true if the goal were to study specific parts of grammatical development (e. g. Maslen et al. 2004). In this case, as argued by Tomasello & Stahl (2004), a non-dense corpus might miss valuable information at crucial

moments in language development, and so a larger dense corpus would probably be more suitable to the task.

- 33 It seems that, in the case of the induction of lexical or syntactic development, having closely time-related recordings does not change the nature of the data available. This is certainly a surprise, which gives interesting information about the nature of syntactic induction from input, or the nature of the input. For example, it may mean that syntactic induction is independent of the amount of data available in memory. Alternatively, it may mean that the variability of input data is low, so that the sampling frequency is not important but rather the sampling size. These hypotheses could be tested in future work involving a close inspection of corpus data.

6. Conclusion

- 34 The results presented above tend to show that using a dense corpus to predict lexical and syntactic development does not require corpora with several hundreds of recordings. This does not go against the argument of Tomasello & Stahl (2004) who emphasized that large dense corpora are needed for studying rare phenomena in child language. Rather, the results provide an upper bound to complement the lower bound of Tomasello & Stahl (2004).
- 35 Good research material has been shown to begin at 30 or 40 recordings. This is already a fairly large corpus, but much more manageable than a corpus of 379 recordings such as the superb Thomas corpus. This also means that the effort put into collecting a longitudinal corpus of 320 recordings could be put into building eight corpora comprising 40 recordings each. Having this type of corpora would greatly improve the information available in dense corpora, which to date is limited to a few children (as least for open corpora). This corpus size (40) was indeed used in work such as that by Dąbrowska & Lieven (2005), but it is interesting to be able to justify working on a certain size of data.
- 36 This does not mean that very large dense corpora with many more than 40 recordings, such as the Thomas corpus, are not useful. They cover a much larger range of material than what smaller corpora could do. They enable one to look at a large variety of grammatical situations, whereas a small corpus is limited to what happens at the time of data collection⁴. And large dense corpora were necessary for the current work to exist. I hope that other larger corpora such as the Thomas corpus will be produced in the future, but other smaller projects have great interest too.
- 37 A surprising result was that the coverage between what was heard and what was produced did not vary much between age 2, 3, 4, or even 5. There was a slightly decreasing trend in the Thomas corpus, but the sampling rate is not the same at the beginning and at the end of this corpus. In the Sarah and the Lily corpus, where the sampling rate does not change, the coverage values did not change. This suggests that the principle of trying to understand how ‘what the child hears’ can explain ‘what the child produces’ is applicable at all ages. It suggests that the child must be creative when young, and that creativity does not increase drastically at a later age. This also suggests that children could be learning language at their own rate, not at a rate constrained by the environment.
- 38 Finally, the bigram-metric presented here might appear too simple to represent true grammatical development. However, this simplicity is as much a strength as a weakness.

Most grammatical theories consider a higher level of organization, either at the utterance level or at the construction level. But for the child, this level of organization is not known in advance. Even in generative theories such as principles and parameters (Snyder 2007), the child has a huge number of possibilities to deal with. In cognitive grammar theories, this is even truer (cf. Croft 2001). For example, in research such as Lieven et al. (2009), they found that reconstructing the child's utterances from previous data was more difficult as the child got older. However, their analysis was based on higher level structures which made their work more difficult and could explain their results.

- 39 Conversely, the bigram-metric is also a good representation of a low level of language organization. Assuming that a child is able to identify elements of language (words or morphemes – e. g. Saffran et al. 1996), she might have lasting access to the type of information represented by bigrams, even if this also leads to more complex grammatical constructions. On this basis, future work based on longitudinal corpora containing at least 30 or 40 one-hour recordings might be effective in trying to understand grammatical development in children.

BIBLIOGRAPHY

- Ambridge, Ben & Rowland, Caroline F. 2013. Experimental methods in studying child language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science* 4(2), 149-168.
- Behrens, Heike. 2006. The input-output relationship in first language acquisition. *Language and Cognitive Processes* 21, 2-24.
- Brown, Roger. 1973. *A first language: The early stages*. Cambridge, MA: Harvard University Press.
- Chater, Nick & Manning, Christopher D. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences* 10:7, 335-344.
- Croft, William. 2001. *Radical construction grammar*. Oxford: Oxford University Press.
- Croft, William, & Cruse, D. Alan. 2004. *Cognitive Linguistics*. Cambridge University Press.
- Dąbrowska, Ewa. & Lieven, Elena. 2005 Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics* 16: 3, 437-474.
- Demuth, Katherine. 1996. Collecting spontaneous production data. In D. McDaniel, C. McKee, and H. S. Cairns (eds.). *Methods for Assessing Children's Syntax*. Cambridge, MA: MIT Press. 3-22.
- Hsu, Anne S., Chater, Nick. and Vitányi, Paul M. B. 2013. Language learning from positive evidence, reconsidered: A simplicity-based approach. *Topics in Cognitive Science* 5: 1, 35-55.
- Lieven, Elena, Salomo, Dorothé & Tomasello, Michael. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics* 20: 3, 481-508.
- Maslen, Robert, Theakston, Anna, Lieven, Elena & 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.

- MacWhinney, Brian. 2000. *The Childes Project: Tools for Analyzing Talk*, 3rd Edition. Psychology Press.
- Peigneux, Philippe, Laureys, Steven, Delbeuck, Xavier, & Maquet, Pierre. 2001. Sleeping brain, learning brain. The role of sleep for memory systems. *NeuroReport* 12(18). A111.
- Peters, Ann M. 1983. *The units of language acquisition*. New York, NY: Cambridge University Press.
- Roy, Deb, Patel, Rupal, DeCamp, Philip, Kubat, Rony, Fleischman, Michael, Roy, Brandon, & Gorniak, Peter. 2006. The Human Speechome Project. In P. Vogt, Y. Sugita, E. Tuci, & C. Nehaniv (Eds.). *Symbol Grounding and Beyond*. Vol. 4211. 192–196. Springer Berlin Heidelberg.
- Saffran, Jenny R., Aslin, Richard N., & Newport, Elissa L. 1996. Statistical learning by 8-month-old infants. *Science* 274. 1926–1928.
- Snyder, William. 2007. *Child Language: The Parametric Approach*. Oxford: Oxford University Press.
- Song, Jae Yung, Sundara, Megha, & Demuth, Katherine. 2009. Phonological Constraints on Children's Production of English Third Person Singular -s. *Journal of Speech, Language, and Hearing Research* 52: 3. 623–642.
- Tomasello, Michael & Stahl, Daniel. 2004. Sampling children's spontaneous speech: how much is enough? *Journal of Child Language* 31: 1. 101–121.

NOTES

1. Downloaded at <https://childes.talkbank.org/access/Eng-UK/Thomas.html> <https://doi.org/10.21415/T5JG64>
2. Downloaded at <https://childes.talkbank.org/access/Eng-NA/Brown.html> <https://doi.org/10.21415/T5HK5G>
3. Downloaded at <https://phonbank.talkbank.org/access/Eng-NA/Providence.html> <https://doi.org/10.21415/T5R30X>
4. First, it would be difficult to predict at what age the child data would be relevant to the study, and so it would be difficult to organize successful data collection for a small corpus. Second, with a small corpus, there are many infrequent grammatical situations that cannot be studied. You are limited to what happened at the time of data collection.

ABSTRACTS

Dense corpora have been put forward as necessary tools for corpus studies of language acquisition. Despite their great interest, they are not yet frequently used, probably because of the high cost involved in their creation. The goal of the present study was to predict the optimal size of a dense longitudinal corpus when used to infer, manually or automatically, the details of lexical or syntactic development in child language. The results show that corpora of at least 30 to 40 one-hour recordings are necessary, but that longer corpora using the same protocol provide little new information. Dense corpora are indeed very useful, but do not need to be overly large to study grammatical development. This has important consequences for corpus-building

projects, which can be optimized. The existence of a limit to the amount of information provided by large corpora also has important consequences for linguistic theory, as this helps locate the threshold between learning frozen forms and generalizing knowledge about language structure.

Les corpus denses sont souvent présentés comme des outils incontournables dans les études d'acquisition du langage. En dépit de leur grand intérêt scientifique, ils ne sont pas souvent utilisés en raison de leur coût important. Le but de cet article est de prédire la taille optimale d'un corpus dense longitudinal utilisé pour modéliser, de manière automatique ou non, le développement langagier. Les résultats montrent que des corpus d'au moins 30 à 40 sessions sont nécessaires, mais que de plus grands corpus utilisant le même protocole de recueil n'apportent pas beaucoup plus d'information. Il apparaît donc que les corpus denses sont très utiles, mais n'ont pas besoin d'être immenses. Ce résultat a des conséquences importantes pour la mise en place de projets scientifiques, qui peuvent de ce fait être optimisés. Il a également des conséquences pour les théories langagières, car il permet de pointer la frontière entre l'apprentissage massif de formes figées et la capacité de généralisation des connaissances langagières.

INDEX

Mots-clés: Corpus dense, taille optimale d'un corpus, développement lexical et grammatical

Keywords: Dense corpus, optimal corpus size, lexical and grammatical development

AUTHOR

CHRISTOPHE PARISSE

INSERM, MoDyCo, CNRS & Université Paris Nanterre