Child language cuts capers
Dijk, Marijn Wouter Geertje van

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2004

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 5

Disentangling behavior in early child development

Interpretability of early child language and the problem of filler syllables and growing utterance length

M. van Dijk and P. van Geert
Submitted for publication

Abstract
The study of early child development is often hampered by problems of interpretability of behavioral categories. This article discusses problems with interpreting early child language. We show that the solutions for solving ambiguities in observations may have considerable effects on the developmental curves that result from our observations. We offer a procedure for investigating the magnitude of the effect of decisions about how to categorize uninterpretable utterances. This so-called “what-if procedure” compares the effects of worst-case scenarios, i.e. scenarios where either all or none of the uninterpretable utterances are counted as real words. We show that the language development curves are relatively insensitive to how exactly uninterpretable utterances are scored. We end with a number of suggestions regarding the generalization of our procedure to other fields of early development.
5.1 Introduction

5.1.1 Interpretability of early child speech
In the study of early development, the interpretability of behavior poses a major methodological problem. Take for instance the development of smiling (see de Weerth, van Geert & Hoijtink, 1999, for an example). Whereas the smile of a twelve-month-old infant will definitely and unmistakably be interpreted as a smile, “smiles” of a newborn or a one-month-old may pose considerably more problems. Is the observed raising of the corners of the mouth really a smile? While counting the number of smiles in a very young infant, the investigator has to decide whether a specific grimace actually counts as a smile or not. If in doubt, the investigator may decide not to count the observed facial pattern as a smile. Such doubts will in general no longer occur with older infants. Since smiling is known to develop during the first months of life, we expect the frequency of smiles to increase over that period. However, with the first smiles being considerably more difficult to interpret and assuming that doubtful smiles are not incorporated in the frequency counts, we might end up with the undesirable situation that the observed growth in the number of smiles reflects the increasing interpretability of facial expressions instead of the actual increase of the number of smiles. The problem of interpretability is often solved—or should we say by-passed—by training observers until they reach an acceptable level of consensus (see van Geert and van Dijk, 2003, chapter 6 of this thesis). However, consensus does not solve the interpretability problem, it only solves the problem of agreement between observers.

In this article, we will focus early child language where the problem of interpretability is considerable and widely recognized. We will suggest a simple and generalizable way to solve this problem.

5.1.2 Utterance length as an index of language acquisition
The concept of utterance length as a rough index of the child’s level of acquisition, is often used in the study of early child language. In this context the most widely accepted measure is Mean Length of Utterance (or MLU). Brown proposed MLU in 1973 as the best approximation of grammatical complexity in early child language (Shaffer, 1989). He also provided a set of guidelines to calculate MLU by dividing the total number of morphemes (excluding imitations, yes/no answers, ritualized speech such as songs and rhymes, etc.) by the total number of utterances in the first 100 utterances in a sample of spontaneous speech. MLU can also be calculated in words (MLU-w). While MLU in morphemes (MLU-m) and MLU-w are conceptually certainly not identical, research has shown that the two measures are highly correlated in normally developing children (e.g. Thordardottir & Weismer, 1998; Arlman-Rupp, van Niekerk de Haan, van de Sandt-Koenderman, 1976). Because MLU-w is much simpler (both theoretically and in practice), it is often considered the preferred measure of the two (Thordardottir & Weismer, 1998).

Since their introduction, both MLU-m and MLU-w have been widely used. According to Rosenthal-Rollins, Snow and Willett (1996), one of the reasons for this popularity is that MLU is sensitive to a wide range of language aspects, for instance in the fields of morphology, semantics and syntax. Up till
today, MLU is a largely used measure, probably mainly because of its simplicity. For instance, in their study on subject omission, Valian, Hoeffner and Aubry (1996) use the MLU-3 boundary to test two competing hypotheses. McGregor and Johnsen (1997) also use the MLU stages in their study on the development of stress. Dunn (1996) considers MLU in combination with the total number of structural errors and age a solid diagnostic predictor for developmental disorders. There are also numerous studies in which MLU is not the focal point, but where it is used to describe the subjects (e.g., Watson & Scukanec, 1997; Rescorla, Roberts & Dahlsgaard, 1997). Furthermore, many studies on language disorders use control groups that are MLU matched (e.g., Hansson, 1997; Rescorla et al., 1997; Conti-Ramsden & Jones, 1997).

As an alternative to MLU, sentence length is also used to express linguistic development. Where MLU only specifies the average utterance length, sentence length states how many utterances contained only one word, how many contained two words, three words, etcetera. This measure is somewhat more informative because the distribution of the sentence lengths is also explicated.

Naturally, these quantitative measures have also been criticized. This criticism is especially centered around the fact that complexity is mainly expressed in structural elements. By simply counting morphemes or words, this structural complexity is being ignored. Therefore, MLU (and sentence length) is often considered to be an inadequate complexity measure (Frijn & de Haan, 1994). Bates, Bretherton and Snyder (1988), however, state that MLU is an important measure of syntactic development, especially up to the third year of life. After this third year, the reliability of MLU declines because the acquisition of new syntactical knowledge is no longer reflected in a growing utterance length (Rosenthal-Rollins et al., 1996).

There is however, one additional problem with MLU and sentence length, and that is the way they are calculated in early child language. In this article we will argue that there are two reasons to assume that the way they are traditionally counted may result in an under-inflated picture of development. The first reason is that early child language often consists of many partially intelligible utterances of which some are included and others are excluded from calculation. The second reason is the use of the “&”-code for so-called “filler-syllables”, which status is yet unclear.

5.1.3 Partially Interpretable Utterances

Early child speech is often difficult to understand and interpret. From experience, we know that it takes transcribers many hours to “get into” the individual pronunciation pattern of a specific child. However, even with extensive training, there is always a set of utterances that remains uninterpretable. Moreover, there is also a group of utterances in which some elements are uninterpretable, but in which other elements can be interpreted. For instance, in Dutch a child might say “Ik wil xxx” (I want xxx), where the first two words are perfectly understood, while the third element remains unintelligible. We will call this type of utterances Partially Interpretable Utterances (PIUs).

The Childes handbook suggests various solutions to the problems involved in the coding of these uninterpretable elements. First, the Childes’ CHAT-manual (MacWhinney, 1991) recommends the use of the “xxx” code for a set of unintelligible lexical elements with an unclear number of words, and the use of the “xx” code for a single unintelligible word. By doing so, one leaves room for ambiguity in the
transcripts which is definitely preferred over over-interpretation on the basis of an adult language model. The Childes handbook states: “The most difficult bias to overcome is the tendency to map every spoken form by a learner (be it a child, an aphasic of a second language learner- onto a standard lexical items in the adult language” (pp. 4 of the online Childes manual). Transcribers tend to assimilate non-standard learning strings to standard forms of the adult language (MacWhinney, 2003). Thus, we must avoid forcing child speech into adult language categories and instead consider the fundamental differences between both.

An important question is: how can a transcriber decide if the unintelligible utterance contains one or multiple words, and thus choose between “xx” and “xxx”? The transcriber will probably do this by focusing on the number of distinguishable phonetic elements (sounds). As a consequence of this procedure, only the shorter uninterpretable elements will be coded as “xx”. For the phonetically longer uninterpretable units, the transcriber will probably resign to “xxx”.

It is at this point that the eventually problematic consequences of the transcriber’s choice become manifest. In calculating MLU, Childes includes xx- but excludes xxx-coded tags. This leads to the situation that the shorter unintelligible utterances are included while the longer fragments are excluded. It might be argued that context may also contribute to a transcribers tendency to attribute “xx” versus “xxx”. Some contexts may suggest that the child utters only one word. For instance, if a child says “Ik wil xxx” (I want ?), while pointing at something, the transcriber might be inclined to attribute only one word, since the object is clearly intended. However, how can the transcriber be sure that the child did not say “Ik wil xx hebben” (I want xx have)? Therefore, we pose the question: how can the transcriber reliably assign any number of words to a part of the utterance he or she did not understand? In principle, this task amounts to an intrinsic impossibility, which we believe is reflected in often arbitrary choices which may undermine the reliability and validity of the measures of utterance length. To be more specific, the resulting curves of the increase of MLU over age may show a distorted picture of development, in particular since interpretability itself improves considerably over the course of development. Before proceeding to the question of how to solve this difficulty, we will first discuss another source of interpretation problems, namely filler syllables.

5.1.4 Filler syllables
In CLAN there is one other transcription code that might be problematic when employing quantitative measures. This is the “&”-code, which is meant for phonological elements with no lexical meaning, such as stutters and exclamations. However, it is often used for the transcription of so-called “filler-syllables” (also called Prefixed Additional Elements, PAE, see for instance Veneziano and Sinclair, 2000). Filler syllables are monosyllabic, often vocalic or nasalized elements that children add to their word-like productions, usually in the early period of acquisition (Veneziano & Sinclair, 2000). The question is whether these filler syllables can be considered grammatical elements or not. Filler syllables appear in the period of early language acquisition and their possible sources and functions has long been a subject of interest for many researchers. Researchers have been warning against crediting the child with early grammatical knowledge on the basis of the appearance of these elements. For instance, Braine (1963) concludes in his study on subject Steven “while it is quite likely
that these elements are an interesting distillate of the unstressed and phonetically often obscure English articles, prepositions and auxiliary verbs, there is no basis for giving them the morphemic status at this stage in Steven’s development” (as reprinted in 1973; pp. 415). One group of authors (e.g. Dolitski, 1983; Peters, 1990; Veneziano, Sinclair & Berthoud, 1990; Scarpa, 1993; Simmonsen, 1993; Kilani-Schoch & Dressler, 2000) link filler syllables more specifically to the child’s development of grammatical morphemes, considering them “an intermediate form on the way to grammatical morphemes” (Veneziano & Sinclair, 2000; pp. 463).

The discussion on filler syllables has received increased attention since Peters (2001) published a paper on the status of these elements in emerging grammar. In this paper, she claimed that it is especially difficult to integrate these filler syllables into theories of language acquisition, since they do not neatly fit into linguists’ notions about modules of language. Furthermore, Peters suggest that it is time to propose a reasonable set of criteria for identifying them, and suggests an approach to further studying them. The author suggests to distinguish two types of filler syllables: (1) premorphological, whose presence in the utterance is motivated by purely phonological considerations, and (2) protomorphological, which functions as placeholders for grammatical morphemes and eventually differentiate into various grammatical morphemes. It is also emphasized that the status of filler syllables has theoretical implications, for instance the contrasting predictions from nativist versus constructivist accounts. As a reaction to the Peters’ paper, Dabrowska (2001) suggests the presence of a third type of filler which sometimes appears when a child begins to generalize over a set of related words. Just as protomorphological fillers can provide valuable information about how children acquire function words, these generalized fillers may offer insight in how children form categories of function words.

The interpretation of filler syllables is also related to the problem of interpretability, we discussed above. For instance, in Dutch a child might say “ik &6 wil &6 bal”, the &6 representing the “schwa”. However, it might be the case that this second “&6”-element is an early article, since the schwa is phonetically closely related to the article “de”. The basis for assigning “&6” versus “de” can in practice be very small, since it depends on the presence of a sometimes very subtle “d”-sound. Naturally not all &6-elements bear a possible meaning, but excluding all of them over the wide scope of an entire transcript would most definitely result in an underestimation of utterance length. For instance, in Dutch, the presence of this &6 category is sometimes abundant, especially in the stage from two-word sentences onwards.

5.1.5 Consequences for calculating MLU and sentence length

As we mentioned before, Childes conventions exclude the strings “xxx” and “&” for further calculation. This means that the utterance “ik wil xxx” will be regarded as a 2-word utterance. We know, however, that this is most likely an underestimation of the real number of words. The utterance most likely contains three, four or more words, and not two. However, although we can be sure that the utterance contains more than two words, we have no idea how many more. In the analysis of the development of the number of 1-word, 2-word, and 3-word utterances, we should expect to find a degree of underestimation to the advantage of the categories that specify smaller numbers of words. For
instance, we suspect that in the category of 3-word utterances, there are also a number of utterances that actually contain 4 or 5 words, and in the 2-word category there are a number of utterances that contain 3 words, etcetera. We have no reasons to expect that, intrinsically, one of the categories is more susceptible to this underestimation than the other. Relatively speaking, the 2-word category contains as many “false” classifications as the 3- or 4-word category. However, we do expect that the language of the child becomes better interpretable with age. Because pronunciation improves, the frequency of “xxx”-classifications declines with age. It is also known that the number of filler syllables declines when the syntactical abilities become more advanced. And because the 3-, 4- and more-word utterances appear relatively late in the developmental trajectory, we might expect that especially the earlier recordings (for instance before the age of 2) show this underestimation, while the effect diminishes with age.

5.2 An empirical study of the growth of 1- and 2-word utterances

As an illustration we will show the developmental trajectory of 1- and 2-word utterances of one child. The subject is Heleen, a Dutch girl who participated in a longitudinal study on variability in early language development (see for details van Geert & van Dijk, 2002).

5.2.1 Method

In the course of a year (from age 1;6 to 2;6) Heleen’s language development was followed employing 60-minutes-recordings of spontaneous speech. These speech samples were collected in the child’s home, while child and parent engaged in normal daily routine. Heleen had access to her own toys and could move freely across the room. Recordings were made by means of a video-camera that was placed in a corner of the room, overviewing much of the living room space. A separate “wide-angle” microphone was connected to the camera in order to improve recording quality. The quality of the recordings turned out to be fairly good, only a small portion of fragments (less than 5%) was interrupted by environmental noise (such as a garbage truck loading outside). Such fragments were excluded from further analysis.

In this case-study, the independent inter-observer reliability was calculated as inter-observer overlap of exact utterance length of individual utterances. In a subset of all transcribed material this amounted to 237 utterances of the total set of 268 utterances (88%), which, in our opinion, is fairly good.

5.2.2 Subject

Heleen is the first-born (and, during the observation period, only) child of middle class parents. The family lives in a suburban neighborhood in an average-size city in the North of the Netherlands. Heleen was raised in a monolingual Dutch environment. The family does not speak any apparent dialect. Heleen’s general cognitive development was tested with the Bayley Developmental Scales 2/30 (van der Meulen & Smrkovský, 1983) a few months before her second birthday. She scored
within the normal range (OI=100). In the beginning of the study, Heleen predominantly used one-word utterances, while at the end of the study her language showed various characteristics of the differentiation stage (see for characteristics of the Dutch differentiation stage Gillis and Schaerlaekens, 2000).

5.3 Results

Figure 5.1 shows the development of the relative occurrence of 1- and 2-word utterances in 25 two-weekly samples of Heleen’s language development. In summary, figure 5.1 shows a decline of the relative number of 1-word utterances and an increase of the number of 2-word utterances. Considering the underestimation problem we discussed before, we do not know how to interpret these curves. So far, it is possible that the trade-off between 1- and 2-word utterances is caused by the increased interpretability of the child’s utterances. It might be the case that while we heard the child say “xxx bal” (? ball) in the earlier recordings and interpreted this as a one-word sentence. In the later stages we hear the same utterance as “dat is mooie bal” (that is beautiful ball), which we interpret as a four-word utterance. In this particular case, the increase in interpretability causes the number of 1-word utterances to decline and the number of more-word utterances to increase. To put it differently, the increase in more-word sentences is at least partly, an artifact of the increase in phonetic interpretability in the child.

![Figure 5.1](image-url)  
**Figure 5.1.** The proportional occurrence of 1- and 2-word utterances (1 is 100% of all utterances) of Heleen (1;6-2;6) (1 represents 100% of all utterances).
5.3.1 Solutions for eliminating the interpretability problem

A first suggestion for dealing with this problem might be to eliminate all PIUs from the transcripts and only analyze the remaining utterances in which all the words were intelligible. CLAN offers this possibility using the postcodes [+ PI] (p. 100 of the online Childes manual). Although this approach is considered an adequate solution, we believe that this approach has some fundamental drawbacks. When using this technique for very young children, for instance, the remaining sample would most likely be relatively small, since many utterances contain at least one uninterpretable component or filler. In our case-study we calculated that of the earliest files, 62% contained either unintelligible units, or instances of the &-code that might as well have contained fillers. Of the intermediate and, later files, around 40%, and 46% of all utterances respectively belonged to this category. Although the practice of excluding PUIs can be refined, so as to exclude a smaller percentage of utterances than we reported, we still have to question to what degree the remaining sample is selective to other linguistic factors. More importantly, if one considers the ambiguity a fundamental characteristic of child speech (an opinion shared by a constructivistic and dynamic systems approach) we would remove one of the essences from child speech out of the samples. The elimination of this characteristic would result in a very unrepresentative sample of child speech, and the question is whether the analysis of this sample has any relevance. However, if we state that the PIUs should be included in our sample, what status do the uninterpretable units have?

5.3.2 The use of the what-if procedure, in particular the worst-case-scenario

Although it might not result in a cut-clear solution to the problem, it is worthwhile to investigate how large the under-inflation effect might be. In case of a worst-case scenario, where all the excluded codes would have been actual words, how different would the resulting developmental trajectory be? We can address these questions by using a procedure that actually sketches this worst-case scenario, which is a procedure that assigns a fixed value of one word to every uninterpretable unit and possible filler in the PIUs. For instance, we can assign the value of one word to each appearance of xxx in the transcripts. In this case the utterance “ik wil xxx” would be considered a 3-word utterance, and the utterance “ik &6 wil &6 bal” a five word utterance. This kind of approach is more commonly known as a what-if procedure, which is often applied in case of uncertainty about important parameters of a process or future event. The use of a worst-case scenario is comparable to the use of confidence intervals: by taking empirically unlikely but theoretically tenable boundaries for uncertain parameters, one obtains an idea of how seriously such unexpected but possible extremes can affect the outcomes of an investigation.

Because we can still never be sure that each instance of xxx and & refers to only one word, a degree of uncertainty remains even after adding the xxx- and &-forms to our word count. However, we should note that &-strings are used to represent many different things, ranging from exclamations, stutters, to filler syllables as reported by Veneziano and Sinclair (2000). Assigning a fixed value to these units most probably leads to an (empirically unlikely) overestimation of the number of words in each utterance. However, with this procedure we only want to get an impression of the under-inflation effect.
in a worst-case scenario. Therefore, the new values acquired using this procedure can only be interpreted in combination with the original values.

In summary, the fixed value procedure assigns the value of one word to every occurrence of xxx and &. We know that this most likely results in an over-inflation of many utterances, while we are sure that the original scores underestimate the utterance length. Plotting these values together in a graph provides an estimation of the magnitude of the interpretability problem. If the trendline of the original values differs considerably from the values after assigning fixed values, we know that the problem might be large, and that we need to be careful in the interpretation of the resulting curves. If the trendlines are close together, the problem might be smaller than expected. Analyzing these lines as a bandwidth of scores also facilitates the analysis of the developmental trajectory.

5.4 Illustration: a what-if or worst-case analysis of our language data

In order to check the effect of the different ways of counting the number of words and utterances, we applied the combined (original and fixed values) procedure to Heleen’s language data.

5.4.1 Utterance length

Figure 5.2 depicts the resulting trajectory of utterance length. Visual inspection shows that the differences between the original scores and the fixed-value scores seem relatively small. We also see only slight widening or narrowing of this bandwidth, which means that the interpretability effect remains roughly the same for the entire developmental trajectory. This indicates that the results based on the original values and the results based on the fixed-value procedure are—surprisingly enough—not very different. The result is even more contra-intuitively if one considers the fact that after assigning the fixed values, the resulting total number of utterances on which the calculations are based increases sometimes dramatically. This is the case because after assigning a value to a dummy (an utterance consisting of only uninterpretable units) this unit ceases to be a dummy and adds one to the total number of utterances (which is the divisor, used to calculate mean length of utterances). The differences between these totals range from 62 up to 471 utterances, which is large. Also, this difference declines with age. Obviously there are more totally uninterpretable utterances (dummies) in the earlier files, which is not too surprising, since early child language is known to become more understandable with age. Very early child speech (before the age of two) often contains many uninterpretable utterances.

How can the finding that the two trajectories are surprisingly similar be reconciled with the fact that early utterances are considerably less interpretable than later ones? The answer to this question probably lies in the fact that by applying the fixed value procedure, one not only increases the number of words in the early utterances, but also the number of one-word utterances (transformed dummy utterances). Thus, on average, the ratio of 1- over 2- and 3-word utterances does not change. Instead,

---

11 The largest differences were, naturally, found in the earliest samples.
there is a balance in the trade-off between the different utterance categories. In itself it is comforting to know that the differences between the over-inflated (acquired after assigning fixed values) and the underestimated (based on the original values) are small. This can only imply that our initial data analysis is fairly adequate in the sense that the observed decrease of 1-word utterances and the increase of 2- (and 3-) word utterances remain intact. We can now assume that this is not solely caused by an increased degree of interpretability, as we speculated before, but that it reflects a real growth in utterance length.

The same reasoning applies to MLU-w (Mean Length of Utterance in words, see Brown, 1973). We calculated Heleen’s MLU-w using the same procedure as described above, first calculating the original values and thus excluding the uninterpretable units, and second, calculating MLU-w after assigning a fixed value to these units.

5.4.2 Mean Length of Utterance

Figure 5.3 shows the results of both procedures (MLU-w based on the original values and after assigning the fixed values). Again, the resulting trajectories are very close together, which means that the differences in the results of the original MLUs and the MLUs-after-adding-fixed-value are small. Note that the slight difference between both lines can go in either direction: at some points MLU-original is slightly lower than MLU-fixed, and at some points it is higher. Although it seems strange that MLU-original (which was assumed to underestimate utterance length) can be the highest of the two, one should realize that there is always a trade-off between 0-, 1-, 2-3-, and more-word utterances. In the case were the MLU is lower after assigning fixed values, there are proportionately a large number of original 0-word utterances (dummies) that have become 1-word utterances. These were not included in the original calculation, but are included after assigning this fixed value. In this case, the MLU-divisor (which is the total number of utterances) increases more than the MLU-denominator (the
total number of words), thus resulting in a smaller MLU-w. However, visual inspection of figure 5.3 shows that there is an almost perfect tradeoff between the two calculations. Again, it is comforting to know that the increase in MLU is not solely caused by an increased degree of interpretability.

![Figure 5.3](image-url) Heleen’s development of MLU-w based on the original and fixed value scores.

### 5.5 Conclusion

With regards to a solution to the problem of interpretability, filler-syllables and growing utterance length we outlined above, we suggest that this procedure of combining two trajectories might be fruitful. The two trajectories can be conceptualized as a bandwidth, with lower bounds based on interpretable words only and upper bounds based on fixed corrections for uninterpretable words. In our illustration, we used the most simple and rough way to overestimate utterance length. We did this in order to outline the results of a worst-case scenario. The procedure can (and probably should) be refined in order to win credibility, since naturally not all the elements to which we assigned a fixed value can reasonable be assumed to be meaningful. For instance, with the use of postcodes we can define which -codes are definitely not lexical (e.g. exclamation), and which ones are possible fillers. This way, we can refine our transcripts by identifying which uninterpretable elements are likely to contain words.

Finally, even while we cannot be sure that the resulting bandwidth of curves gives an accurate picture, we think that it gives us a better understanding of the effect of the interpretability problem on the quantitative representation of early child speech. We have also shown that, in the case of our subject, the bandwidths are relatively narrow if the proportions of 1- over 2- and more word sentences are plotted. A similar conclusion holds for MLU, which is also a proportional measure. This suggest that in this case, one might adequately use the original values of utterance length to represent to developmental trajectory. However, this might not always be the case. Some children speak more slurry than others, and some recordings have a better quality than others, depending on the recording
circumstances. Also different languages may differ with regard to the interpretation of filler syllables. In these cases, the sketching of the interpretability-bandwidth as we suggested above, might provide a better understanding of to what degree interpretability and filler-syllables distort the trajectory of development.

We began this article by observing that the study of early child development can be seriously hampered by the fact that early behavior is sometimes difficult to interpret. Choices must be made about how uncertainties about the categorization of behaviors must be solved. Since interpretability itself depends on the developmental level of the behavior under observation, such choices may seriously affect the outcomes of our investigations and lead to misrepresentations of the underlying growth curves. Since we believe that problematic interpretability is an essential feature of early behavior, it cannot be solved by additional training of observers or ever increasing refinement of criteria for categorizing behaviors. Instead, we argue for a simple procedure that is often applied in cases where uncertainty about essential parameters occurs, namely the application of what-if procedures, more in particularly based on worst-case scenarios. In our example from child language, we defined the worst-case scenarios as those in which either none or all of the uninterpretable utterances of the uninterpretable utterances referred to a word. We compared the curves resulting from these extreme cases with each other and found that they did not contain serious differences. This finding supported the conclusion that our results are relatively insensitive to the choices we have made – in accordance with the Childes recommendations – with regard to transcribing the uninterpretable utterances. With our worst-case scenarios, we in fact specified confidence intervals for the curves that result from our observations.

We will conclude our discussion with three remarks aimed at generalizing the method presented in this article. First, the procedure used to test what-if questions and in particular worst-case scenarios depends on the nature of the observed phenomena and the peculiarities of the observation procedures. Thus, a procedure that works well with language data is not necessarily adapted to problems with observing emotional expressions or cognitive and motor skills. Further work needs to be done to develop what-if and worst-case procedures for fields other than language development.

Second, testing a worst-case scenario is akin to looking for statistical confidence intervals. Whereas such intervals can be calculated from known or assumed statistical distribution properties, the specification of worst-case scenarios requires a combination of theoretical and empirical insights. A worst-case applies to a condition that is theoretically possible (let us say, “reasonable”) but empirically unlikely (similar to, for instance, a 97.5 percentile case in a 95% confidence interval). For instance, we assumed that in the worst case every uninterpretable phoneme counts as one word, which is theoretically possible but quite unlikely. We did not assume that every such phoneme counts as three words, for instance, which would also be unlikely but also theoretically unfeasible. The specification of adequate worst-case scenarios for other variables than language requires further attention.

Finally, what if the application of worst-case scenarios leads to the conclusion that the resulting growth curves – or anything else one wishes to extract from the data – are highly sensitive to the decisions made to solve the ambiguities and lack of interpretability in the data? With our language data, we
found that the results were quite insensitive to such decisions, which improves the credibility of our conclusions. If results depend very much on the way in which the interpretability problems have been solved, the researcher is faced with a problem of reliability, but possibly also with a problem of validity in the sense that it is possible that the categories chosen for the interpretation of the behaviors are just not adequate.