Available online at www.sciencedirect.com
SCIENCE DIRECT®
Infant Behavior & Development 28 (2005) 99–117
Infant Behavior & Development

Disentangling behavior in early child development: Interpretability of early child language and its effect on utterance length measures

Marijn van Dijk [a,*], Paul van Geert [b]

[a] *Open University of the Netherlands, Regional Study Center Zwolle, Campus 2-4, 8017 CA Zwolle, The Netherlands*
[b] *University of Groningen, Department of Psychology, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands*

Received 7 September 2004; received in revised form 23 November 2004; accepted 10 December 2004

Abstract

Early child speech is often difficult to understand and interpret. Usually, these unintelligible units are not included in quantitative measures, such as MLU. In this paper, we claim that these interpretation problems have an unknown effect on utterance length measures (such as MLU), since we have no knowledge on how the unintelligible units are distributed across the speech sample. We offer a procedure for investigating how big the effect of specific coding decisions is on quantitative language measures (number of sentences, MLU and sentence length). 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 explore whether the application of such a worst-case scenario leads to different results in the (longitudinal) language data of two infants. These data show that there are obvious inter-individual differences, a finding which implies that the effect of interpretability should not be ignored a priori. We discuss the potential meaning of these differences for understanding the underlying developmental processes and end with a number of suggestions regarding the generalization of our procedure to other fields of early development.

Keywords: Interpretability; Early child language; Filler syllables; Mean Length of Utterance; Methodology; Transcription procedure

* Corresponding author. Tel.: +31 38 4658333; fax: +31 38 465 8224.
E-mail addresses: marijn.vandijk@ou.nl (M. van Dijk), vangeert@inn.nl (P. van Geert).

0163-6383/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.infbeh.2004.12.003

1. Introduction

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 & Hoitink, 1999, for example). Whereas the smile of a 12-month-old infant will definitely and unmistakably be interpreted as a smile, "smiles" of a newborn or a 1-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 probably occur much less 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 & van Dijk, 2003). However, consensus does not solve the interpretability problem, it only solves the problem of agreement between observers (e.g. observers can be taught to agree on how to divide a continuous grayscale into black and white). Even if maximal agreement between observers is reached, it is likely that a certain part of behavior will remain uninterpretable, and will consequently not be included in further analyses. In this article, the focus is on *early child language*, where the problem of interpretability is considerable and widely recognized. We will investigate to what degree arbitrary coding decisions influence the developmental trajectories of absolute measures, such as sentence counts and proportional measures, such as MLU, by suggesting a simple and generalizable way to approach this problem. In the conclusion, some possible extensions of our method to other domains of infant research will be addressed.

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. A widely accepted measure is Mean Length of Utterance (or MLU). Brown (1973) proposed MLU 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. MLU is still a widely used measure, probably mainly because of its simplicity. For instance, in their study on subject

omission, Valian, Hoeffner and Aubry (1996) use the MLU-stages to test two competing hypotheses concerning subject omission in an elicited imitations task (to be more specific, to test a competence-deficit versus a performance-deficit hypothesis). McGregor and Johnson (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 as 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 (total numbers or percentages of all utterances) contained only one word, how many contained two words, three words, etcetera. This measure is more informative in the sense that the distribution of the sentence lengths is also made explicit.

Both quantitative measures have been criticized. This criticism is especially centered on the fact that complexity is a matter of structural components and not of quantitative magnitude. By simply counting morphemes or words, this structural complexity is being ignored. Therefore, MLU (and sentence length) is often considered 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, also a *methodological* problem with MLU and sentence length, which concerns the way they are calculated in early child language. In this article, we will argue that there are two reasons for assuming that the way they are traditionally counted may result in a distorted picture of their development. The first reason is that early child language often consists of many Partially Intelligible Utterances in which the unintelligible sounds are not counted as words, which implies that part of the utterance is excluded from calculation. The second reason is that some uninterpretable forms are considered to have a special function or status, namely that of filler-syllables, the status of which is yet unclear (i.e., it is unclear to what extent they are real words). We speculate that the exclusion of both unintelligible speech, and filler syllables (which is common in child language research) might lead to an underestimation of sentence length measures, such as MLU.

3. Partially Interpretable Utterances

Early child speech is often difficult to understand and interpret. It is known from experience 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. These utterances will be called *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 code "xx", for a single unintelligible word. By assigning these codes, one leaves room for ambiguity in the transcripts and this is definitely preferred over over-interpretation on the basis of an adult language model. The online CHILDES manual states: "The most difficult bias to overcome is the tendency to map every spoken form by a learner (be it a child, an aphasic or a second language learner) onto a standard lexical items in the adult language. Transcribers tend to assimilate non-standard learning strings to standard forms of the adult language" (pp. 4) (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 transcriber's 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 xxx), 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 is impossible to accomplish, and trying to pursue it anyway is reflected in often arbitrary choices that 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, another source of interpretation problems, namely filler syllables, will first be discussed.

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 have 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, in his study on subject Steven, Braine (1963) concludes "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; 415). One group of authors (e.g. Dolitsky, 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; 463).

The discussion on filler syllables has received increased attention since Peters (2001) published an article on the status of these elements in emerging grammar. Peters (2001) 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 suggests that it is time to propose a reasonable set of criteria for identifying them, and proposes an approach to further studying them. The author suggests to distinguish two types of filler syllables: (1) pre-morphological filler syllables, whose presence in the utterance is motivated by purely phonological considerations, and (2) proto-morphological filler syllables, which function 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 for the contrasting predictions from nativist versus constructivist accounts. As a reaction to the Peters' article, Dabrowska (2001) suggests the existence of a third type of filler that sometimes appears when a child begins to generalize over a set of related words. Just as proto-morphological 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. For instance, in Dutch a child might say "ik &6 wil &6 bal" (I &6 want &6 ball), 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 the article "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 will 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 2-word sentences onwards.

In summary, a filler syllable is a kind of precursor of a word: it is a word and it is not a word. This issue of ambiguity is discussed in an earlier article (Van Geert & van Dijk, 2003) in which the present authors invoked notions of fuzzy logic to account for ambiguity. Ambiguity is expressed in terms of (identifiable) class membership, i.e. a degree of membership to the category "word". The existence of fuzzy or ambiguous cases, filler syllables, for instance, sheds another light on the issue of countability: should a word be counted as a word if it is neither a real word nor a real non-word?

5. Consequences for calculating MLU and sentence length

As was mentioned before, CHILDES conventions exclude the strings "xxx" and "&" for further calculation. This means that the utterance "ik wil xxx" will be interpreted as a 2-word utterance. However, this interpretation 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. The same reasoning applies to the utterance "ik &6 wil &6 bal", which is counted as a 3-word utterance, but may also contain three or even

four words. If this is the case, we should also expect to find a degree of *underestimation* in utterance length measures. For instance, we suspect that in the category of 3-word utterances, there are also a number of utterances that actually contain four or five words, and in the 2-word category there are a number of utterances that contain three words, etcetera.

Researchers have no a-priori knowledge of whether one of the categories is more susceptible to this underestimation than the other. Relatively speaking, the 2-word category may contain as many "false" classifications as the 3- or 4-word category. Also, in advance, researchers have no idea of the impact of the 0-words category (dummies), which are utterances that contain no intelligible elements. Children who speak more slurry than others, may have a relatively large number of dummies. For these children, the influence on MLU can be relatively big. On the other hand, it is likely that the language of a 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, it might be expected that especially the earlier recordings (for instance before the age of 2) show this underestimation. It is therefore reasonable to assume that the effect of a large dummy category diminishes as the child grows older.

5.1. Solutions for eliminating the interpretability problem?

A first suggestion for dealing with this problem is 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] (pp. 100 of the online CHILDES manual). Although this approach is considered an adequate solution, it has some fundamental drawbacks. When using this technique for very young children, for instance, the remaining sample will 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 (when subject Heleen was 18 months old), 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 partially interpretatble utterances can be refined, such 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 ambiguity is indeed a fundamental characteristic of child speech (an opinion shared by constructivist and dynamic systems approaches) excluding such PIU's 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.

5.2. Missing values procedures

The problem of interpretability of Partially Interpretable Utterances can eventually be considered as a problem of *missing values.* Missing values are very common in psychological studies (surveys, experiments, etc.). Usually, a data set is conceived of as a rectangular data matrix, with the rows representing the subjects and the columns representing the variables of the study. The values in the matrix can be real numbers (for instance age), or numbers representing categories (such as gender and ethnicity). The entries in this matrix that are missing because the values at issue have not been observed, are called missing

values. In further statistical analyses, missing values usually receive a specific code (for instance the value of 9 for "not observed", 8 for "do not know", etc.). Statistical packages, such as SPSS, often exclude missing values from further analysis, although this is considered inappropriate by many statisticians (see Little & Rubin, 1987) because inferences are made about the entire target population and not a sub-group. Missing values can be either random (in the sense that there is no systematic drop-out, for instance a random group of subjects forgot to check a box) or non-random (drop-out is possibly systematic, for instance low income subjects refuse to report their monthly income). The statistical analysis of missing values has been given increasing attention since the classical work of Little and Rubin (1987). They review methods proposed in the literature of Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972), Dempster, Paird, and Rubin (1977) and Little (1982), and have grouped these into (1) procedures based on completely recorded units, which thus discard incomplete records and analyze only the units with complete data, (2) the imputation-based procedures where the values are filled in and the resultant complete data are analyzed by standard methods, (3) weighting procedures, and (4) the model-based procedures which are based on a generated *model* for the partially missing data and are basing inferences on likelihood under that model (with parameters estimated by procedures such as maximum likelihood).

However, the problem is that it is extremely difficult to verify the randomness of the unknown data (Navarro Pastor, 2003). In the last 20 years, many researchers have assessed the requirements of different methods for the analyses of incomplete data, showing that different types of strategies (such as single imputation, complete-case or list-wise analysis, maximum likelihood and multiple imputations) require, for generalizable results, that the missing values are missing at random (e.g. Graham, Hofer & Piccinin, 1994). Navarro Pastor (2003) adds that "In practice it is difficult to check the randomness of missing data because it requires knowing the unknown variables" [pp 364].

In child language, we can consider the total set of sounds as the dataset. All transcribed words can be conceived of as values in the data matrix (for instance, every transcribed word has the value of 1). In this matrix, the rows represent a language sample for which utterance length is calculated, and the columns are all the sound units that were transcribed. The uninterpretable units either receive the xx, xxx or the &-code, depending on the type of sounds, and are ignored in calculating sentence length. Because sentence length and MLU are proportional measures, ignoring such units is not a problem *if and only if* missing values are randomly distributed across the 1-, 2-, 3-(etc)word categories. However, we would first have to establish if this is the case, or to speak in terms of Navarro Pastor (2003), we would first have "to acquire knowledge of the unknown variables". In summary, there is no readily available missing data procedure that will solve the problems of Partially Interpretable Utterances (PIUs). In the remainder of this article, a procedure will be introduced that focuses on the effect of different interpretation methods on the resulting images of developmental trajectories, rather than on trying to infer or impute missing, unintelligible or ambiguous utterances.

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

In order to acquire a first impression of how the unintelligible utterances (missing values) are distributed across the utterance categories, the following question is asked:

> *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 call it the "worst-case" because it is the case that implies the greatest possible counting error). We can address this question by using a procedure that actually implements 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. In this case, we assign the value of one word to each appearance of xxx and every &- string 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 5-word utterance. We can compare this worst-case scenario to the original data, that might be considered a *best-case scenario*, a situation in which none of the unintelligible units are meaningful units (which we call the "best-case" because it implies that our original word count has no errors).

The use of a worst- or best-case scenario 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. What-if procedures compare empirically unlikely (or not very likely) but theoretically tenable boundaries for uncertain parameters in order to obtain an idea of how seriously such unexpected but possible extremes can affect the outcomes of an investigation. Because we can still not 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 overestimation of the number of words in each utterance. However, the aim of this assignment procedure is *not to* provide a better (i.e. more correct) estimation of the number of words, but to give an idea of an upper bound, which is the number of utterances that corresponds with the worst-case scenario. That is, this procedure aims at acquiring a rough impression of the degree of underestimation of the number of words (or any other counted or measured variable, for that matter) in a worst-case scenario. The information it provides is only meaningful if it is combined with the original values acquired from the standard procedure (in which none of the xxx- or &-strings are included in the analysis).

In summary, the worst-case scenario assigns the value of one word to every occurrence of xxx and every &-string. We know that this most likely results in an *overestimation* of many utterances, while we have argued that the original scores possibly *underestimate* the utterance length. Comparing the values of both procedures provides a preliminary "knowledge of the unknown", as introduced in the discussion of the missing variables issue. If the original counts differ considerably from the counts after assigning fixed values, we know that we need to be careful in the interpretation of the original data. If the values are closer together (or even identical), the original values can probably be used for further analysis as a reliable estimation of the child speech variable at issue. In addition to providing an idea of how reliable the data are, the worst-case scenario can eventually provide information about underlying aspects of development.

6. An empirical study of the growth of 1-, 2- and 3-plus words utterances

6.1. Subjects

The subjects in our case study are Heleen and Jessica, two Dutch infants who participated in a longitudinal study on variability in early language development (see for details Van Geert & van Dijk, 2002). In the course of a year (from age 1;6 to 2;6) the language development of these infants was followed

employing 60-min recordings of spontaneous speech. Heleen's and Jessica's families live in a suburban neighborhood in an average-size city in the North of the Netherlands and were raised in a monolingual Dutch environment. The families do not speak any apparent dialect. Heleen's and Jessica's general cognitive development was tested with the Dutch version of the Bayley Developmental Scales 2/30 (Van der Meulen & Smrkovsky, 1983) a few months before their second birthday. Both subjects scored within the normal range. In the beginning of the study, the infants predominantly used 1-word utterances, while at the end of the study their language showed various characteristics of the differentiation stage (see for characteristics of the Dutch differentiation stage Gillis and Schaerlaekens, 2000).

6.2. Method

All child language was transcribed according to Childes conventions, (MacWhinney, 1991). This was conducted by the first author (who is an experienced transcriber) and two graduate students (one per subject), who had received an intensive training. 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 0.84 (Heleen) and 0.88 (Jessica), which is adequate. All files were checked by the first author. Mean Length of Utterance in words (MLU-w) was calculated by dividing the number of words in the total samples by the total number of utterances (all samples contained over 100 utterances, most samples counted 200–400 utterances). Sentence length was determined by counting the frequencies of all utterances with 1-, 2,- or 3-plus words and dividing it by the total frequency of all utterances in the sample. Thus, the sentence lengths express which part of all utterances contain 1-, 2- and 3-plus word utterances. For instance a value of 0.5 of the 1-word utterances means that 50% of all utterances contain only one word.

Next, two procedures were applied to Heleen's and Jessica's language data. First, the original procedure was used which excludes uninterpretable utterances, direct imitations, and yes/no-answers, songs and imitation games. Secondly, we applied the worst-case scenario to the same transcripts and gave all uninterpretable units (xxx strings) and fillers (&-strings) the value of one word.

7. Procedure

The speech samples were collected in the child's home, while child and parent engaged in normal daily routine. Subjects had access to their own toys and could move freely across the room. Sixty minute 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.

7.1. Illustration: a what-if or worst-case analysis of linguistic productivity, sentence length and MLU

7.1.1. Linguistic productivity

We begin with a very simple characteristic of the child's language, which is the child's linguistic productivity measured in number of sentences spoken during each observation session. This measure can

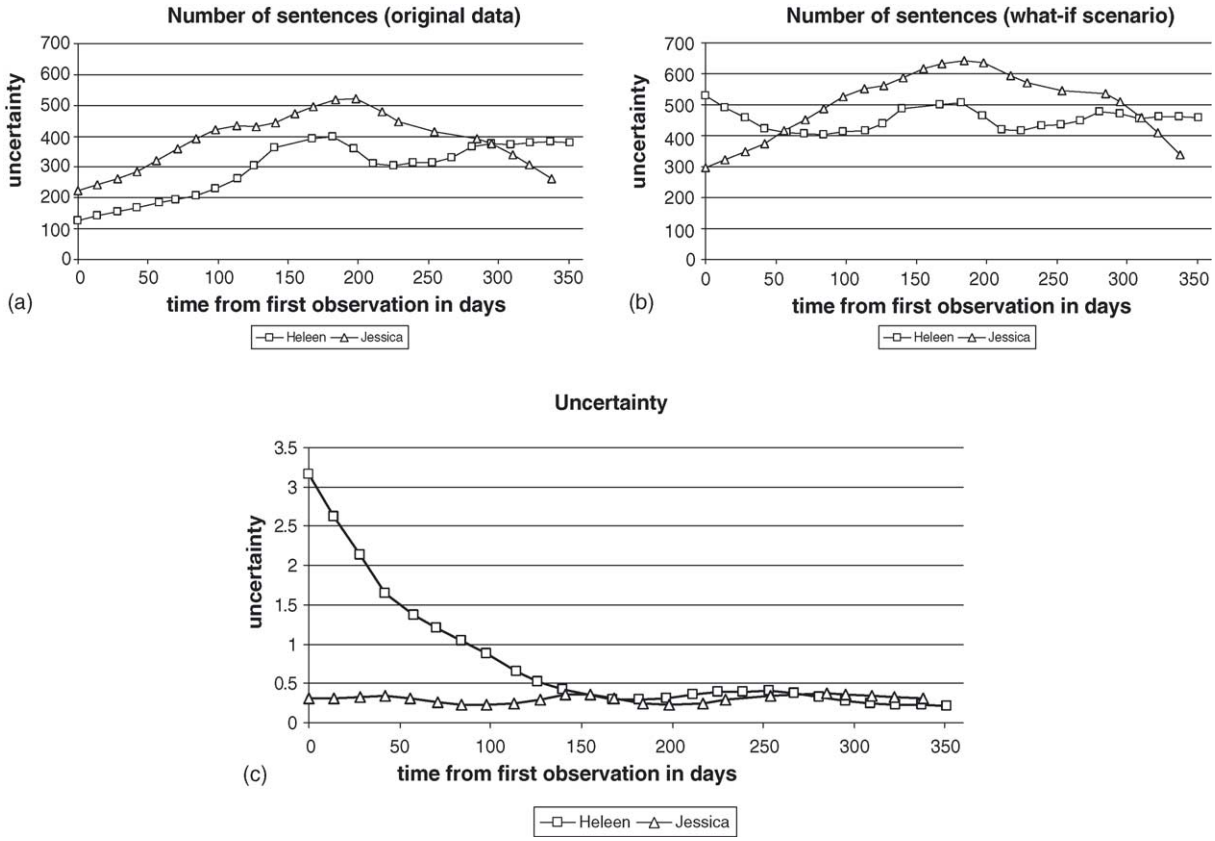

Fig. 1. Linguistic productivity and uncertainty based on original and worst-case counting scenarios of subjects Heleen and Jessica.

tell us something about the eventual growth of productivity, but also about differences between children (see Shore, 1995). In order to simplify comparison, the raw data have been smoothed by means of a so-called Loess smoothing procedure (Simonoff, 1996). Inspection of Fig. 1a that is based on the original data (excluding uncertain cases, fillers etc.) suggests that Heleen's productivity increases and then roughly stabilizes around day 180. Jessica on the other hand shows an inverted U-shape, with a peak around day 200. Fig. 1b is based on the "worst-case" scenario and assigns the value of 1-word to any ambiguous case, which changes the number of sentences relative to the effect of this assignment (e.g. it might affect the number of 1-word sentences more than the number of 3-word sentences). Inspection of the figure still reveals an inverted U-shaped pattern for Jessica, but a pattern of more or less stable productivity for Heleen, which is different from the original case. Note that the difference between the two figures is not a matter of which is "right" and which is "wrong". Both aim at providing a correct picture of the growth of productivity, but they do so under different assumptions (one under a "strict"—the best-case scenario—and the other under a "tolerant"—the worst-case scenario—interpretation of the number of words, and thus, of sentences). Let us call the difference between the original data ("best-case scenario") and the data under the what-if ("worst-case") scenario the margin of uncertainty (since it refers to uncertain

words, both in the sense of unintelligible and ambiguous). It is possible to quantify this uncertainty margin as the proportion of uncertain words over the certain ones. In this particular case, we are interested in the effect of this margin on the number of sentences. We thus calculate the uncertainty margin as the difference between the number of sentences under the what-if scenario and under the original, "strict" interpretation, divided by the number of sentences under the "strict" interpretation. Fig. 1c shows that in Jessica's case, the uncertainty hardly changes over the course of the observed period of about 1 year. Heleen, on the other hand, shows a marked drop in the uncertainty measure and stabilizes onto a level that is similar to Jessica's around day 140 (days are counted after the first observation date).

The difference provides a quantitative indication of a difference in style between the two children, with the observational data suggesting that Heleen is of a more "expressive", Jessica a more "referential" style (Nelson, 1975). Thus, the uncertainty or "interpretability" index potentially provides a quantitative measure of a difference in linguistic style between children. Second, the difference in the trajectory of this index shows that our original conjecture, that interpretability increases with age is not true in all cases, but is also likely to be an individual property.

7.1.2. Sentence length

Fig. 2 displays the resulting trajectories of utterance length of Heleen (top) and Jessica (bottom). Inspection of the figure shows that the data of Heleen and Jessica differ rather considerably. The differences between Heleen's original scores and her worst-case scenario are relatively small. There is only a slight widening or narrowing of the bandwidths, 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 what-if or worst-case procedures are not very different. Remind that a considerable difference was found between the two counting procedures if the absolute numbers of sentences were taken into account. The *proportions* of sentences however, are considerably less dependent on the counting procedure chosen. This finding can be explained by 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 1-word utterances (transformed dummy utterances). In the case of Heleen, the resulting proportions or ratios of 1- over 2- and 3-plus word utterances change only slightly. That is, there is a balance in the trade-off between the different utterance categories.

The data of Jessica, on the other hand, show much greater differences between the two procedures. The difference is greatest in the 1-word category. The figure shows that there are relatively more 1-word utterances and less 2-word utterances in the original sample than in the worst-case scenario. This means that the original values actually underestimate sentence length, as was expected in the introduction. Furthermore there seems to be a decrease in the differences between the sets, i.e. the lines resulting from the different procedures tend to grow closer together.

What differences do these counting scenarios make if we wish to describe the general trajectories of 1-, 2- and 3-plus word sentences? The general trajectories were inferred from the data by applying a smoothing procedure as introduced earlier.

Fig. 3 shows the smoothed trajectories under the two scenarios for Heleen and Jessica, respectively. As expected, the qualitative patterns are rather different in Jessica's case and highly similar with Heleen. In particular, the pattern of change in Jessica's 1- and 2-word sentences is quite different in the worst-case scenario, the latter showing that 2-word sentences are relatively stable throughout the entire observation period. The original counting scenario results in a pattern of 2-word sentences showing a relatively rapid increase in the beginning.

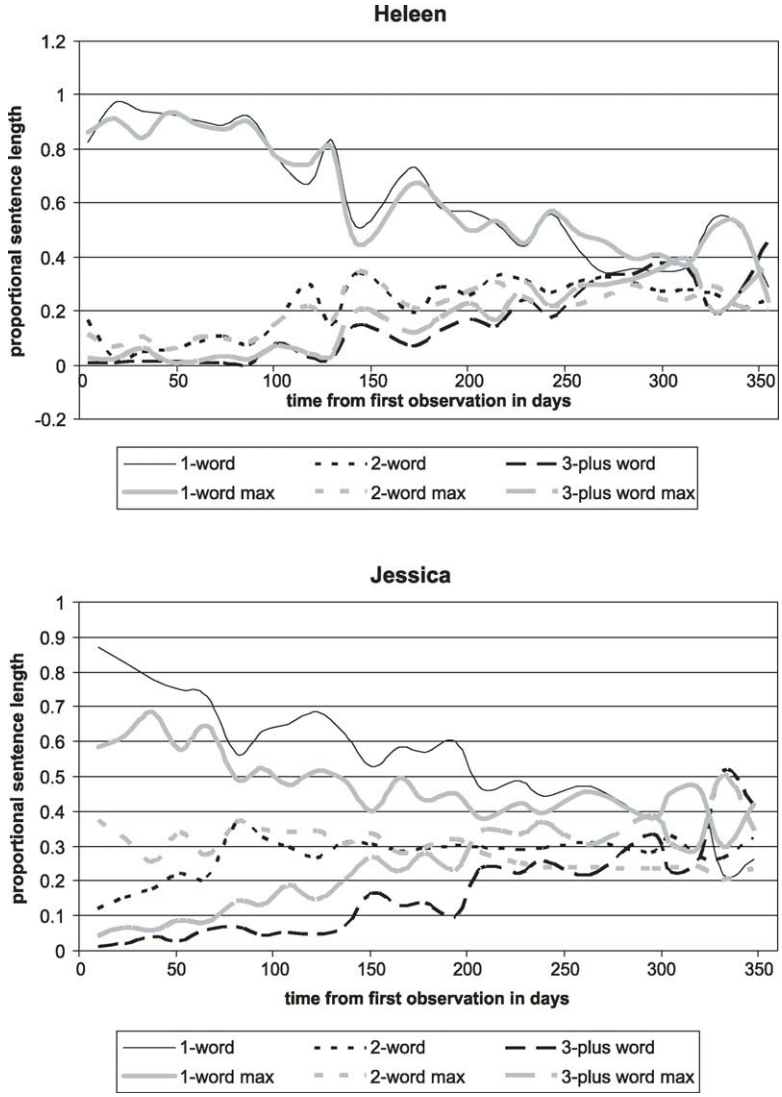

Fig. 2. 1-, 2- and 3-plus word utterances under the original and worst-case ("max") counting scenarios.

As discussed earlier, the difference between the original and worst-case scenarios can be quantified in the form of an uncertainty measure, which, for each observation session, was defined as the absolute difference between the worst-case number and the original (best-case) number, divided by the original number. However, since the current numbers are proportions and thus have a common scale (the unit interval between 0 and 1), division is no longer necessary and so eventual scaling problems resulting from heteroscedasticity can be avoided. Besides, if instead of taking the absolute difference, we subtract the original count from the worst-case count, uncertainty can be either negative (in which case the worst-case count is lower than the original count) or positive. The resulting uncertainty values can be smoothed in the same way as the data, resulting in a general trend over time. Fig. 4 shows the trends of uncertainty with regard to the proportions of sentences of different length for Jessica and Heleen, respectively.

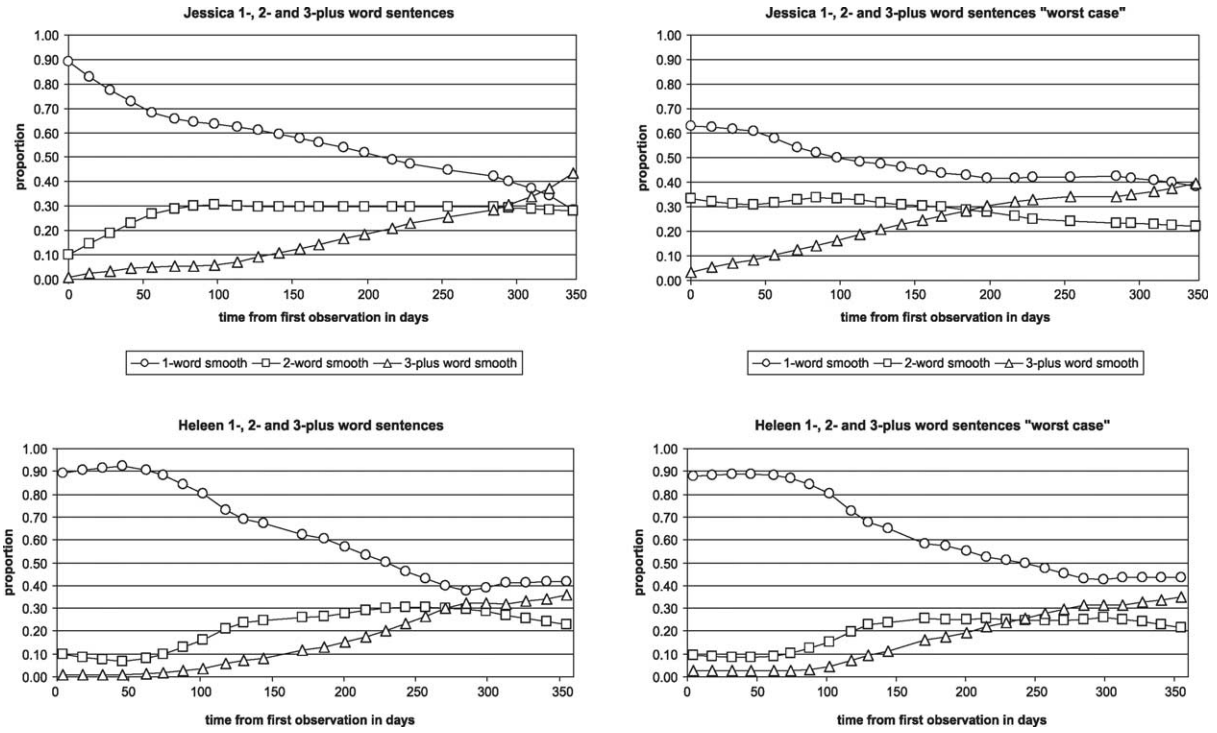

Fig. 3. Smoothed developmental trends for 1-, 2- and 3-plus word utterances for Jessica and Heleen under the original and the worst-case counting scenarios.

In spite of the fact that Heleen showed a major quantitative difference (in terms of total numbers of sentences) under the two counting scenarios (best-case and worst-case), the difference in terms of proportions of sentences is small. Jessica, on the other hand, who showed relatively little quantitative differences for the two counting scenarios, shows considerable differences for the proportions of sentences. This difference might suggest that, overall, Jessica's and Heleen's uninterpretable utterances are of a different kind. If the uninterpretable utterances are basically of the same classes or categories as the interpretable ones, adding the uninterpretable ones to the pool of utterances will not change the proportions between types of sentences (1-, 2- and 3-plus word). If the uninterpretable words are different, in the sense that they tend to be "fuzzy in-betweens", adding them is likely to change the proportions of sentence-types in the sample. Of course, this hypothesized categorical difference needs further testing. However, if this difference indeed turns out to correspond with different strategies in which children construct their language on the go, it illustrates how information from comparing the original with the worst-case counting scenario might be of help in acquiring a better understanding of the process of language development.

7.1.3. Mean length of utterance

Fig. 5 shows the results of both procedures to MLU-w. While for Heleen, the lines are very close together, the data of Jessica show more differences. This means that in the case of Heleen, the differences in the results of the original MLUs and the MLUs in the worst-case scenario are small. Note that the slight

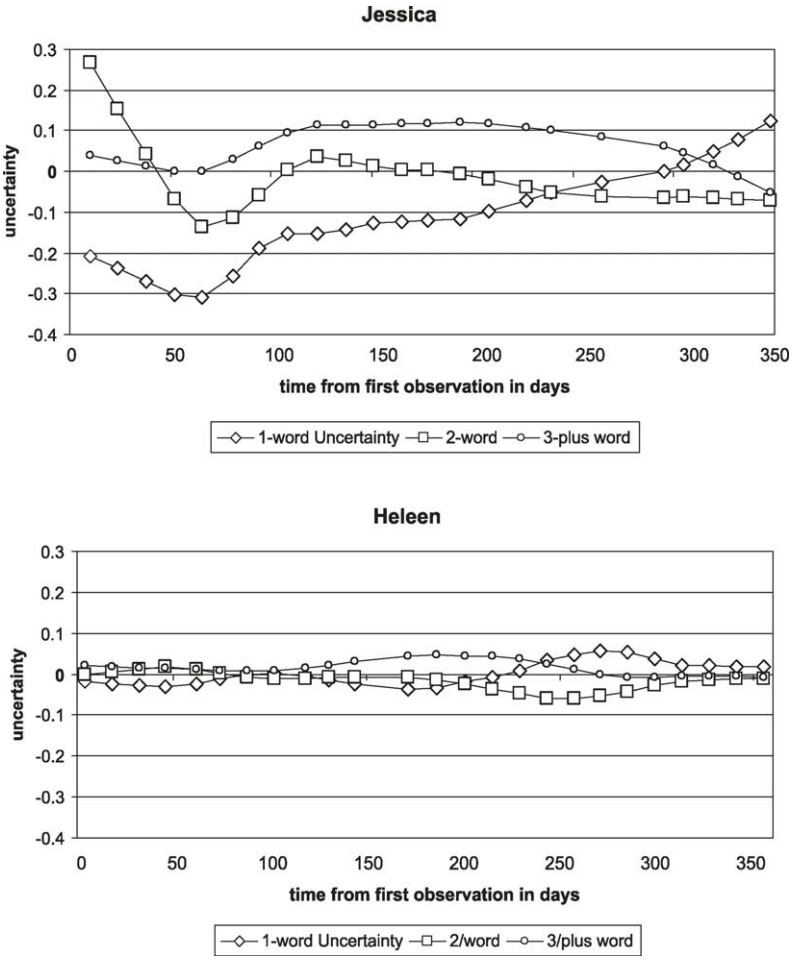

Fig. 4. Uncertainty of 1-, 2- and 3-plus word utterances for Jessica and Heleen.

difference between both lines can go in either direction: at some points MLU-original is slightly lower than MLU-max (based on the worst-case scenario), and at some points it is higher. Although it seems strange that MLU-original (which was assumed to underestimate utterance length) can be the higher of the two, this is the result of a trade-off between 0-, 1-, 2- 3-, and more-word utterances. In the worst-case scenario, there are a proportionally large number of original 0-word utterances (dummies) that have become 1-word utterances. 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 Fig. 5 shows that there is an almost perfect trade-off between the two calculations.

In the case of Jessica, on the other hand, the differences are much greater. The figure clearly shows that MLU based on the worst-case scenario is positioned around .20 above the original MLU almost to the full end of the trajectory. Based on our previous calculation of sentence length, this is in line with expectations. If the uncertainty according to the first uncertainty equation is calculated (which is: abs(max − min)/min),

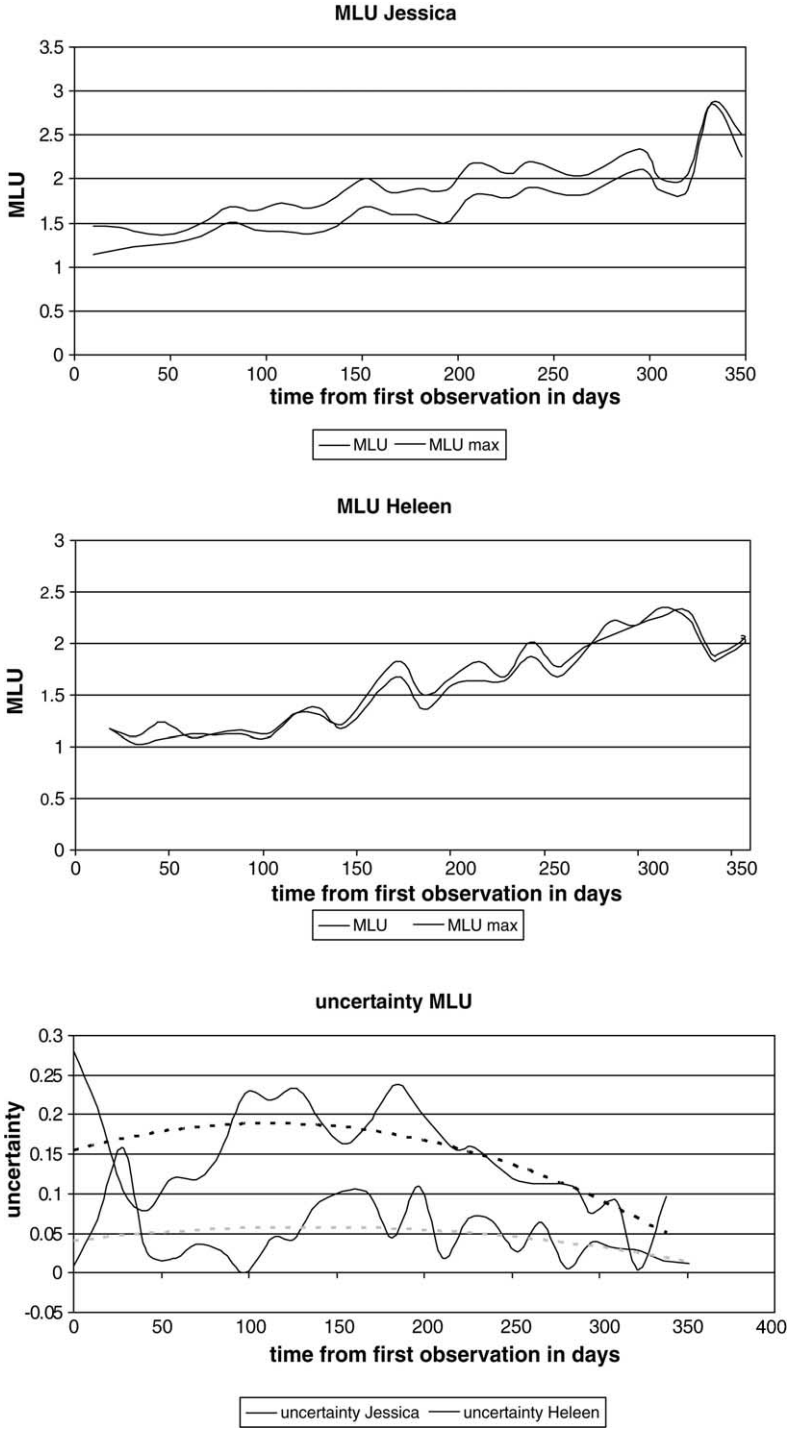

Fig. 5. MLU for Jessica and Heleen under the original and worst-case (“max”) counting scenarios; (bottom) comparison of uncertainty for Heleen and Jessica from first day observation on, with quadratic trendline.

the speech of both children becomes on average more intelligible with time, although Jessica's sample poses more interpretation problems than Heleen's (as far as MLU is concerned).

8. General conclusion: what-if and worst-case scenarios as a way to help solve the interpretability problem

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. Difficulty of interpretation might entail a form of misunderstanding (e.g. we cannot recognize the child's pronunciation of a particular word) or a form of fuzziness or ambiguity (e.g. a sound represents a word or a proto-word, e.g. a filler). Choices must be made about how uncertainties about the categorization of behaviors should 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 particular based on worst-case scenarios. In our example from child language, the scenarios were defined as those in which either *none* or *all* of the uninterpretable utterances referred to a word. We compared the curves resulting from these extreme cases with each other and found that the results of two subjects were very different. This finding supports our expectations that there are individual differences in the effect of unintelligible speech (including ambiguous words or filler syllables) on utterance length measures. It also implies that we must be careful with interpretations regarding "errors" or missing data: unintelligible speech is not distributed equally across utterances lengths, it is not distributed equally across subjects and the effects of unintelligible speech are also not equally distributed across quantitative measures (simple counts) and qualitative measures (proportions of categories). We acknowledge that the procedure should probably be refined in order to win credibility, since not al the elements to which we assigned a fixed value can be assumed meaningful. For instance, with the use of postcodes in CHILDES 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.

The discussion will be concluded 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. One suggestion on how to develop such procedures relates to an issue that the current authors have discussed in the context of observer (dis)agreement in the categorization of linguistic forms. Observation-based studies, for instance of motor behavior or emotional expression, involve the classification of behavior in categories. Given the fuzziness and continuous character of much behavior, attempts at categorization leads to disagreement among observers. We argued (Van Geert & van Dijk, 2003) that if the observers are competent and well-trained, such disagreement contains information about the categorized behaviors. For instance, it

eventually refers to the fact that the categories are still immature, or under (developmental) construction. Hence, in observational studies of developing behavior, one might make a distinction between frequency counts based on "strict" interpretations where both observers agree, and "worst-case" scenarios that take the (divergent) interpretations of the observers for granted (in which case one would have one worst-case scenario per observer).

Second, having read our plea for different scenarios, the reader may ask: which, if any, of the two scenarios provides the "true" picture, the true number of words or sentences, for instance? If we know that neither of the scenarios provides the true picture, why not invest our energy in trying to find the scenario that gives the correct number of sentences, for instance? Our answer is that it is a matter of principle that no such "true" representation exists, at least not in the unconditional sense. In a developing system, information about existing categories can be unclear (e.g. children lacking pronunciation skills to allow for clear determination of the word pronounced) or the categories themselves can still be under construction and therefore ambiguous or fuzzy (see Van Geert & van Dijk, 2003). Fuzziness and lack of information are part of the complexity of the system. There is probably no other way to represent a complex system than to represent it from different "perspectives". The "strict" and "worst-case" scenarios represent such different perspectives, and it is by comparing information from these two sources that a better specification of the phenomenon at issue can be provided. A comparable approach, which however uses an entirely different type of perspective, is Fischer's distinction between optimal and functional levels of development (Fischer & Bidell, 1998). Instead of searching for a single, "true" representation of a person's developmental level, the best representation one can give is by showing both the functional level (based on independent performance) and the optimal level (based on help and assistance from a more competent person). The situation is comparable to providing a photograph of a three-dimensional object, such as a house: no single photograph is the only true one, it is by combining photos, that is to say, different perspectives, that one arrives at a better representation of the house.

Finally, what if the application of worst-case scenarios leads to the conclusion that the resulting growth curves—or anything else one whishes to extract from the data—are highly sensitive to the decisions made to solve the ambiguities and lack of interpretability in the data? That is, what if the perspectives provided by two different "photographs" are highly dissimilar? If results depend very much on the way in which the interpretability problems have been solved, the researcher can indeed be faced with a problem of reliability, but possibly also with a problem of validity. It is possible that the categories chosen for the interpretation of the behaviors are just not adequate and that better ones must be provided. This means that if child speech is truly—intrinsically—ambiguous or fuzzy, the entire concept of a "true score" of utterance length or MLU should have to be reconsidered. However, if different procedures produce highly different results, the problem is not necessarily one of less adequate methods or variables. It is likely that it hints at an interesting aspect of the data themselves that would have remained concealed if we had not bothered to look at it from different perspectives. The complexity and richness of developmental processes requires that all possible sources of information are used, including the information contained in the observers' uncertainties and ambiguities.

References

Afifi, A. A., & Elashoff, R. M. (1966). Missing observations in multivariate statistics I: review of the literature. *Journal of the American Statistics Associations*, *61*, 595–604.

Arlman-Rupp, A. J. L., Van Niekerk de Haan, D., & Van de Sandt-Koenderman, M. (1976). Brown's early stages: some evidence from Dutch. *Journal of Child Language, 3*, 267–274.

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar: individual differences and dissociable mechanisms*. Cambridge: Cambridge University Press.

Braine, M.D.S. (1963/1973). The ontology of English phrase structure: the first phase. *Language, 39*, 1–13 [Reprinted in: Furguson, C.A., Slobin, D.I., (Eds.). *Studies of Child Language Development*. New York Holt, Rinehart & Winston].

Brown, R. (1973). *A first language: The early stages*. London: Allen & Unwin.

Conti-Ramsden, G., & Jones, M. (1997). Verb use in specific language impairment. *Journal of Speech and Hearing Research, 40*(6), 1298–1313.

Dabrowska, E. (2001). Discriminating between constructivist and nativist positions: fillers as evidence for generalization. *Journal of Child Language, 28*(1), 243–245.

De Weerth, C., Van Geert, P., & Hoitink, H. (1999). Intraindividual variability in infant behavior. *Developmental Psychology, 35*(4), 1102–1112.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistics Society, B39*, 1–38.

Dolitsky, M. (1983). The birth of grammatical morphemes. *Journal of Psycholinguistic Research, 12*, 260–353.

Dunn, M. (1996). The use of spontaneous language measures as criteria for identifying children with specific language impairment: an attempt to reconcile clinical and research incongruence. *Journal of Speech and Hearing research, 39*(3), 643–654.

Fischer, K. W., & Bidell, T. R. (1998). Dynamic development of psychological structures in action and thought. In R. M. Lerner (Ed.), *Handbook of child psychology. Theoretical models of human development: 1* (5th ed., pp. 467–561). New York: Wiley.

Frijn, J., & De Haan, G. (1994). *Het taalerend kind [The language learning child]*. Dordrecht: ICG Publications.

Gillis, S., & Schaerlaekens, A. (Eds.). (2000). *Kindertaalverwerving, een handboek voor het Nederlands [Child Language Acquisition, a handbook for Dutch]*. Groningen Martinus Nijhoff Uitgevers.

Graham, J. W., Hofer, S. M. & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In L. M. Collins & L. A. Seitz (Eds.). *Advances in Data Analysis for Prevention Intervention Research*. NIDA Research Monograph. Series (#142), Washington, DC: National Institute on Drug Abuse.

Hansson, K. (1997). Patterns of verb-usage in Swedish children with SLI: an application of recent theories. *First Language, 17*(50), 195–217.

Hartley, H. O., & Hocking, R. R. (1971). The analysis of incomplete data. *Biometrics, 27*, 783–808.

Kilani-Schoch, M., & Dressler, W. U. (2000). Are fillers precursors of morphemes relevant for morphological theory? In W. U. Dressier, O. E. Pfeiffer, M. Pochtrager, & J. R. Rennison (Eds.), *Morphological analysis in comparison, current issues in linguistic theory: 201*. Amsterdam, Philadelphia: John Benjamins.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Little, R. J. A. (1982). Models for non-response in sample surveys. *Journal of the American Statistics Association, 77*, 237–250.

MacWhinney, B. (1991). *The CHILDES project, tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

MacWhinney, B. (2003). CHILDES Child Language Data exchange System. http://childes.psy.edu/manuals/CHAT.pdf.

McGregor, K., & Johnson, A. (1997). Trocaic template use in early words and phrases. *Journal of Speech and Hearing Research, 40*(6), 1220–1231.

Navarro Pastor, J. B. (2003). Methods for the analysis of explanatory Linear Regression Models with missing data not at random. *Quality and Quantity, 37*, 363–376.

Nelson, K. (1975). Individual differences in early semantic and syntactic development. In D. Aaronson & R. Rieber (Eds.), *Developmental Psycholinguistics and Communication Disorders* (pp. 132–139). New York: New York Academy of Science.

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1* (pp. 697–715).

Peters, A. M. (1990). *From phonology to morphology: the transformation of filler syllables. Fifth International Congress for the study of Child Language*. Hungary: Budapest.

Peters, A. M. (2001). Response to comments. *Journal of Child Language, 28*(1), 283–289.

Rescorla, L., Roberts, J., & Dahlsgaard, K. (1997). Late talkers at 2: outcome at age 3. *Journal of Speech and Hearing Research, 40*(3), 556–566.

Rosenthal-Rollins, P., Snow, C. E., & Willett, J. B. (1996). Predictors of MLU: semantic and morphological developments. *First Language, 16*(47), 243–259.

Scarpa, E. (1993). *Filler-sounds and the acquisition of prosody: sound and syntax. Sixth International Congress for the Study of Child Language*. Italy: Trieste.

Shaffer, D. (1989). *Developmental psychology, childhood and adolescence*. California: Brooks/Cole Publishing Company.

Shore, C. M. (1995). *Individual differences in language development*. Sage: Thousand Oaks.

Simmonsen, H. G. (1993). *Models for the description of Phonological acquisition. Sixth International Congress for the Study of Child Language*. Italy: Trieste.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer Verlag.

Thordardottir, E., & Weismer, S. E. (1998). Mean length of Utterance and other language sample measures in early Icelandic. *First Language*, *18*, 001–032.

Valian, V., Hoeffner, J., & Aubry, S. (1996). Young children's imitation of sentence subjects: evidence of processing limitations. *Developmental Psychology*, *32*(1), 153–164.

Van der Meulen, B. F., & Smrkovsky, M. (1983). *BOS 2-30. Bayley ontwikkelingsschalen, handleiding [Bayley Developmental Scales Manual]*. Lisse, the Netherlands: Swets & Zeitlinger.

Van Geert, P., & van Dijk, M. (2002). Focus on variability; new tools to study intra-individual variability in developmental data. *Infant Behavior and Development*, *25*(4), 340–374.

Van Geert, P., & van Dijk, M. (2003). Ambiguity in Child Language. The problem of inter-observer reliability in ambiguous observation data. *First Language*, *23*(3), 259–284.

Veneziano, E., & Sinclair, H. (2000). The changing status of "filler syllables" on the way to grammatical morphemes. *Journal of Child Language*, *27*(3), 461–500.

Veneziano, E., Sinclair, H., & Berthoud, I. (1990). From one word to two words: repetition patterns on the way to structured speech. *Journal of Child Language*, *17*, 633–650.

Watson, M., & Scukanec, G. (1997). Profiling the phonological abilities of 2-year-old; a longitudinal investigation. *Child Language Teaching Therapy*, *13*(1), 3–14.