Chapter 2
Child Language Cuts Capers

a study on intra-individual variability in language development


Abstract

Traditionally, empirical developmental psychology has emphasized the gradual character of developmental processes. Recently, dynamic systems theory (and related theories) caused an increase in the interest in dynamic properties of change (Thelen & Smith, 1994; van Geert, 1994). One of the aspects that is currently being reconsidered is intra-individual variability. Lack of stability and the presence of fluctuations in data have long been seen as the result of measurement error or environmental factors. However, dynamic systems theory views this type of variability as an important developmental characteristic.

This article uses the language data of two young children (age 2;4 and 2;8) to illustrate the problem of intra-individual variability, in the form of a study of short-term variability in Mean Length of Utterance as an important language measure (MLU, Brown, 1973). The data of both children showed considerable variability in MLU, even between measurement days.

In addition to illustrating intra-individual variability, this article deals with the meaning of that notion and also addresses the question of how “noise” can be distinguished from developmentally meaningful variability.
Chapter 2

2.1 Introduction

2.1.1 Stability and variability
When you open a randomly chosen journal article on child development, most of the articles will present the variables under investigation in the form of regular curves based on averages of groups of subjects. In addition, many studies are based on a restricted number of measurement points per individual. The fact that the resulting curves show regular patterns of development is not surprising, since averaging over groups has a smoothing effect on irregularities. Moreover, drawing a line between only a restricted number of points easily leads to a smooth developmental curve.

Time serial developmental research is still relatively rare, although the effort invested in such studies is increasing. This type of research shows a dramatically different picture of development, especially when studies are based on a larger number of measurement points. The lack of regularity in the developmental curves is often striking. A recent example is taken from the study of de Weerth, van Geert and Hoijtink (1999) on emotion related behaviors in the infant's first 15 months of life. The fluctuations and irregularities immediately catch the eye (see figure 2.1).

![Infant E](image)

Figure 2.1. Crying behavior (as a percentage of the observed time) of one infant (Infant E) (source de Weerth, 1998).

The study of Ruhland (1998) on transitions in early language development also shows a capricious developmental pattern: a general increase of function word use, with strong fluctuations from measurement to measurement (see figure 2.2).
The present article discusses fluctuations in early language development. We address the following questions: is variability (as found in recent time serial research) also present in early language development? How can this variability be explained? In answering these questions we will refer to our own study on Mean Length of Utterance (MLU), a widely used measure for grammatical complexity.

### 2.1.2 Variability in early language development: how stable is MLU?

In 1973, Brown proposed Mean Length of Utterance as the best estimation of grammatical complexity in early child language (Shaffer, 1989). He also proposed to calculate MLU by dividing the total number of morphemes by the total number of utterances in the first 100 utterances in a sample of spontaneous speech. In addition to these criteria, Brown constructed a classification of early language...
development into MLU-stages. These stages vary from MLU-1 (for MLUs between 1.0 and 2.0) in four equal steps (of 0.5 MLU point) to MLU-5 (between 3.5 and 4.0). Since Brown’s pioneering study, MLU and its MLU stages have been widely used, a success that is partly due to the simplicity of this concept. Rosenthal-Rollins, Snow and Willett (1996) state that 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 has been criticized as a measure for grammatical complexity. This criticism focuses especially on the fact that language complexity is mainly expressed in the form of structural elements. By simply counting morphemes, this structural complexity is ignored. Therefore, MLU is often considered to be an inadequate complexity measure (Frijn & de Haan, 1994). Bates, Bretherton and Snyder (1988), however, state that MLU is still an important measure of syntactic development, especially up to the third year of life. After the 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).

Until today, MLU is a widely used measure, both for children under the age of three and older children. 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 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 a solid diagnostic predictor for developmental disorders. In the literature, there are also numerous studies in which MLU is not the focal point, but where it is applied to describe the subjects (e.g. Watson & Scukanec, 1997; Rescorla, Roberts & Dahlggaard, 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).

Few studies pay explicit attention to intra-individual variability in MLU. So far, most studies measure MLU only once. If MLU turns out to be highly variable within individuals, these studies should be viewed in a different perspective. Only two studies pay attention to variability between days (Minifie, Darley & Sherman, 1963; Chabon, Edolf & Kent-Udolf, 1982). Minifie et al. investigated the consistency of language from day to day, by means of several language measures, among which average sentence length (similar to the later MLU by Brown). Variability (named “temporal reliability” in the article) was operationalized as test-retest reliability: the consistency of an individual, on the same test after repeated measurement. Two groups of subjects were tested, the first was aged around 5;6 years, the second aged around 8 years. A semi-structured test situation was used in which subjects were asked to tell a story with three pictures. The first 50 utterances of these stories were used for further analysis. This test was administered at three different occasions within three weeks. The results show that none of the language measures had a high temporal reliability. Minifie et al. (1963) conclude that none of the language measures, based on the 50 utterance samples, were consistent.

In a study on temporal reliability of language measures, Chabon et al. (1982) investigated MLU (as defined by Brown, 1973) as a language measure for children aged 3;6 to 9;6 years. They found a lack of temporal stability of individual MLU values measured on three distinct occasions. The authors
caution against the use of one single MLU assessment as a diagnostic indicator. A temporal decrease or increase could wrongly be ascribed to a (lack of) progress, while it is merely a reflection of the low temporal reliability of MLU. According to Chabon et al, it would be better to use averages over repeated measurements, for diagnostic purposes.

Both Minifie et al. (1963), and Chabon et al. (1982) do not elaborate on the source of the variability they have found in their studies. However, the use of the term “reliability” points in the direction of measurement error, since this term is associated with true score theory. This suspicion is also supported by the recommendation of Chabon et al. to work with averages. True score theory (e.g. Lord & Novick, 1968, Cronbach, 1960) -which is a deeply rooted axiom in psychology- states that every measurement is subject to error. Because this error is assumed to be randomly distributed around the true score, the true score can best be approached by working with averages. Therefore, we assume that the recommendation to work with averages points in the direction of true score theory.

2.1.3 Variability in other language measures
As we mentioned previously, there are only a few studies that focus on intra-individual variability in utterance length variables such as MLU. Two studies, however, have paid attention to variability in other language variables.

First, Wells (1985) investigated variability in language productivity within one day. This productivity was measured as the total number of utterances produced during a recording. The subjects (n=125) were recorded for a short duration, on average 18 times a day, distributed over equal intervals. Variability in productivity turned out to be significant: the children were more productive in the morning. Thus, language productivity is apparently variable within the course of a day.

In the second study that focuses on variability in language development, Ruhland (1998) analyzed transitions in the development of function words. Function words are more or less abstract words that have a more syntactical than referential meaning. They are also called “closed class words” because they belong to a limited set that no longer expands. Instances of function words are articles, modals, pronouns, and prepositions. It is assumed that the use of functions words is related to the syntactical abilities of young children. When these abilities increase, the number of function words in spontaneous speech also increases. Ruhland analyzed the time serial language data of six subjects from the Groningen Dutch Corpus (Childes database). These subjects were followed from the one-word stage to the stage in which the essential features of the Dutch syntactical system were acquired (ages from around 1;6 till around 3;0). Frequency counts of articles, modals and pronouns were made in two-weekly recordings of spontaneous speech. These specific function words are considered as an adequate indicator for the use of all types of function words. These, in turn, are seen as a measure for syntactic and semantic complexity of the child’s utterances (Ruhland, 1998). Figure 2.2 shows the developmental curves of two of the subjects in this study. Although the variability between subjects is large, the variability within subjects (fluctuations) is also striking. Instead of turning to measurement error, Ruhland considers this variability as meaningful for detecting developmental discontinuities. We will get back to this explanation at a later stage.
2.1.4 Explanatory models

The literature we have discussed so far suggests that early language development might be characterized by intra-individual variability. The question rises how this variability can be explained. First, there is the possibility that the fluctuations are caused by measurement error and contextual factors, as we have discussed previously. A child might be less fluent in his or her language because he or she is tired, while on another day the same child might speak fluently, due to a good night’s rest. The factor “fatigue” leads to fluctuations that distort the true language level. This classic way to approach variability stems from the previously mentioned “true score theory”. True score theory rests on the assumption that every psychological measurement is subject to “noise” (measurement error). Scores acquired with any type of measurement procedure are conceived as the true score plus or minus this “noise”. By definition, noise is always independent of the measurement itself, and thus, noise can be assumed to be normally, or at least symmetrically, distributed around a central tendency. This central tendency is consequently considered to be the best estimation of the errorless measure. In the linguistic sciences, a similar principle is used. Here, the question is whether on the basis of some language output, it can be derived if a child has acquired a specific linguistic rule or not. This problem is solved by the competence-performance distinction (see for instance Chomsky, 1986) in which competence is considered as the true underlying acquisition of a specific linguistic rule. Performance, on the other hand, is the sum of competence and a number of interfering factors. Both models of explanation seem to rest on the same general idea. Apparently, thinking about variability in terms of noise and interference is deeply rooted in both psychology and linguistics.

As we mentioned before, there is an increased interest in variability in the last decade. This impulse is caused by a radically new perspective in developmental psychology provided by dynamic systems theory. Dynamic systems models, which were initially applied to other branches of science such as physics, are currently being applied to human development with great success. Simple iterative growth functions turn out to be capable of simulating and modeling development. The two central concepts in this approach are: 1) the concept of self-organization and 2) the concept of an attractor-value. Simple (mathematical) dynamic systems models have shown that chaos, seemingly random variation within a certain range, can be the product of the developmental process itself and not something that is added to this process externally.

An second alternative approach to variability is posed by catastrophe theory (Thom, 1975; van der Maas & Hopkins, 1998), which is closely related to dynamic systems theory. Catastrophe theory focuses on discontinuities in processes (including developmental processes). One of the questions catastrophe theory raises is whether a given developmental curve is continuous or discontinuous. Catastrophe theory offers a set of criteria, called “catastrophe flags”, which can be applied to empirical data to detect discontinuities. In addition to the flags “sudden jump” and “bimodality”, “anomalous variance” is mentioned as an indication for a discontinuous transition. The criterion of anomalous variance refers to the expectation that variability increases in the vicinity of a developmental jump.

---

2 This assumption of normality is reasonable if noise is an additive process of independent factors which is in fact the simplest model of noise.
because of the loss of stability in the equilibrium. It takes some time before the next equilibrium is reached and before the fluctuations stabilize again. It reaches beyond the scope of this article to elaborate on the theoretical background of catastrophe theory. The central issue is that this theory does not automatically consider variability to be the result of measurement error. Instead it conceives of it as additional information, in the sense that its presence (at a specific moment in time) indicates a discontinuous transition.

Inspired by these new theories, several empirical studies have been conducted. The domains of research vary widely. Most of the research on variability is conducted in the field of motor development (see Smith & Thelen, 1993; Thelen & Smith, 1994). For instance, Wimmers (1996) studied transitions in the development of reaching and grasping and used the catastrophe flags to detect discontinuities. Also, Berthenthal (1999) discusses the function of qualitative variability in the development of crawling in infants. He states: “[..] this variability is not merely a correlate of change but instead a contributor to the change itself” (pp. 105). (also see Bertenthal & Clifton, 1998; Newell & Corcos, 1993). He goes on stating that variability offers flexibility, which drives development following Darwinian principles. Principles of variation and selection make that the most successful behaviors will be repeated and stored more frequently than the less successful ones.

In the field of emotion-related behaviors in infants, the study of de Weerth and van Geert (1999; de Weerth, van Geert & Hoijtink, 1999) also focuses on variability. In this study, four infants were followed on a two-weekly basis from age 0 to 15 months. Large variability in the emotion-relates behaviors was found in the period from 0 to 5 months, and 5 to 10 months, but not in the period from 10 to 15 months. Instead of discarding these fluctuations as measurement error, de Weerth and van Geert point at a possible adaptive strategy. Intra-individual variability, they claim, ensures the infant of continuing maternal care: “[..]mother and infant try out new ways of communicating with each other, and also change them over time. [..] [They] tune into each other and influence each other with their moods attitudes and developing skills etc.” (pp. 11). De Weerth and van Geert also point at the theoretical arguments of Goldsmith (1993), who advocates a dynamic systems approach to the development of emotions. Goldsmith considers emotions to be the result of a self-organizing process, which results from the dynamic interaction between many different elements, which include cognitive, motor and social components.

In the field of cognitive development, Alibali (1999) applies the framework of dynamic systems theory in a study of the acquisition of strategies children use to solve mathematical problems. She states that conceptual change is a cyclical process, which results in periods with more variability alternated with periods with less variability. In this case, variability is a predictor for the type of strategy change the child is going through. The expectation is that children with low initial variability will generate many new strategies, which causes the variability to increase. At the same time, children with high initial variability are expected to drop strategies, which cause the variability to decrease. The effect of variability was different for each condition of instruction type. Both initial variability and the instruction type contributed to the mathematical strategies the children generated.
In the field of language development, ideas from dynamic systems theory are applied on a small scale. For instance, the study of van Geert (1991) showed that the data from the lexicon growth of Keren (as published in Dromi, 1986) could be modeled with logistic growth equations. These growth curves show a smooth development, but this is largely caused by the fact that cumulative numbers were used. The analysis of the growth rate in the same data (as carried out by van Geert), shows large fluctuations (van Geert, 1994). This illustrates that the way variability is revealed is critically dependent of the way the developmental processes are analyzed.

As shown in the illustration, the data of Ruhland (1998) were highly variable. In his study he uses dynamic systems theory and catastrophe theory to explain these fluctuations. The catastrophe flag “anomalous variance” was applied to the variability of the use of pronouns (one of the function words) by comparing the first half hour with the second half hour of the sessions. In two out of six children, within-session variability of the use of pronouns (one of the function words) coincided with the jump-wise development, while in four children no such temporal relation was found. However, visual inspection of the developmental trajectories of all subjects showed large fluctuations between measurements. Although within-session variability did not correlate systematically with developmental transitions, between-session variability might still show such a pattern after further analysis. Ruhland speculates that between-session variability can be considered as a developmental characteristic and that knowledge about this phenomenon is important for a more thorough insight in developmental processes.

In summary, there are several alternative explanations for the existence of variability, besides the traditional measurement error assumption. Studies inspired by dynamic systems models and Catastrophe theory point at two different options: (1) variability as an indicator of a specific moment in development (Ruhland, 1998; Wimmers, 1996), and (2) variability as a causal factor for development (Bertenthal, 1999; de Weerth & van Geert, 1999).

It should be noted that dynamic systems theory does not claim that measurement error does not exist. The definition of measurement error is, however, more specific. For example, an error of measurement in language research might be a transcription error. For instance, the child says “My Pant-aehr dirty”, which the transcriber interprets as “My pants are dirty” (a four-word sentence). It might very well been the case that the child talks about his panther-toy: “My panther dirty” (a three word sentence). Frequent occurrences of this type of transcription error might result in an MLU that is either too low or too high. In that case, MLU is susceptible to measurement error. However, when a child has a high MLU in certain circumstances (for instance when it is experimenting with language) and a lower MLU in other circumstances (for instance when the attention of the child is drawn to something else or a different developmental domain), this is not conceived of as measurement error. Both values, in this case, stem from the abilities of the child in the dynamic interaction with the environment. The fact that the child’s MLU is higher under optimal circumstances is a reflection of the child’s potential and expresses the fact that the child is still in the acquisition process. We expect that when the abilities are fully acquired, achievement is less variable in different circumstances.
The question remains how these two types of variability (measurement error and intrinsic variability) can be distinguished in developmental data. How can we decide whether a variable curve is the result of measurement error, or a reflection of true, meaningful variability? We will elaborate on this question on the basis of the language data of two toddlers.

### 2.2 Methods

#### 2.2.1 Subjects and procedure

Language data were collected from two monolingual Dutch children. These subjects are named Jan and Eva, and are aged 2;8 and 2;4 years respectively. Speech samples of these toddlers showed that both of them were in the so-called “differentiation stage” of language acquisition (for characteristics of the Dutch differentiation stage see Frijn and de Haan, 1994). This stage is characterized by the elaboration and sophistication of the language aspects. For instance the child learns to form plurals, and to inflect nouns and verbs. Also the child begins to use function words such as articles, modals and pronouns.

Recordings of spontaneous speech were made at the child’s home by the second author. During these recordings the child was free to play and interact with at least one parent and the observer. Language samples consisted of at least 100 utterances each, a standard that is commonly used in language research (see for instance Lahey, Liebergott, Chesnick, Menyuk & Adams, 1992). The recordings were made within a period of three weeks. The dataset of Jan consisted of 12 recordings, the set of Eva of 14 recordings. The time between consecutive observations ranged between one and four days. Shortly after recording, the audio-tapes were transcribed according to Childes-conventions (MacWhinney, 1991). Of each recording, more than 100 child utterances (ranging from 115 to 135) were transcribed (by the second author). The child utterances were interpreted by the transcriber (also using context for a richer interpretation) and transcribed in standard (adult) language. If the child pronunciation deviated from the adult form, the child pronunciation was coded between brackets. In order to explore inter-transcriber reliability a second transcriber (the third author) listened to the tapes and made notes where his interpretation of the language differed from the transcriptions. On the basis of these notes, the percentage of inter-transcriber agreement was calculated as overlap percentages for Jan and Eva individually. This percentage was 96 in both children, which is high.

Mean Length of Utterance was computed based on these transcripts, using a definition closely related to Brown’s (1973). MLU was calculated in morphemes, excluding yes/no answers to closed questions. Imitations, on the other hand, were included because we consider them to be an essential element in child speech.
2.3 Results

2.3.1 Visual inspections of the developmental curves

Figure 2.3 shows the MLU values of both children against time. It can be seen that Eva’s MLU is generally slightly higher than Jan’s. On day 4 this difference is over 1 MLU point (Eva has an MLU of 3.47, Jan of 2.06). However, on day 10 Jan’s MLU exceeds Eva’s (Jan 2.91, Eva 2.28). Next to these obvious inter-individual differences, the intra-individual differences are also clearly observable. Jan’s scores range from 2.03 to 2.97, Eva’s from 2.39 to 3.47. The first impression is that Eva’s scores are more variable than Jan’s.

![Figure 2.3. MLU for Jan and Eva.](image)

2.3.2 Boxplot

A simple way to depict the spread of scores, and to compare the score ranges of the two children, is to transform the scores into the format of a boxplot. This figure displays information about the median and range. In a boxplot, 50 percent of all observed values are depicted as a box. The upper boundary of this box depicts the boundary of the third quartile, while the lower boundary of the box depicts the end of the first quartile. The whiskers above and below the box reach to 1.5 times the boxlength, in observed values. Values that exceed this boundary are called outliers (up to 3 times the boxlength) or extreme values (over 3 times the boxlength).
Figure 2.4 shows the boxplots for both children’s MLUs. Visual inspection of these boxplots reveals that roughly the top 50 percent of Jan corresponds with the low 50 percent of Eva, which is quite large. However, the inter-individual difference is large, both regarding the level and the shape of the curve. In addition to the fact that a great part of Jan’s values are below Eva’s, the boxplot is qualitatively different. While the distribution of Eva’s MLU values looks symmetric, Jan’s MLU distribution seems positively skewed, which means that the values above the median show a greater spread than the values below the median.

When comparing the intra-individual variability in the shape of the score distribution, the following things are noticeable. Half of Eva’s MLUs are between 2.7 and 3.2, while half of Jan’s MLUs fall between 2.3 and 2.6. Thus, for both children, the highest value is around one-and-a-half times as large as the lowest value (Jan 2.03-2.97, Eva 2.39-3.47). This means that measuring MLU once can yield large deviations from measurements taken at other (nearby) occasions depending on the timing of the measurement. This affects the predictive value of MLU. If we apply Brown’s MLU stages to these data we see that both children fall within different stages depending on the day of measurement. In this short period of a few weeks, Jan falls within stage MLU-2 (with his lowest value of 2.03) and MLU-3 (with his highest value of 2.97 which is almost MLU-4). Eva, on the other hand, alternates between MLU-2 (with her lowest value of 2.48), MLU-3, and MLU-4 (with her highest value of 3.47 almost MLU-5). Considering Brown’s initial intentions with the MLU stages, which were to give a global indication of the syntactic level of the child, this range of 2 to 3 stages is large.
2.3.3 Meaningful variability

So far, we have established that the development of MLU shows large intra-individual short-term fluctuations. This leads us to the question how this should be interpreted. Is this variability the result of measurement error (for instance transcription errors), or can this be the meaningful developmental variability we discussed above?

First, it should be emphasized that the variability found in these case studies is large. Purely intuitively, it seems unlikely that the fact that both children fall within several MLU-stages in the period of only a few weeks, is caused by for instance transcription errors. In these case studies, the inter-observer reliability was more than adequate. Also, the observations were very similar in the sense that each time the children played freely in their own homes. On each occasion the same persons were present and the children had the opportunity to play with their own toys. Thus, it seems unlikely that the large amount of the variability as found in the study is caused by pure “measurement” error. Other factors such as fatigue, the occurrence of unusual situations and so forth may very well have contributed to the large variability that was found. It is questionable whether this type of situational influences should be considered as measurement error. In fact, dynamic systems theory considers the sensitivity of the developmental grower (in this case MLU) to situational factors an important developmental characteristic. This sensitivity can increase and decrease during the process of development. For instance, it is assumed that perturbations have a larger effect when they occur near a developmental transition. This assumption has been empirically tested and confirmed in the domain of motor development. Thelen and Ullrich (1991) showed that the effect of perturbations decreased with the age of the infants, in a study on the development of infant stepping on a treadmill. As the infants, grew older, they became less sensitive and these perturbations resulted in less variability. For this reason, we argue that the effect of the environment should not automatically be associated with error, but can also provide information about the process of development.

Besides the intuitive idea that the variability found in the study cannot sufficiently be explained by measurement error, there are other starting-points to study the measurement error hypothesis more thoroughly. The first one is to analyze the distribution of the variability. Using the simplest model of noise as an additive process of independent factors, we may assume that there is a central tendency (true score) with a normally distributed spread (noise) around that central tendency. By studying the distribution of empirically found variability, we can analyze whether the variability in the data is in fact normally distributed or not. If the measurement error hypothesis applies, a bell shaped curve of variability is expected to be present in the data. We can study whether this is the case. Is the majority of fluctuations small and are larger peaks and wells rare? Is the spread of scores symmetrical? To answer these questions, we will return to the skewness of the fluctuations in the case study of Jan. While Eva’s distribution is largely symmetric, Jan’s distribution turned out to be positively skewed. This means that the outliers above the trendline or central tendency are more extreme than the outliers below the trendline. We argue that this positive skewness indicates that the variability in Jan’s case is potentially meaningful, in the sense that it may reveal the degree of consolidation of a new language
rule. When a new language rule is only just appearing and the child uses it in outbursts, we expect more extreme fluctuations upwards, which results in a positive skewness. On the other hand, if a language rule is very well consolidated we expect the child to use this new rule predominantly, and only use the old strategy infrequently. This old strategy will extinguish slowly in the output (the child utterances), which results in a negative skewness. A positive skewness, as can be seen for Jan’s MLU, may therefore indicate the recent acquisition of one or more linguistic rules, which cause the more extreme positive outliers.

A second way of studying the distribution of fluctuations is by analyzing the different shapes variability may take in the course of development. If the variability is solely caused by measurement error, we expect it to be randomly distributed, and thus be approximately equal in the entire developmental trajectory. If there is an increasing trend, we expect variability to increase proportionally. A sudden increase of variability on the other hand, cannot be explained by the measurement error hypothesis. In order to analyze these developmental effects, the data should have a scope that is wide enough to capture —at least part of — a developmental process. The case studies of Jan and Eva are too restricted in that sense.

To illustrate the approach we suggested above, we will apply it to the preliminary results of a more extensive study (van Geert & van Dijk, 2002, chapter 3 of this thesis). This concerns the data of Heleen, a monolingual Dutch girl who’s language development was followed from age 1;6 to age 2;6. During this period language is developing rapidly. In the beginning the child uses only one-word utterances, while at the end of this period the child predominantly uses more-word utterances that show early characteristics of the differentiation stage. The methods in this study are comparable to the methods of Jan and Eva, with the exception that larger speech samples were used. The speech samples of Heleen consisted of a full hour, (often more than 200 utterances) while the samples of Jan and Eva consisted of a little over 100 utterances. Furthermore, the measurement design in the Heleen study was irregular: two-weekly measurements were alternated with periods of more intensive measurements (for instance every other day). Heleen’s MLUs during this period were also computed, but for practical reasons we used MLU in words (MLU-w). This measure is quite similar to MLU in morphemes (MLU-m, which was used for Jan and Eva) but is easier to compute. It should be noted that it seems unlikely that the slight difference between MLU-w and MLU-m influences the results on variability. Figure 2.5 depicts the development of Heleen’s MLU, which also shows large variability.

---

3 Naturally, in order to conclude that this is in fact to case, we need longitudinal data of more subjects to corroborate this relation.
Figure 2.5. Heleen’s MLU in words, age 1;6 to 2;6, including a polynomial trendline (degree 2). The triangles indicate the intensive periods.

In order to estimate the range of scores, we determined a bandwidth of scores in Figure 2.6. This was done by using “moving windows” that compute the minimum and maximum within a selected timeframe (for instance 5 measurements). After that the timeframe moves up one position (and thus retaining a relatively large overlap with the previous timeframe) and computes the minimum and maximum again, etcetera. By depicting these minima and maxima in a graph, we can get an indication of how the bandwidth changes in the course of development.

Figure 2.6. Moving minimum, maximum and average of Heleen’s MLU (time frame 5 measurements).
Studying this bandwidth leads to the following conclusion. First, the range shows an increasing trend. Both the central tendency and the minima/maxima increase in the course of the observed year. Secondly, this bandwidth also seems to increase. In the earlier observations, the differences between the minima and the maxima are smaller than at the end of the trajectory. However, this is not surprising since a larger mean is often associated with a larger spread.

The variation in the bandwidth can be more irregular, which is illustrated with the results from spatial prepositions, a second language variable that was analyzed in the study of Heleen. The use of spatial prepositions was studied as follows: in each speech sample we counted the total number of spatial prepositions the child used. When the same spatial preposition (context) was uttered repeatedly, it was counted as only one spatial preposition. It is beyond the scope of this article to elaborate on the linguistic meaning of this category, we merely want to use the data to illustrate to developmental pattern of its variability which is shown in figure 2.7.

![Frequency of spatial prepositions in Heleen's language.](image)

Visual inspection of the curve leads to the observation that the developmental trajectory consists of two distinct parts. In the first part prepositions rarely occur, with small fluctuations. The second part on the other hand shows a tremendous increase in preposition use in combination with large fluctuations (from around November onwards). Figure 2.8 shows the moving minima and maxima of Heleen’s preposition use.

It is now possible to inspect the bandwidth between the moving minima-maxima graph for further widening or narrowing. In the case of Heleen’s preposition use, the impression that the bandwidth consists of two distinct stages is visually confirmed. There seems to be a widening from around November. While the minimum remains constant around the entire trajectory, the maximum suddenly
increases from this point onwards. Finally, at the end of the curve, we see a slight narrowing of the curve.

Figure 2.8. Moving minima and maxima for Heleen’s prepositions (age 1;6 tot 2;6) (using moving windows of 5 positions).

Thus, Heleen’s preposition use shows a bandwidth in which a sudden widening occurs, followed again by a slight narrowing. However, according to the measurement error hypothesis, the proportion between the central tendency and the range should be equal during the entire path of development (because error is assumed to be independent of the specific moment in development). Measurement error alone cannot satisfactorily explain these observations.

It should be noted that the methods discussed above are exclusively based on visual inspection of developmental curves. In recent years, several new statistical techniques have been developed that make testing of these observation possible, for instance resampling procedures (Efron & Tibshirani, 1993; Good, 1999) and the critical moment method (Verheul & Geuze, 1999). It is beyond the scope of this article to elaborate on these techniques. It should be noted, however, that testing is possible. The observations from visual inspection are important for generating hypotheses that can be tested later on.

2.4 Conclusion and discussion

2.4.1 Variable variables

In the case studies of Jan and Eva, the fluctuations in MLU are salient. In this period of a few weeks both children show a range that is large enough to place them in more than one MLU-stage (Jan in two stages and Eva in three). In the literature a single measurement is standard for the estimation of
MLU as a global syntactical indicator of the syntactic development of a subject. On the basis of this single computation, predictions are made about the development of other variables or the development at a different point in time. For this reason, we can conclude that the variability found in this study is large. Dynamic systems theory shows that there is an alternative to the measurement error hypothesis approach to variability. In developmental psychology, this approach to variability is influencing an increasing number of researchers.

The question remains however: why does MLU show such large fluctuations? The following reasoning might provide an explanation. At the moment of the study, the two children were in a stage in which language is developing at great pace. In this period, the so-called differentiation stage the child’s language is greatly accelerated (for characteristics of the Dutch differentiation stage see Frijn & de Haan, 1994; Goorhuis & Schaerlaekens, 1994). There is a large growth in the semantic and morphological abilities, and sentences become increasingly longer and more syntactically complex. Utterances grow from being linearly ordered to being hierarchically ordered. Also, children can be very resourceful in applying language, and invent new combinations constantly. We have established that the language of Jan and Eva is rapidly developing at the time of the investigation. If Mean Length of Utterance is an adequate indicator of syntactic development (see Bates, Bretherton & Snyder, 1988), we expect that this measure is affected by this rapid development.

It is assumed that MLU in young children is strongly related to the acquisition of many linguistic rules. For instance, MLU is known to increase when children learn to order their utterances hierarchically. However, it might be that the acquisition of linguistic rules does not guarantee that the child applies this new rule in each utterance. Sometimes, the child successfully applies the rule, while at other times it falls back on the previous, less advanced expressions. This idea is compatible with the concept of partial knowledge from the theory of Connectionism (see for instance Elman, Bates, Johnson & Karmiloff-Smith, 1996). This theoretical approach is based on the notion that development is driven by the interaction between nodes in a neural network. In the case of partial knowledge, both the old rule and the new rule are simultaneously present in the system because of the dynamical interaction between the nodes. If this new linguistic rule is not consolidated yet, the output is variable. Qualitative changes can thus lead to quantitative variability. This variable output occurs most likely in situations with rapid development, that leads to the expectation that developmental processes which occur in a short period of time are more susceptible to large fluctuations (see for instance Alibali & Goldin-Meadow, 1994). When children are in the differentiation stage, they acquire new linguistic rules, which lead to fluctuations. Only when syntactical development is no longer expressed in an increasing MLU (around the third year of life, see Rosenthal-Rollins et al., 1996), this measure is expected to stabilize.

An explanation that sees variability as something that is intrinsic to development is in fact related to the way Brown (1973) considered short-term variability. In addition to MLU, he defined a second measure for syntactical development: the Upper Bound (UB). This is the largest number of morphemes in the longest utterance in the language sample. In the literature, this measure is applied much less, but was initially constructed to form an index of syntactical development, together with MLU. Moreover, Brown provides a table (Brown, 1973; pp. 56) in which he estimates how large the
Upper Bound is for the different MLU-stages. Apparently, Brown considered positive outliers within a session to be meaningful and regarded them as an indicator of syntactical development. If the concept of Upper Bound has been a developmental measure in linguistics, it also makes sense to add the concept of a Lower Bound. If positive outliers are meaningful, it might be the case that negative outliers are meaningful as well because they bear information on the degree of consolidation of the linguistic rule. The expectation is that when consolidation increases, the child ceases to use the old rule. In this case, the area between Upper and Lower Bound can be considered as a range or bandwidth of utterance lengths that are produced in this specific context. The concept of bandwidth of scores fits in very well with the method of moving minima and maxima we discussed above. Note, however, that the Lower Bound can be operationalized in many different ways and does not necessarily need to be the absolute number of morphemes in the shortest utterance in the speech sample, because it is highly likely that this value is 1. Even in adult speech there are utterances that have only one morpheme. What is important is the distribution of utterance lengths. Are these lengths normally distributed or are they skewed? The Lower Bound can in this case be conceptualized as the “lowest range” of utterance lengths. For instance, what does the lowest 20% look like? How many of these utterances consist of only one morpheme, how many of two morphemes, etcetera? Not only the absolute values in this bandwidth are important, but especially the distribution of scores. This distribution can be analyzed for skewness, as we illustrated with the data of Jan and Eva. In the microgenetic literature, skewness has sometimes been conceptualized in relation to bimodality (Alibali & Goldin-Meadow, 1994). Bimodality refers to a situation in which two equilibrium levels exist. In the case of the acquisition of linguistic rules, the first equilibrium represents the old “rule” and the second equilibrium the new rule. Bimodality can be an indication for a period of rapid change, or a developmental transition between two stages (van Geert, 1998; van der Maas & Molenaar, 1992; Hosenfeld, van der Maas & van der Boom, 1997).

Furthermore, the idea of a score range fits in very well with the conception of scores as fuzzy numbers. This approach originates from Fuzzy Logic (see for instance McNeill & Freiberger, 1994; Nguyen & Walker, 1997, Ross, 1995) which offers the possibility to quantify vague boundaries. This can be done by conserving the fuzzy characteristics, which are essential to human behavior and the competence that directs this behavior. Fuzzy logic offers the possibility to assign a degree of “characteristicness” (to scores. While certain scores are very characteristic for an individual, other scores may be very uncharacteristic. This does not mean that for this individual these scores cannot appear. The chance that they will appear is, however, much smaller. In the range of observed scores there are many scores that are characteristic of that individual. At the same time, part of the scores are much less characteristic, but have nonetheless occurred under certain circumstances. For instance, a child can have an MLU of 2.5, which is very characteristic of that child, while an MLU of 4 is much less characteristic. This score may have occurred under rare, optimal circumstances (van Geert, 2002).
2.4.2 Implications for diagnostic use of MLU

This overview has shown that variability in developmental data must not be underestimated. This article illustrated the phenomenon of variability with data from early language development. From the literature, we know that similar variable results are also found in other developmental domains, such as motor development, early infant emotions, etcetera. Naturally, variability also has its practical implications, for both scientific and practical applications. At least, variability stresses the need for repeated measurement of the developmental grower in question. From a classical viewpoint, taking only a single measure of behavior that displays large intra-individual variability jeopardizes test-retest reliability. The illustration from MLU is crystal clear. Both Jan and Eva could be placed in several MLU stages in the period of only a few weeks. It speaks for itself that a single calculation of MLU is not adequate. In the discussion of the literature on MLU, we stated that in most cases only a single estimation of MLU is standard practice. The predictive value of this single MLU estimation is questionable, to say the least. Variability as found in this study also stresses the need for repeated measurement. In our opinion this does not mean that the average over several different measurements necessarily gives a better indication of the “true” MLU level. The average can have useful predictive value, as an approximation of the location of acquired scores (for instance is this band high or low?). On the other hand, it should be realized that variability could also carry information that can be used for diagnostic purposes. There are theoretical and empirical indications that variability is a normal characteristic of healthy development. A lack of variability might be a predictor for developmental problems. For instance, infant research found that normal infants show more variability in crying behavior than infants whose parents sought help because their infant cried excessively. Empirical research will have to show whether a similar pattern also applies to language development.

In conclusion, we have shown that the intra-individual variability that exists in other time-serial research, also applies to Mean Length of Utterance. This is demonstrated by the fact that the two children in the case studies fall into several MLU-stages within the period of several days. We have also shown that variability is not always sufficiently explained by the measurement error hypothesis. Instead, it could be analyzed as an additional source of information, which might also be applicable in the context of diagnostics and prediction. In the literature variability has been noted as both the source of development (for instance Berthenthal, 1999) and an indicator of a specific moment in the developmental process, namely in the presence of a developmental transition (van der Maas & Hopkins, 1998). In order to visualize these meaningful aspects of variability we have suggested several new methods. These methods rest heavily on visual inspection, and can be used to generate hypotheses. These hypotheses can be tested using new statistical techniques such as resampling (see Good, 1999). In this way it is possible to distinguish “noise” from meaningful aspects of variability. Although the research on variability is still taking baby steps, this approach offers an intriguing new perspective on irregularities in development.