Notes on Ingram's whole-word measures for phonological development *

Helena Taelman, Gert Durieux and Steven Gillis

University of Antwerp

Address for correspondence:

Helena Taelman, Gert Durieux and Steven Gillis

University of Antwerp,

Dept. of Linguistics – GER,

CNTS,

Universiteitsplein 1,

B-2610 Wilrijk,

Belgium

Helena.Taelman@ua.ac.be Gert.Durieux@ua.ac.be Steven.Gillis@ua.ac.be

0032 - 3 820.27.89

Title note

The research reported in this paper was supported by a GOA grant 'Computational Psycholinguistics'. The first author was supported by the National Science Foundation – FWO.

Abstract

In this note we discuss pMLU, a whole-word measure for phonological development that was proposed by Ingram (2002). Ingram's rules for calculating pMLU are analyzed and we point at the crucial role of the level of transcription for making pMLU measurements comparable over different corpora. The main aim of the paper is an assessment of the reliability and the validity of pMLU. The assessment is accomplished using a computational tool for measuring pMLU on two large Dutch CHILDES corpora. We propose minimal sample sizes for reliable measurements relative to the stage of phonological development.

INTRODUCTION

How can children's progress in language acquisition be measured? Probably the best-known yardstick is the Mean Length of Utterance, or MLU. The development of this measure has been a long process. According to Ingram (1989), Margaret Nice (Nice, 1925) introduced the Average Length of Sentence as a measure for quantifying syntactic development on the basis of spontaneous speech samples. In his 1973 book, Roger Brown took the measure one step further, and formalized rules for calculating MLU in speech samples of children acquiring English. Brown cut up the MLU continuum in stages of equal size, and related each stage to specific qualitative developments, such as the acquisition of particular morphemes. Further studies established the reliability and validity of MLU (e.g. Rondal, Ghiotto, Bredart & Bachelet, 1987), and investigated the relation with other developmental indexes such as age (e.g. Miller & Chapman, 1981) or lexicon size (e.g. Bates, Bretherton & Snyder, 1988). At present MLU is a standard index of children's morphosyntactic proficiency.

In his paper 'The measurement of whole-word productions', Ingram (2002) introduces a measure for phonological development similar to MLU, called phonological mean length of utterance (pMLU). The main aims of the measure are (1) to quantify phonological development in a straightforward way, and (2) to focus on the child's whole-word productions instead of the production of specific segments. pMLU reflects the length of the child's words and the number of correct consonants, and is formally defined as the length of the child's word productions (in segments) plus the number of correct consonants in each production divided by the total number of word tokens. Thus,

if a child says 'nana' for 'banana', the score for this particular utterance would be six pMLU points, i.e. four for the length in segments, plus two for the number of correct consonants (the two /n/'s). PMLU is computed by averaging the score over a number of utterances (preferably 25 or more, cf. infra).

In a vein similar to Brown's (1973) formalization of the rules for calculating MLU, Ingram proposes a number of guidelines for computing pMLU: he discusses criteria for selecting words from spontaneous speech samples, provides recommendations on sample size, and points out some common difficulties in dealing with transcribed speech. For the sake of convenience, Ingram's guidelines are listed in Table 1; they will be discussed later on.

Just like Brown (1973) cut up the MLU continuum in stages of equal size (Early Stage I = MLU 1.00 - 1.49, Late Stage I = 1.50 - 1.99, Stage II = 2.00 - 2.49, etc.), Ingram proposes to divide phonological development into six stages using pMLU as his yardstick. These stages are given below:

(1)

Stage	Range	Midpoint
I	2.5 – 3.5	3.0
II	3.5 – 4.5	4.0
III	4.5 – 5.5	5.0
IV	5.5 – 6.5	6.0
V	6.5 – 7.5	7.0
Beyond V		

Ingram also defines a number of other measures, which complement pMLU. The first of these measures is the Proportion of Whole-word Proximity (PWP), which relates the complexity of the child's productions to that of the attempted adult targets. It is obtained by dividing the child's pMLU score by the pMLU of the adult target forms. Two children with highly comparable pMLU scores may differ quite dramatically in the kinds of words they are trying to produce, since unsuccessful attempts at longer words and perfect attempts at shorter forms will yield similar pMLU scores. For example, a child that attempts a number of trisyllabic or longer words, but truncates the majority of them, will obtain a pMLU score similar to that obtained by a child that only attempts mono- and bisyllabic targets and prodoces them correctly. By bringing the complexity of the attempted targets into the equation, PWP is able to reflect this difference. A further measure is the Proportion of Whole-word Correctness (PWC), which is the ratio of correct attempts over the total number of productions. This measure provides a straightforward means to assess the overall correctness of the child's productions with respect to the adult targets. The final measure proposed is the Proportion of Whole-word Variation (PWV), which is meant to give an indication of the consistency (or lack thereof) with which target forms are produced. Together, these measures cover correctness, complexity and consistency of whole-word productions.

Ingram demonstrates the value of the pMLU measure by applying it in a wide range of contexts: these include a comparison of monolingual children, a comparison across languages, and the diagnosis of impairment or delay. Despite the apparent

usefulness of the measure, its practical application remains cumbersome if pMLU is to be calculated by hand. At present this is a major obstacle barring further development and promotion of the measure. Furthermore, in contrast to well-established measures such as MLU, little is known about the reliability and validity of the pMLU measure. The aim of this paper is twofold: (1) In the first section, we will sketch a procedure for computing pMLU automatically, on the basis of standard CHAT transcriptions. (2) In the second section, the computational procedure will be used to assess the reliability and the validity of pMLU on longitudinal data. Both the computer program and the empirical test will lead to a critical assessment of Ingram's calculation rules.

Calculating pMLU automatically

Ingram's guidelines for computing pMLU are listed in Table 1. Three types of guidelines can be distinguished. (1) The third guideline defines the *units of analysis*: it spells out what counts as a word. (2) The first two guidelines, the Sample-Size Rule and the Lexical-Class Rule concern the *selection* of target words to be analyzed. (3) The last three guidelines discuss the *model-replica analysis* of target words and productions. To facilitate the application of these guidelines on large data sets, a semi-automatic procedure was developped. All guidelines except the Sample-Size Rule are incorporated in our computer program.

Insert Table 1 about here

The input required is a standard CHAT transcription, containing an orthographic transcription and a phonetic transcription of the child's utterances, and a phonetic transcription of the adult targets attempted by the child. This input is processed in three steps. In the first step, a list is constructed of the attempted targets together with their different realizations. One of the standard CLAN tools, MODREP, is used for this purpose. The units of analysis, as spelled out in the Compound Rule, are defined in this first step: the model-replica analysis automatically analyzes all strings of characters as single words. In the second step, particular words can be excluded from the analysis on the basis of Ingram's Lexical Class Rule. As this step relies on the analyst's appreciation of particular cases, manual intervention is required here. Finally, in a third step, pMLU is calculated on the basis of a program that was specifically developed for the calculation of pMLU. The program incorporates the guidelines concerning the model-replica analysis: the Variability Rule, the Production Rule and the Consonants Correct Rule. This program together with a more detailed discussion of the calculation procedure is available on request from the authors.

Although the calculation rules laid out by Ingram are quite detailed, there are a number of issues in need of clarification. The first of these concerns the proper treatment of metathesis. By way of example, the Dutch word '*wesp*' (/wEsp/, wasp) is commonly rendered as '*weps*' /wEps/ in child forms. Following one line of reasoning, the PMLU score for this form would be seven i.e. four from the production rule, which counts the length in segments, plus three from the consonants correct rule: all three consonants turn up correctly, although not in the correct order. Under an alternative interpretation, this

form would yield a pMLU score of six, discounting one of the final consonants as out of sequence. This is the solution adopted in our program.[i] It is not entirely clear which of these interpretations is intended by Ingram, although we believe that giving up on the strict ordering constraint may give rise to many spurious correspondences, such as /lip/ and /pil/ as equivalent renditions of the target form /lip/. If the child utters 'lip' correctly this would of course yield identical results in both interpretations. But if we apply the strict ordering constraint /pil/ would yield a score of three (for the three segments). If we do not apply the strict ordering constraint we arrive at a score of five: three points for the three segments and an additional two points for the /p/ and /l/ that are produced though not in the correct place.

Another issue concerns the granularity of the phonetic or phonemic transcription. The level of phonetic detail exerts an influence on the resulting pMLU score, and should therefore be standardized in some way. Comparison of children's productions and the resulting pMLU scores may become awkward if one transcription is a 'broad phonemic' one and another a 'narrow phonetic' one. Since the empirical investigation in this paper is based on two corpora which use different conventions, we decided to recode all corpus files that contained a narrow phonetic transcription. This was done by reducing acceptable allophones to their standard phonemic form. All reported results are based on this recoded version.

On a related note, there is the issue of accuracy of transcription. Ingram (2002: 718) states that "... transcribers have more difficulty reaching agreement on the correctness of vowels than consonants", and therefore chooses to bias pMLU more

towards the latter. While this may be well founded for English, with its complex vowel system, one wonders whether this also holds for languages with far simpler vowel systems, e.g. only cardinal vowels. In general, however, it seems obvious, that reliability of transcription and level of phonetic detail are inversely related: sacrificing some phonetic detail will generally reduce disagreement among transcribers. Using a narrow phonetic transcription has its own advantages, though. A narrow phonetic transcription introduces in the pMLU measure the child's mastery of allophonic variation in addition to phonemic distinctions. In general, measuring pMLU at different levels of phonetic detail may result in a better diagnosis of the child's phonological proficiency.[ii]

An empirical evaluation of pMLU

Data

Our evaluation of pMLU will be based on two large phonemically transcribed databases of Dutch child language, which are available through CHILDES (MacWhinney, 2000): the first is the MAARTEN-corpus (Gillis, 1984), the second one is the CLPF-corpus (Fikkert, 1994; Levelt, 1994). The Maarten-corpus consists of 19 observation sessions of a single child between 1;9.29 and 1;11.15. The average duration of an observation session was two hours. The CLPF-corpus contains similar longitudinal, naturalistic data of 12 children. The children were followed for approximately one year; their age at the onset of the observations was between 1;0 and 1;11. Observation sessions lasted 30-45 minutes.

The data in the MAARTEN-corpus are transcribed phonemically, whereas the CLPF-corpus contains a narrow phonetic transcription of the children's utterances. For the purpose of this study, the phonetic tier in the CLPF-corpus was translated into a broader, phonemic transcription, viz. the one also used in the MAARTEN-corpus.

Reliability

The reliability of a measure refers to the consistency with which test items yield comparable indices of the ability being assessed. Reliability can be affected by a number of factors such as, in the case of pMLU, the conditions of recording, the transcriber's accuracy, and the size of the sample. In this paper, we focus on the latter: when we calculate a pMLU value based on a particular sample of utterances, we would like to be reasonably certain that the obtained value is close to the value that would have been obtained had we chosen a different sample. In other words, suppose we take a sample of 25 utterances out of a file of our corpus, and we calculate the pMLU value, and suppose we calculate pMLU again on a different selection of utterances from the same file, how close would these pMLU values be? If pMLU is a reliable measure given a selection of 25 utterances, the two values should be 'very close'. Intuitively, we expect smaller differences the larger the samples are. Therefore, to obtain reliable results, it is necessary to include a sufficiently large number of test items. Ingram suggests that at least 25 words, and preferably 50 words, are required to arrive at reliable pMLU calculations. He illustrates this sample size rule by calculating pMLU on three different 25-word samples taken from a corpus of a single child. The obtained values range from

6.2 to 6.6, which leads Ingram to conclude that the proposed sample size is indeed adequate for the purpose at hand. But a single example seems a rather small basis for firm guidelines. In the following, we will pursue the issue in a more systematic way, and verify whether the proposed guidelines hold up under closer scrutiny. The way we will go about this is as follows: first, we will propose a way to measure reliability, which is an important first step in coming to terms with the issue. Next, we will determine what degree of reliability we would like to obtain. Finally, we will explore some variations in both sample size and proficiency level, in order to determine the minimal required sample size for the desired degree of reliability.

When we select a 25-word sample from a session, we obtain the pMLU value by averaging over individual scores. Selecting a different sample would yield a comparable, but slightly different value, as Ingram's example shows. The question is whether we can find a way to determine the bounds within which these values vary. One way to go about this is by repeating Ingram's exercise a large number of times, say, one thousand times. We could then look at the lowest and highest value obtained, and these determine the bounds of the variation. Alternatively, we could discard a number of values, say 25 (or 2.5%) at either end of the scale, and thus obtain the bounds within which 95% of the values lie. Technically, this procedure is called the bootstrap (Efron, 1979; Stine, 1990), and has as its major advantage the fact that it does not need to make any assumptions about the distribution of values. The disadvantage is that the procedure is computationally expensive, and obviously only possible if automated. There is however another way, which relies on what is known about the distribution of values in one sample. This

method is based on the Standard Error of the Mean (SEM)[iii]; SEM is computed as the standard deviation of values in the sample, divided by the square root of the number of observations. The lower and upper bounds of the variation between different measurements can be derived from the SEM on the basis of the formula in (4) (Blalock, 1985: 180-186; Woods, Fletcher & Hughes, 1986: 103).

(4) mean ± t x SEM where t stands for the appropriate value in the t distribution, (e.g. t=2.06 when n=25)[iv]

The formula determines 95% confidence intervals around the sample mean, which is equal to the pMLU of that sample. Suppose that the pMLU value of a 25 word sample is 4.5 and the SEM is 0.5. On the basis of the formula in (4), we estimate the 'real' pMLU to be in between the lower bound of 3.47 (4.5 - 2.06 x 0.5) and the upper bound of 5.53 (4.5 + 2.06 x 0.5) with a probability of 95% (see also Klee & Fitzgerald, 1985 for an application of this method to MLU).

To illustrate the procedure, we will calculate pMLU values and confidence intervals for Robin, one of the children in the CLPF-corpus. In Figure 1 the development of Robin's pMLU is presented over a period of 9 months (1;7.13 - 2;4.28). The pMLU values are based on random samples of 25 words extracted from 19 sessions. The bars indicate 95% confidence intervals, derived from the SEM using formula (4). Figure 1 shows rather large differences in the width of the resulting confidence intervals: the narrowest confidence interval is ±0.46 at age 1;7.27, the largest confidence interval is

±1.55 at age 2;4.28. The trend is for confidence intervals to grow larger with increasing age and pMLU: the first 5 data points have an average confidence interval width of ±0.55, the last 5 data points have an average confidence interval width of ±1.15. Thus, the younger the child and the lower the pMLU measure, the higher the precision of the pMLU measure.

Insert Figure 1 about here

The question now crops up whether the width of these confidence intervals is acceptable, and if not, whether more appropriate confidence intervals can be found. In order to answer these questions, it is important to keep the purpose of the measure in mind. Since the measure is intended to serve as the basis for a developmental scale of children's whole word proficiency, the measure should be precise enough to capture children's progress even at relatively short developmental spans. Any assessment of the measure's reliability should therefore be made relative to this intended purpose. If e.g. pMLU is found to increase slowly over the course of development, higher precision is needed than if pMLU were a fast-growing measure. Thus, before determining the desired bounds of the confidence intervals, we need to know the average growth rate of pMLU over the course of development in normal children.

To do this, we will widen the scope somewhat, and calculate pMLU values for Robin as well as for eight other children. The data selection encompasses all children in the CLPF-corpus for which at least 10 datapoints were available, spanning the entire

range of available sessions. Table 2 identifies the children and their age ranges, the number of sessions analyzed, and the pMLU values for the first and the last session. The next columns relate to the analysis of the rate of development. Given the pMLU values for each session and the child's age (in days) at each session, a linear regression was performed through the pMLU values. The column in Table 2 headed by 'F value' shows that the linear fit was significant for 8 out of 9 children. The strength of the correlation between age and pMLU is measured in the next column, headed by 'R'. The high correlation coefficients (range = 0.421 - 0.973) indicate that a large proportion of the variability in pMLU scores can be accounted for by the linear regression equation. The last column in Table 2 provides crucial information concerning the increase of pMLU per day and lists the x-coefficient in the linear regression equation, which yields the increase of pMLU per day. The average growth rate is 0.006 points per day (SD=0.004) or 0.18 points per month (SD = 0.12).

Insert Table 2 about here

On the basis of the average growth rate, we can translate confidence intervals into time intervals: a confidence interval of pMLU ±1 corresponds to a time interval of ±6 months. This means that the child's actual level of proficiency may lay 6 months ahead or before the measured proficiency. The MLU, the morphosyntactic equivalent of the pMLU is much more precise. Miller & Chapman (1981) find a growth rate of 0.1 morphemes per month, Rondal et al. (1987) report on confidence intervals smaller than ±0.25 for MLU <

2 and smaller than ±0.5 for MLU < 3.5 (confidence interval = 2 times SEM). When combining these two findings, we obtain a confidence interval of less than ±2.5 months for MLU< 2 and of less than ±5 months for MLU<3.5. If we want pMLU to achieve the same degree of reliability as MLU, we should aim at similar confidence intervals. We will use the time span of ±3 months (or pMLU 0.5) as our criterion.

Returning to the development of Robin's pMLU (see Figure 1), only 2 out of 19 datapoints have a confidence interval at or below ±0.5 pMLU, or ±3 months. Clearly, the selected sample size is inadequate to obtain the desired reliability. The question is whether more appropriate sample sizes can be found.

In order to determine the influence of sample size on the width of the confidence intervals, we extracted random samples of 25, 50 and 100 words from all observation sessions with sufficient word types in the CLPF-corpus and the MAARTEN-corpus. For each sample, pMLU and confidence intervals were computed. Since we know on the basis of Robin's data that confidence intervals may increase with increasing pMLU, this factor was taken into account in the analysis: the results were grouped into the developmental stages established by Ingram (2002, see (1)): (1) $3.5 < \text{pMLU} \leq 4.5$, i.e. Stage II in Ingram's model; (2) $4.5 < \text{pMLU} \leq 5.5$, i.e. Ingram's Stage III, (3) $5.5 < \text{pMLU} \leq 6.5$, i.e. Ingram's Stage IV, (4) $6.5 < \text{pMLU} \leq 7.5$, i.e. Ingram's Stage V. Data for other stages were unfortunately not (or insufficiently) available.

Insert Table 3 about here

Table 3 shows per stage the mean 95% confidence intervals calculated for all three sample sizes (25, 50 and 100 words), as well as their standard deviations (SD) and the number of sessions (i.e. files) they were computed on. The widths of the confidence intervals in Table 3 range from ±0.36 to ±1.19. As expected, the confidence intervals are larger in smaller samples. In addition, the size of pMLU exerts influence: the size of the confidence intervals increases with approximately 0.50 from Stage II to Stage V. The data in Table 3 permit us to pinpoint the (minimal) sample size for achieving an acceptable confidence interval of ±3 months or a confidence interval of pMLU ±0.5. For Stage II ($3.5 < \text{pMLU} \leq 4.5$) a sample size of 25 suffices. For Stage III ($4.5 < \text{pMLU} \leq 5.5$) 50 words are required for a confidence interval of approximately ±0.5. Finally, a minimum of 100 words is needed in order to maintain the same confidence intervals for Stages IV and V. Hence, we propose to change Ingram's sample size rule in the following way: (1) Analyze all available word tokens; (2) Observe a minimum of 25 words in Stage II, a minimum of 50 words in Stage III, and a minimum of 100 words in the later stages; and (3) Report the sample size and the standard deviation (data needed to derive confidence intervals).

Our guidelines were validated by computing the inter-measure reliability. This notion denotes the degree to which two independent samples will yield comparable results. In order to get at this type of reliability, the following steps were taken. From the MAARTEN-corpus and the CLPF-corpus, all sessions (n = 29) were selected that contain at least twice the number of minimally required word types.[v] From each session two independent random samples were drawn, each containing the proposed minimum

number of word types. Next, pMLU was calculated for each sample. The correlation between the pMLU values in both samples was very high: R=0.811. When two 25 word samples were randomly selected from the same sessions, the correlation was considerably lower: R=0.616. Thus, the proposed sample size requirements result in a much higher inter-measure reliability than the minimal sample size of 25 items that Ingram recommends.

The downside of our recommendations is that data requirements are more stringent. Note that the proposed sample size requirements are often, but not always fulfilled in 30-45 minute sessions of spontaneous speech. For instance, the CLPF-corpus contains 212 sessions of 30-45 minutes. One third of the sessions, 81 to be precise, have an insufficient number of word types: all Stage I sessions (12 out of 12), half of the Stage II sessions (22 out of 47), one fifth of the Stage III sessions (17 out of 68), and one third of the Stage IV and Stage V sessions (30 out of 85).

Validity

The practical usefulness of pMLU not only hinges on the reliability of the measure, but also on its validity. Thus, the question we will turn to now is: does pMLU accurately reflect the child's phonological development? One way of answering this question is by comparing pMLU to other measures of phonological development. However, most phonological tests rely on elicited data, and cannot be applied to spontaneous data such as those used in our analysis. Besides pMLU, we know of one other common phonological measure used for spontaneous speech: the percentage of

consonants correct (PCC, Shriberg & Kwiatowsky, 1982). Unfortunately, the PCC is not suitable for an assessment of the validity of pMLU, since the criterion of independence is not met: both PCC and pMLU rely on a count of correct consonants.

An alternative way of assessing the validity of pMLU is to analyze the extent to which pMLU can be considered a pure measure of phonological development. Phrased differently, the question is to what extent pMLU reflects other domains of language proficiency, such as the child's morphosyntactic development. Ingram (2002) seems aware of this issue: in his lexical class rule, he excludes function words from the analysis, because these words are usually short and, hence, lower the overall pMLU score in morphosyntactically advanced children. However, this guideline does not completely exclude the influence of morphosyntactic factors. For English and Dutch, there may be a confound between word length and inflectional complexity: inflected words are often longer than non-inflected ones. For instance, once English-speaking children start to use inflectional morphology, adding a suffix like s in *walks*, increases the length of the child's words and consequently increases the pMLU score. The third person singular suffix s may be argued to reflect the child's growing morphological or morphosyntactic capacity rather than her phonological capacity.

In order to determine the impact of morphological development on Dutch children's pMLU, Robin's pMLU values were computed twice, once using all available word forms (pMLU), and once using only word lemmas (pMLU-lemmas). In the latter case only the bare form of nouns were analyzed, excluding plural nouns, diminutivized nouns, etc. The verbs analyzed were restricted to the infinitive form, and adjectives were

restricted to the undeclined form. The two curves in Figure 2 represent these two ways of arriving at Robin's pMLU.

Insert Figure 2 about here

As expected, both curves are closely intertwined at the very beginning; from 2;1.6 onwards they become clearly divorced, which means that these two pMLU scores indeed measure different aspects of the child's production. At the very end of the observation period, Robin's pMLU-lemmas score is 6.07 and his pMLU computed using all word forms is 6.40. An analysis of the entire database (all children in the CLPF-corpus and all data from the MAARTEN-corpus) confirms this observation. In the 134 sessions analyzed[vi], there is an average difference between both pMLU values of 0.31 ($t=16.0$, $p < .001$). As pMLU increases, the difference between pMLU and pMLU-lemmas increases as well ($F(1, 132) = 75.17$, $p<0.001$). In Stage II the mean difference is only 0.13, in Stage III it is 0.23, in Stage IV it is 0.45, and in Stage V it is 0.59.

This analysis leads us to the conclusion that, indeed, in a language like Dutch, pMLU reflects morphological development in addition to phonological development. Whether or not this is a reason to select only uninflected words for the computation of pMLU, depends on the goal of analysis: it is crucial for studying particular fundamental research questions, such as the influence of phonological proficiency on morphosyntactic development, whereas it is less important for the purpose of diagnosis in delayed children. Furthermore, there are typological considerations: in Dutch and in English there

may be a confound between word length and morphological complexity, but this is certainly not true for all languages.

Conclusion

Ingram (2002) proposes several new measures of phonological proficiency, most notably pMLU, the phonological mean length of utterance. This measure is intended as a yardstick for phonological development, and forms the basis of a developmental scale.

In this paper, we have outlined a procedure for the automatic calculation of pMLU, which greatly facilitates its practical application. The procedure relies on standard CHAT tools, in combination with a custom-made computer program to calculate pMLU. Development of the program uncovered a number of issues concerning the calculation of pMLU which are in need of clarification: the case of metathesis, the level of phonetic detail in the transcription, the exclusion of vowels in the count of segments correct.

A second goal of this paper was to provide an assessment of both the reliability and validity of the pMLU measure. Reliability was tested on two longitudinal databases of Dutch speaking children, the Maarten-corpus and the CLPF-corpus. Our results indicate that the recommendations on sample size from Ingram (2002) are in need of revision. Ingram's original proposal was to include at least 25 - and preferably 50 - words into the analysis. We found that a sample size of 25 words is too small to obtain reliable results for pMLU values greater than 4.5. Even a sample size of 50 words does not result in reliable pMLU values for pMLU scores greater than 5.5. Only for pMLU values less than or equal to 4.5 are 25 words sufficient. As an alternative to Ingram's sample size

rule, we propose to include all word types, observing a strict minimum requirement of 25 words for pMLU≤4.5, a minimum of 50 words for pMLU>4.5, and a minimum of 100 words for pMLU>5.5. Furthermore, we recommend that standard deviations and sample sizes be reported.

Finally, the validity of the pMLU measure was investigated. Studying validity by comparison to other measures of phonological proficiency was found difficult in practice, for want of comparable independent measures. Instead, we chose to assess pMLU on its own merits as a purely phonological measure. This was done by factoring out the contribution of morpho-syntactic development to the calculation of pMLU. Perhaps unsurprisingly, it was found that pMLU is not a pure measure of phonological development, since it partly reflects the child's morphosyntactic proficiency. pMLU values are higher when inflected words are included in the calculation than when only lemmas are considered. Whether or not this confound is sufficient reason to discard inflected words from the computation of pMLU, depends largely on the goal of the analysis. Moreover, not all languages will be the same in this respect.

In the meanwhile, we found the pMLU measure to increase with age in most of the studied children at a rate of about 0.18 per month. Further research is needed in order to establish developmental norms and to determine the correlation between pMLU and other measures of language acquisition.

References

Bates, E., Bretherton, I. & Snyder, L. S. (1988). *From first words to grammar*. New York, NY: Cambridge University Press.

Blalock, H. (1985). *Social statistics*. London: MacGraw – Hill.

Brown, R. W. (1973). *A first language: The early stages*. Cambridge, Mass.: Harvard University Press.

Cormen, T. H., C. E. Leiserson & R. Rivest (1999). *Introduction to algorithms*. (23 ed.). Cambridge, MA: MIT Press.

Efron, B. 1979. Bootstrap methods: another look at the jacknife. *Annals of Statistics* **7**, 1-26.

Fikkert, P. (1994). *On the acquisition of prosodic structure.* Rijksuniversiteit Leiden, Leiden.

Gillis, S. (1984). *De verwerving van talige referentie.* Universiteit Antwerpen, Antwerpen.

Ingram, D. (1989). *First language acquisition: method, description, and explanation*. Cambridge: Cambridge University Press.

Ingram, D. (2002). The measurement of whole-word productions. *Journal of Child Language* **29**, 713-733.

Klee, T. & Fitzgerald, M. D. (1985). The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language* **12**, 251-269.

Levelt, C. C. (1994). *On the acquisition of place*. Rijksuniversiteit Leiden, Leiden.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* (3 ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Miller, J. F. & Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research* **24**, 154-161.

Nice, M. M. (1925). Length of sentences as a criterion of a child's progress in speech. *Journal of Educational Psychology* **16**, 370-379.

Rondal, J. A., Ghiotto, M., Bredart, S. & Bachelet, J.-F. (1987). Age-relation, reliability, and grammatical validity of measures of utterance length. *Journal of Child Language* **14**, 433-446.

Shriberg, L.D. & Kwiatkowski, J. (1982). Phonological disorders II: A conceptual framework for management. *Journal of Speech and Hearing Disorders* **47**, 242-256.

Stine, R. (1990). An introduction to bootstrap methods. Examples and ideas. In J. Fox & J. S. Long (eds.), *Modern methods of data analysis*. Newbury Park/London/New Delhi: Sage Publications.

Woods, A., Fletcher, P. & Hughes, A. (1986). *Statistics in language studies*. Cambridge: Cambridge University Press.

Table captions

Table 1: Ingram's calculation rules

Table 2: Linear regression equations of the relation between age and pMLU in 9 children

Table 3: Confidence intervals around the pMLU in 4 pMLU stages

Table 1: Ingram's calculation rules

1. *Sample-Size Rule*: Select at least 25 words, and preferably 50 words for analysis, depending on sample size. If the sample is larger than 50 words, select a selection of words that cover the entire sample, e.g. every other word in a sample of 100 words.
2. *Lexical-Class Rule*: Count words (e.g. common nouns, verbs, adjectives, prepositions and adverbs) that are used in normal conversation between adults. This excludes child words, e.g. mommy, daddy, tata, etc. Counting child words can inflate the PMLU if a child is a reduplicator.
3. *Compound Rule*: Do not count compounds as a single word unless they are spelled as a single word, e.g. 'cowboy' but not 'teddy bear', i.e. 'teddy bear' would be excluded from the count. This rule simplifies decisions about what constitutes a word in the child's sample.
4. *Variability Rule*: Only count a single production for each word. If more than one occurs, then count the most frequent one. If there is none, then count the last one produced. Counting variable productions may distort the count if there is a highly variable single word.
5. *Production Rule*: Count 1 point for each consonant and vowel that occurs in the child's production. Syllabic consonants receive one point, e.g. syllabic 'l', 'r', and 'n'. (Some transcriptions may show these as two segments, i.e. a schwa plus consonant, e.g. 'bottle' [bɑdəl], but it should be counted as one consonantal segment.) Do not count more segments than are in the adult word. For example, a child who says 'foot' as [hwut] has two consonants counted, not three. Otherwise, children who add segments will get higher scores despite making errors.
6. *Consonants Correct Rule:* Assign 1 additional point for each correct consonant. Correctness in vowels is not counted since vowel transcriptions are typically of low reliability. Syllabic consonants receive an additional point in the same way as nonsyllabic consonants. A child who applies liquid simplification, for example, will get 1 point for producing a vowel, e.g. 'bottle' [bado], but 2 points if the syllabic consonant is correct.

Table 2: Linear regression equations of the relation between age and pMLU in 9 children

name	*age*	*pMLU*	*n*	*F Value*	*R*	*x coefficient*
Cato	1;10.11 - 2;7.4	4.75 - 6.66	16	F(1,14) = 244.5***	0.973	0.006891
Elke	1;8.13 – 2;4.29	4.08 – 5.57	12	F(1,10)=116.2***	0.960	0.0063666
Enzo	1;11.8 – 2;6.11	6.07-6.48	12	F(1,10)=2.2	0.421	0.0014704
Eva	1;4.12 – 1;11.8	4.08-4.67	10	F(1,8)=10.7*	0.757	0.0047304
Jarmo	1;9.9 – 2;4.1	4.29-5.47	12	F(1,10)=20.5**	0.820	0.0057994
Leon	1;10.1 – 2;8.19	5.12-5.74	13	F(1,11)=5.1*	0.561	0.0019619
Maarten	1;9.12 – 1;11.15	4.78-5.75	15	F(1,13) = 125.0***	0.952	0.0159569
Noortje	2;3.7 – 2;11.0	3.85-5.01	14	F(1,12)=21.8***	0.803	0.0061849
Robin	1;7.13 – 2;4.28	4.09-6.40	18	F(1, 16) = 154.0***	0.952	0.0080782

*p <0.05; ** p<0.01; *** p<0.001;

Table 3: Confidence intervals around the pMLU in 4 pMLU stages

sample size	*stage II*	*stage III*	*stage IV*	*stage V*
25	±0.51 (SD=0.11, n=27)	±0.74 (SD=0.17, n=80)	±0.99 (SD=0.19, n=57)	±1.19 (SD=0.20, n=28)
50	±0.36 (SD=0.04, n=5)	±0.52 (SD=0.09, n=63)	±0.71 (SD=0.10, n=75)	±0.87 (SD=0.12, n=10)
100	-	±0.38 (SD=0.05, n=17)	±0.50 (SD=0.05, n=49)	±0.58 (SD=0.06, n=9)

Figure captions

Figure 1: The development of Robin's pMLU: values represent the pMLU computed over 1 sample of 25 word forms, with 95% confidence intervals

Figure 2: Development of Robin's pMLU (dotted line) and pMLU-lemmas (straight line)

Figure 1: The development of Robin's pMLU: values represent the pMLU computed over a randomly selected sample of 25 word forms, with 95% confidence intervals

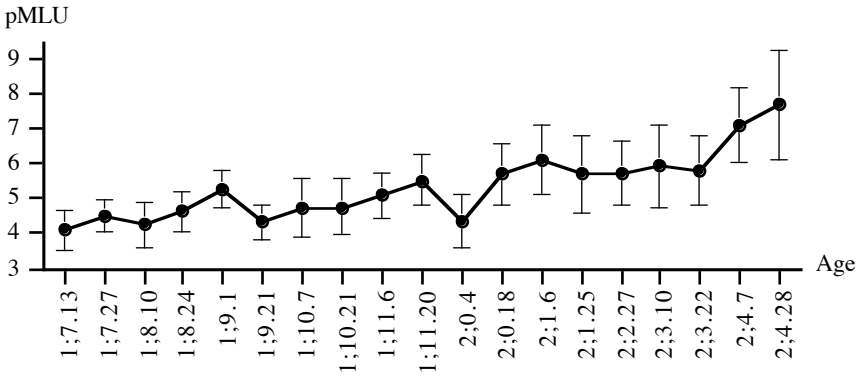
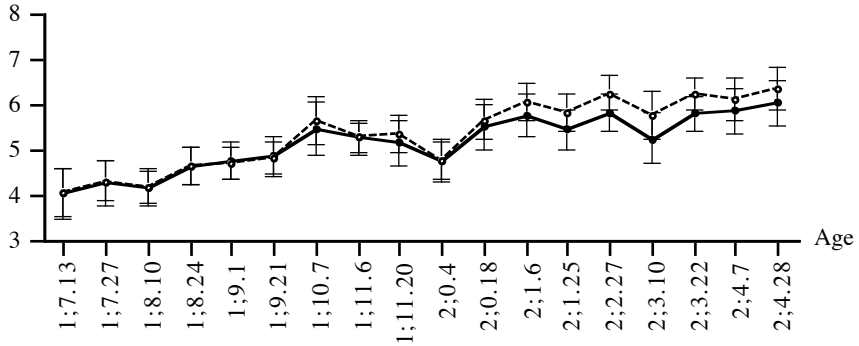


Figure 2: Development of Robin's pMLU (dotted line) and pMLU-lemmas (straight line)



ENDNOTES

i Technically, this is a direct consequence of our implementation of the consonants correct rule. In our program, this rule is cast as an instance of the longest common subsequence problem (Cormen, Leiserson & Rivest, 1999), calculated over representations of child and target forms, from which vowels are stripped. Calculation of the longest common subsequence proceeds using the standard dynamic programming solution to this problem.

ii We thank an anonymous reviewer for pointing this out.

iii A comparison of the two methods reveals only slight differences. Whereas the SEM method yields confidence intervals of equal length, the confidence intervals in the bootstrap analysis are not always symmetrical: for pMLU the length of the positive confidence interval is always lower than the length of the negative confidence interval. Thus, the probability of underestimation is somewhat higher than that of overestimation. Overall, the distance between the positive confidence interval and the negative confidence interval is slightly higher according to the bootstrap method than according to the SEM method. The size of the difference lies between 0.00 and 0.10.

iv In order to obtain a 95% confidence interval, we will use t=1.99 in case of sample sizes of 100 or higher, t=2.01 in case of samples sizes of 50 or higher, and t=2.06 in case of sample sizes of 25 or higher.

v The sessions were selected independently of the children's age. An interesting question for further investigation is whether the pMLU can reliably tap on the variation across children of the same age.

vi These were all files that fulfilled our sample size criteria for both pMLU and pMLU-lemmas.