Article

LANGUAGE
TESTING

Language Testing
2014, Vol 31(1) 3–18
© The Author(s) 2013
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265532213485570
ltj.sagepub.com
SAGE

The reliability of morphological analyses in language samples

Jodi Tommerdahl and Cynthia D Kilpatrick
The University of Texas at Arlington, USA

Abstract

It is currently unclear to what extent a spontaneous language sample of a given number of utterances is representative of a child's ability in morphology and syntax. This lack of information about the regularity of children's linguistic productions and the reliability of spontaneous language samples have serious implications for language testing based upon natural language. This study investigates the reliability of children's spontaneous language samples by using a test-retest procedure to examine repeated samples of various lengths (50, 100, 150, and 200 utterances) in regard to morpheme production in 23 typically developing children aged 2;6 to 3;6. Analyses indicate that out of five morphosyntactic categories studied, one of these (the contracted auxiliary) achieves an ICC for absolute agreement over .6 using 100 utterances while most others (past tense, third-person singular and the uncontracted 'be' in an auxiliary form) fail to reach a correlation above .52 even when samples of 200 utterances are compared. The study indicates that (1) 200-utterance samples did not provide a significantly greater degree of reliability than 100 utterance samples; (2) several structures that children were able to produce did not show up in a 200-utterance sample; and (3) earlier acquired morphemes were not used more reliably than more recently acquired items. The notion of reliability and its importance in the area of spontaneous language samples and language testing are also discussed.

Keywords

Acquisition, language samples, morphosyntax, reliability, test-retest

The transcription and analysis of spontaneous language samples are common practice amongst researchers and clinicians working with child language (Stalnaker & Creaghead, 1982; Cole, Mills, & Dale, 1989; Evans & Craig, 1992; Marinellie, 2004; Southwood & Russell, 2004). It is unsurprising that spontaneous samples have withstood the test of

Corresponding author:
Jodi Tommerdahl, Center for Mind, Brain and Education, The University of Texas at Arlington, Box 19545, Hammond Hall 413, Arlington, TX 76019, USA.
Email: joditom@uta.edu

time, as they are arguably the most appropriate way to get an overview of children's naturalistic language production. Spontaneous language samples are used to examine diverse aspects of language including turn-taking, mean length of utterance (MLU), number of utterances produced, vocabulary size, and phonological and semantic knowledge (Cole et al., 1989; Evans & Craig, 1992). They are also used in the analysis of morphological and syntactic structures (Crystal, Fletcher, & Garman, 1989; Marinellie, 2004). Their popularity is further evidenced by the existence of the CHILDES database (MacWhinney, 2000), a free online resource providing spontaneous samples of child language for researchers of child language worldwide to share. To date, over 3000 academic articles based on language samples from CHILDES have been published. Tomasello and Stahl (2004, p. 102) point out that audio and video recordings of interactions between children and another interlocutor are "the main form of naturalistic observation in the modern study of child language acquisition". Even within clinical environments having access to normed and standardized assessments, surveys indicate that a large majority of speech and language therapists use spontaneous samples in their professional practice (Kemp & Klee, 1997; Hux, Morris-Friehe, & Sanger, 1993).

It has been noted that the extensive use of observation and sampling in speech and language research means that dependability of measurement is essential for producing generalizable results (Scarsellone, 1998). However, despite the popularity and usefulness of language samples, there is little work regarding temporal reliability, or the regularity of language production from one occasion to another. This gap has been repeatedly recognized over several decades.

Fifty years ago, Minifie, Darley, & Sherman (1963) called for more information to be made known about the temporal reliability of spontaneous samples, stating that much of the work over the past 30 years was based on the researchers' assumption that "the numbers they obtained with their scales and measures were isomorphically related to "real" language development." Cole et al. (1989, p. 260) emphasized that "although reliability information is basic to the interpretation of test results, this measurement characteristic appears to have been generally overlooked in the area of language sample interpretation". Marinellie (2004, p. 519) reiterated that the reliability of samples in the area of syntax for normally developing children and those with Specific Language Impairment (SLI) is "a relatively uncharted area".

Tomasello and Stahl (2004), acknowledging that much child language data currently used for analysis is likely to represent only a very small portion of a given child's language, warn researchers of the potential risk of error present when analyzing language samples. Muma, in discussing the problems around language samples that are not necessarily representative of a child's typical speech states that "it is peculiar that the clinical fields have been silent about them and maintained a reliance on 50 and 100 utterances" (1998, p. 316).

Despite the lack of work speaking to the degree to which language samples represent the true state of a child's competence, the body of work using samples of spontaneous language continues to grow. In addition, probabilistic models of language processing are common, relying on distributional regularities to account for the relative ease with which different forms are processed (see Bybee, 1995, among others). As the body of work using distributional frequencies grows, many linguistic theories are also shifting their focus to the statistical regularities of language. Stochastic Optimality Theory (Boersma & Hayes, 2001) builds probabilities into the grammar, predicting, at least in some cases, the frequency with which forms may appear. Theories such as this rely mainly on frequency information from large corpora to support their claims. As this work grows to encompass analyses of not only adult language but also acquisition data, accurate information regarding the frequency and reliability of child language samples will become still more important.

While determining distributional frequencies of adult language is reasonable based on the large number of corpora available (see www.ldc.upenn.edu/), child language is more difficult to assess. As children proceed through stages of acquiring language, we would naturally expect that the frequencies of different forms would change as they acquire more vocabulary and increasingly complex syntactic structures. However, statistics related to how frequently children produce different morphosyntactic items are extremely limited. Many studies provide data from a single child, and even excellent resources such as the CHILDES database do not always contain more than a single sample from a given child. Analyses of these data generally involve taking a child's single sample for analysis, but the underlying assumption of such work may be that the recorded sample is a reliable indicator of the child's competence at the time. However, this leads to the question of how reliable one can consider a spontaneous sample of child language to be. Before moving into a discussion of work specifically focused on this question, we will first address the notion of reliability and its importance.

Reliability and language samples

The notion of reliability is a vital one to professionals concerned with using assessment tools. At its core, reliability refers to how generally trustworthy data is, given a particular measure, in this case a language sample. If a particular measure is reliable, then similar results should be obtained time and time again when used in identical conditions. While this sense of dependability of assessment is somewhat broad, the statistical concept of reliability under Classical Test Theory (CTT) (see Novick, 1966; Lord & Novick, 1968; Allen & Yen, 1979; Streiner & Norman, 1995, among many others) is much more precise. Under CTT, three elements are taken to be of importance for any given score: true ability, measurement error, and the actual observed score of the participant: $X = T + E$ (Allen & Yen, 1979). The participant's observed score, X, refers to the actual score that a participant gets each time that a test is administered. This observed score is composed of the participant's true score and error score. The true score, T, is the hypothetical mean score that the participant would get if they took the same test an infinite number of times.

Consider how this might work with a collection of language samples. Imagine that a language sample is collected from the same adult individual each day for several years and scored for use of particular morphemes. On some days, the individual might use some particular morphemes more than on other days. But overall, the mean of all the different samples would be an indication of that individual's true use of morphemes, and would comprise the observed score, X, of the individual. The random error (E) in CTT is assumed to have a normal distribution, so over the course of the multiple samples, the mean for E would be 0, with a variance of σ^2. By subtracting the error E from the observed score, we are left with the True score (T) of that individual. While this argument is based on multiple sampling of the same individual, Allen and Yen (1979) show that these assumptions hold for sampling of multiple individuals once, rather than multiple sampling of a single individual. It follows, then, that multiple language samples collected from different individuals will provide us with the same standard error of measurement. This can be figured using the variance for X, T, and E:

$$\text{Var (X)} = \text{Var (T)} + \text{Var (E)} \text{ (Kline, 2005, p. 93)} \tag{1}$$

The variance of the observed scores of individuals (X), is equal to the variance of the true scores (T) plus the variance of the error (E). Given (1) above, we can calculate reliability (R) as the ratio of true score variance to the observed score variance:

$$\text{R} = \text{Var (T)/Var (X) (Kline, 2005, p. 93)} \tag{2}$$

Combining (1) and (2) above, it is then clear that reliability is the ratio between the true score variance and the true score variance plus error variance:

$$\text{R} = \text{Var (T)}/[\text{Var (T)} + \text{Var (E)}] \tag{3}$$

Given (2) and (3), reliability will be high if the true score variance is high in relation to the observed score variance, but low if the true score variance is low in relation to the observed score variance. This ratio allows a measure of reliability in which perfect reliability would have a value of 1, while no reliability at all would have a value of 0. Looking back at the ratio given in (3), it can be seen that when the error variance is high, reliability will be lower. In general, reliability improves as the variance within subjects decreases, and reliability falls as the variance increases. Therefore, error variance lies at the heart of the reliability question.

While it is impossible to know what the true scores are of participants, it is possible to calculate reliability in several ways. One such way is through the use of a test-retest methodology, in which individuals are tested twice under the same conditions. The results are then compared to determine the degree of agreement between them. There are two types of agreement that could be considered here: relative and absolute. Relative agreement would rank the participants in the same way, but would allow each individual's scores to vary. For instance, given five individuals whose raw scores on a preliminary assessment are 1, 2, 3, 4, and 5, and whose scores are respectively 6, 7, 8, 9, and 10 on a second assessment, relative agreement would be perfect because all the participants' scores are ranked the same in relation to each other, even though their individual scores varied. Absolute agreement, on the other hand, would require that the test and retest scores of an individual participant agree, rather than simply ranking the participant scores in relation to one another. It is this absolute agreement that we are interested in here in terms of reliability. The appropriate measure for such a comparison is the intraclass correlation coefficient (ICC) (Shrout & Fleiss, 1979), which is the ratio of between-subjects variance to total variance. Using a test-retest, this would then be the sum of between-subjects variance (the differences between the subjects from the mean of all subjects) + between-time variance (the differences between the mean of the two time points) + within subjects variance, where ‘variance within subjects’ is the variation within one subject, which would be unexplained variance, or the error (assumed to be N(0, σ^2), where the data is normally distributed with a mean of 0 and variance of σ^2:

$$\mathrm{ICC} = \sigma^2_{\text{subjects}} / (\sigma^2_{\text{subjects}} + \sigma^2_{\text{time}} + \sigma^2_{\text{error}}) \qquad (4)$$

While the ICC is expected to give an accurate measure of the reliability between the language samples, there are additional factors that may affect the reliability of an assessment. For instance, reliability is positively influenced by the number of items that a test measures, with a greater number of items leading to higher reliability. Therefore, items with an expected high frequency may be expected to reach a significant correlation in a shorter sample than less frequent items. As Tomasello and Stahl (2004) show, the capture probabilities for targets of different frequencies is quite different, as highly frequent targets are quite likely to appear in limited samples, but the probability of less frequent targets appearing is much lower.

There are also other sources of random error that must be considered. In much research in speech-language pathology, the reliability of the observer (inter-rater reliability) appears to take the most prominent role in determining the reliability of observational data. Cordes (1994) surveyed 83 different research reports and found that 80% of them addressed reliability issues, but the reliability was, for the most part, determined by correlations or some other measure of agreement between observers. Furthermore, in data unrelated to language samples, such as that of animal behavior studies, inter-observer reliability has also been argued to be of great importance (Kaufman & Rosenthal, 2009). In the area of second language proficiency testing, Davidson (2000) notes that essay exams and oral interviews in L2 testing often give rise to disagreements that affect reliability, including discrepancies among raters.

While reliability of the observers is crucial to overall reliability, inconsistencies between raters are not the only cause of random error in CTT. Several other sources of random error are possible, including inconsistencies across different versions of an assessment or across difficulty levels of different items on an assessment, as well as inconsistencies of occasion, or differences in time and place. Davidson (2000) cites test tasks and time-of-day as factors that affect reliability of second language proficiency. In an investigation of generalizability from one occasion to another, Hernandez-Lloreda and Colmenares (2006) observed that failing to consider the time of day of sampling can lead to biased estimates. Inconsistencies of occasion such as these may result in an increase in error variance, and lead to low test-retest reliability.

In terms of language samples themselves, these same factors must be taken into consideration in determining whether a particular language sample is representative of a child’s actual language competence. For instance, a child who is tested at an odd time of day, when not feeling well, or with an unfamiliar interlocutor, may not produce language that is typical of his or her daily use. In addition, if an elicitation task is unfamiliar to a child, it may not effectively elicit the targeted forms. While these factors must all be considered, an additional question arises: to what degree can we expect a spontaneous language sample to be representative of a child's language? This is the question to which we now turn.

Previous work on reliability of spontaneous language samples

The challenge of establishing reliability in spontaneous language samples is complicated by the fact that language use will vary by multiple factors including sample length, age, and language status. In examining any particular language sample, the researcher needs to know whether or not conclusions can be drawn about the child's competence at the time of sampling. Of the few studies that have sought to determine the temporal reliability of spontaneous (meaning non-elicited) samples, most examine such global measures as mean length of utterance (MLU), with just a few considering the reliability of particular morphosyntactic items.

In the area of global measures, Fisher (1934) concludes that 50 utterances are not enough to be representative of a child's language in regard to mean length of response (MLR). However, Minifie et al. (1963) report that MLR reached a correlation of .82 and .77 for younger and older groups respectively with only 50 utterances. Despite the high degree of MLR reliability found previously, Barlow and Miner (1969) find that the Length Complexity Index (*r* = .80) outperforms MLR (*r* = .65) in terms of temporal reliability over three language samples of 50 utterances, each taken within 10 days in total. Cole et al. (1989) looked specifically at mean length of utterance (MLU), question use, morphological production of the past tense, regular plural and present progressive. Findings show high reliability for MLU, with a robust rank-order correlation of .92, but other items had much lower degrees of reliability with none over .50.

Additional work argues specifically that sufficient reliability of some global measures is reached with even samples of only a minute. Heilmann, Nockerts, & Miller (2010) compared samples of one, three, and seven minutes for both spontaneous and elicited language, looking specifically at total number of utterances, WPM, NDW and MLUm. The degree of reliability reached a minimum of .70 at 1 and 3 minutes when compared to the longer 7-minute sample, with 3-minute samples being slightly more reliable than 1-minute samples. These results are supported by the findings of Tilstra and McMaster (2007), which indicate that strong correlations from .70 to .94 were reached with very brief narratives elicited from a single picture and a highly standardized protocol.

Despite the positive results for shorter samples seen above, the question of how long a language sample must be in order to be reliable is one that is still at issue when additional measures are considered. For instance, Gavin and Giles (1996) examined total number of words (TNW), NDW, MLU-m and mean syntactic length (MSL) for samples of 12 and 20 minutes, showing that reliability improved overall with increased length of sample, but concluding that 175 utterances were needed in order to reach a level of acceptable reliability.

Johnson and Tomblin (1975) examine the specific morphosyntactic structures of indefinite pronouns, personal pronouns, main and secondary verbs, negatives, conjunctions, interrogative reversals, wh- questions and an overall Developmental Sentence Score. The highest degree of reliability was found in the use of personal pronouns, which reached a correlation of .92 at only 25 utterances. Due to the lower reliability of other items in shorter samples, the authors suggest the use of 175 utterances to improve reliability to a coefficient of 0.91.

Pushing the minimum length of sample even longer, Muma (1998) claims that the error rates in 50 utterance samples are as high as 55%, and at 100 utterances may still be at 40%. He argues that samples of at least 200–300 sentences are needed to accurately assess grammatical production.

The overall body of work carried out on the temporal reliability of spontaneous language samples to this point has primarily focused on elements which take the language sample in its entirety into account such as MLU or MLR, total number of words, number of different words, type-token ratio, and so on. A smaller amount of work has been done on the appearance of specific morphemes and syntactic structures. This area is of great importance given our need to use language samples as an accurate indicator of children's capacity for productive language. Only language samples that are understood in terms of reliability can allow us to appropriately research the acquisition and use of morphosyntax. We cannot overstate the claim that knowledge of the reliability of language production is absolutely foundational to all testing using spontaneous language.

The present study provides new information about the temporal reliability of five specific morphosyntactic targets as well as a measure of multi-verb sentences by using a test-retest procedure to determine if different samples from the same child collected under near-identical conditions yield similar results. Within each pair of language samples, we examine the consistency of use of these five morphosyntactic targets at 50, 100, 150, and 200 utterances. In addition, we test split-half reliability by comparing the first 100 utterances of the sample with the second 100, which allows us to see the degree to which the use of these different morphosyntactic targets is consistent across the sample as a whole. If a spontaneous language sample can be considered to be representative of the child's language at the grammatical level, the use of these different morphosyntactic targets should be highly correlated in the two samples. On the other hand, a low degree of correlation would indicate that a single spontaneous language sample may not be representative of a child's typical morphosyntactic production, at least not when looking at specific linguistic forms.

Methods

Participants

Twenty-three children (13 females, 10 males) are included here as participants. Four additional potential participants were excluded from analysis due to not producing the minimum number of utterances required in one or both sessions. The 23 included participants ranged in age from 2;6 to 3;6, with a mean of 35.6 months. All children were from monolingual English backgrounds and were considered to be typically developing based on parental report of the following criteria: normal hearing level; (2) no referral for speech or language therapy; (3) no suspicions of communication difficulties; (4) no

Table 1. List and examples of target morphosyntactic structures.

Set 1: Early acquisition	–ing	he's playing
	plural	two hungry cows
Set 2: Later acquisition	genitive	Mommy's hand
	copula*	donkey is hungry
	multi-verb utterances*	He came and I ate. (2 clauses) It can drive fast. (auxiliary + verb)

*These two targets were also reported in Tommerdahl and Kilpatrick, in press.

family history of language disorders or communication difficulties; and (5) no known neurological disorders.

Procedure

The participants in this study were recorded in two separate sessions within one week of each other under conditions as near to identical as possible. For each session, the participant, accompanied by either a parent or grandparent (hereafter referred to as caregiver), played in a Flexible Learning Room. The same caregiver was always present for both sessions. This room was fitted with a variety of toys, including such items as blocks, vehicles, animals, a tea set, and a toy farm. One wall included a projection of a circus scene, and the rug on the floor provided a setting of roads and a small town. The toys provided were the same for each session and children selected the toys they wanted to play with throughout the session. Interactions in the room were audio and video recorded through the use of five microphones and four cameras.

Each session lasted approximately 35 minutes. The caregivers were asked to play normally with the child during that time rather than attempting to elicit any specific linguistic forms. The one exception to this was that the caregivers were asked to bring up at some point some event that occurred in the recent past. This was thought to provide encouragement to the participants to produce a wider variety of tenses than what might be typical in a normal play session (Crystal, 1982).

Orthographic transcription of each sample in its entirety was completed by a trained Speech and Language Therapist (SLT), who also divided each sample into utterances according to P-units (Loban, 1976). Following Miller and Chapman (2004), the P-unit was limited to a maximum of two independent clauses to avoid run-on sentences. In order to assess the reliability of the transcriptions, approximately 10% of the samples, all of which had been transcribed by the SLT, were transcribed by the first author and then compared with the full transcriptions. Inter-transcriber reliability was 0.88.

A variety of items was selected for examination in this study (see Table 1). Four targets were selected based on their positioning on Brown's table of morpheme acquisition (1973). The first of these four, the plural '*s*' and the *–ing* ending, appear in Stage II, at 28–36 months. The other two, the genitive (-*s* possessive) and the copula, are Stage III morphemes, acquired at 36–42 months. As the ages of our participants range from 30 to 42 months, with a mean of 35.6 months, the Stage II morphemes are likely to have already been acquired by many of the participants. The Stage III morphemes, on the other hand, are likely to still be in the process of acquisition. Thus it might be expected that the Stage II morphemes will show more stability, and a higher degree of reliability, than the Stage III morphemes. By including Stage II and Stage III morphemes, we can then compare an array of structures with varying frequencies.

The final target, the multi-verb utterance, includes a variety of syntactic structures and was constructed as an experimental tool that could be counted on to produce a category where high frequency was predictable in order to test the hypothesis that high frequency leads to high reliability. Although not an established grammatical category of its own, it was constructed here to represent a target squarely in the category of morphosyntax, representing a certain degree of syntactic complexity, and predicted to have a relatively high frequency count. All utterances composed of more than a single verb, regardless of type, were counted. Multi-verb utterances are expected to show robust measures of reliability, and thus allow a comparison of reliability with the less-frequent Stage II and III morphemes.

In the analysis of the language samples, a minimum of 225 utterances per session were necessary. In order to give the child time to become accustomed to their surroundings, the first 25 utterances of each child were excluded from analysis, and utterances over 225 were excluded as well, providing 200 utterances per child in each session. Within these 200 utterances, target morphemes were identified and counted. If a morpheme was repeated due to fluency issues (he ran … ran into the woods), the morpheme was only counted once, but otherwise, morphemes that appeared multiple times in the same utterance were counted multiple times. No matter how many verbs they included, multi-verb utterances were only counted once.

Target morphemes were analyzed based on two different categorizations of the data: nested blocks (test-retest reliability) and independent blocks (split-half reliability). In the nested blocks analysis, target morphemes were counted in 50-utterance nested blocks, resulting in counts for 50 utterances, 100 utterances, 150 utterances, and 200 utterances. The frequency of the target structures for each child at first visit were then compared with the frequency for the second visit, in order to determine whether the frequency of that particular target was reliable across the two language samples. In the split half analysis, the first set of 100 utterances were compared directly with the second set of 100 utterances taken at the same time, resulting in a single comparison for Time 1 and a single comparison for Time 2.

We take the Intraclass Correlation Coefficient as our measure of interest for both analyses in determining the absolute reliability between the two samples. If target items are produced at the same frequency in compared samples, the correlation coefficient will be high, while varying frequencies in the compared samples will result in lower degrees of correlation. This value ranges from 0 to 1, with a correlation of 0 indicating a complete lack of similarity, and a correlation of 1 meaning that the samples are identical. Taking into consideration the work discussed in the preceding section, we take as an acceptable degree of correlation a significant value of $r > .6$.

Results

As shown in Table 2, the frequency of the different targets varied greatly. Multi-verb utterances were the most common, followed by the use of *be* as a copula. Plural and –ing

Table 2. Mean frequency of each structure in samples 1 and 2 for all lengths of samples with standard deviation given in parentheses, listed in order of frequency at 200 utterances.

	50	100	150	200
Multi-verb 2	11.0 (5.8)	21.8 (10.6)	32.7 (15.2)	44.3 (19.7)
Multi-verb 1	7.9 (4.3)	17.4 (8.3)	26.7 (10.8)	37.2 (15.2)
Copula 1	6.7 (4.5)	12.5 (7.2)	17.8 (10.4)	24.0 (13.0)
Copula 2	5.7 (3.7)	11.3 (6.5)	15.9 (8.3)	20.1 (9.4)
Plural 1	3.1 (2.7)	6.4 (2.9)	9.1 (3.6)	12.3 (4.8)
Plural 2	2.5 (2.1)	5.7 (2.9)	8.7 (4.3)	12.1 (5.0)
–ing 2	2.9 (2.8)	5.8 (5.4)	8.7 (6.8)	11.0 (7.4)
–ing 1	2.1 (2.1)	4.5 (4.3)	7.4 (5.5)	9.8 (6.2)
Genitive 1	0.22 (.60)	0.52 (.79)	0.91 (1.2)	1.1 (1.3)
Genitive 2	0.17 (.39)	0.43 (.89)	0.61 (.94)	0.74 (.96)

Table 3. Correlation of samples 1 and 2 for all children for four lengths of samples, listed in order of degree of correlation at 200 utterances.

	50	100	150	200
Multi-verb	0.56**	0.70**	0.70**	0.73**
Copula	0.32	0.64**	0.70**	0.63**
Plural	0.22	0.24	0.35	0.49**
–ing	0.26	0.39*	0.52**	0.47*
Genitive	0.21	0.12	0.01	−.05

*$p \le 0.05$.
**$p \le 0.01$.

were the next most common, but both were roughly a third to one half as frequent as the copula. Past tense and genitive were the least common, with few instances even at 200 utterances.

Table 3 shows the correlation index for frequency of use for different structures at Time 1 and Time 2. With the exception of multi-verb and copula constructions, correlations are generally below .5 regardless of the number of utterances in the sample. While –ing and plural both have significant correlations, these correlations fall below .5 for the most part. Only the most frequent structures of multi-verb and copula reach a correlation of .6, and both of these reach this degree of correlation at just 100 utterances.

Because analysis of nested samples may mean that some anomaly early in a block affects the rest of the sample, analysis of independent blocks of 100 utterances to show split-half reliability was also performed in order to determine whether the same results would hold. In this analysis, the first 100 utterances of Time 1 were compared with the last 100 utterances of the same session, and the same was done for Time 2, providing two additional comparisons of 100 utterance blocks. Table 4 shows the frequency with which the target structures appeared in the four different 100-utterance blocks analyzed.

As shown in Table 5, results for split-half reliability are similar to the test-retest results, with multi-verb and copula constructions obtaining correlations above .6. Results

Table 4. Frequency for 100-utterance blocks, ordered by frequency.

	1st 100,T1	2nd 100,T1	1st 100,T2	2nd 100,T2
Multi-verb	17.4 (8.2)	19.8 (8.2)	21.8 (10.4)	22.5 (10.5)
Copula	12.5 (7.1)	11.6 (6.6)	11.3 (6.4)	8.9 (4.5)
Plural	6.4 (2.9)	5.9 (2.9)	5.7 (2.8)	6.4 (3.4)
–ing	4.5 (4.2)	5.3 (2.6)	5.8 (5.3)	5.3 (3.7)
Genitive	.52 (.77)	.57 (.88)	.43 (.88)	.30 (.55)

Table 5. Correlations for split-half reliability at Times 1 and 2, ordered by degree of correlation.

	T1	T2
Multi-verb	.63**	.72**
Copula	.73**	.36
–ing	.47*	.27
Plural	.33	.27
Genitive	.15	.17

*$p \leq .05$.
**$p \leq .01$.

are also similar for –ing, plural, and genitive, where either significance is not reached or the correlations fall below .5.

The results regarding the use of the copula require further explanation. In the comparison of the samples taken at different times, the correlation at 100 utterances and above is both significant and robust, as is the split-half comparison for Time 1. However, the split-half comparison for Time 2 is neither significant nor robust. It appears that while the frequency of use of the copula overall appears to be reliable, its use is not always evenly distributed across a sample. In other words, at Time 2, participants tended to cluster their use of the copula more into one part of the session rather than across the session as a whole, while at Time 1, the use of the copula across the session was more evenly distributed.

Discussion and implications for further research

It is clear from these results that samples collected in near-identical conditions do not necessarily produce similar frequencies of language structures. The degree of reliability appears to be determined by a combination of frequency and length of sample, though the two are obviously related. Frequent structures, such as multi-verb utterances and the use of *be* as a copula, are more reliable than less frequent structures overall, reaching significant correlations with smaller sample sizes. As sample size increases, the degree of correlation for these structures becomes more robust.

A primary finding of this study is that the longer samples of 200 utterances did not provide a significantly greater degree of reliability than shorter samples of 100 utterances. Of the lengths tested, a sample size of 100 utterances appears to be the optimal size required for the specific morphosyntactic features examined. Shorter samples do not allow enough frequency of items for a reliable analysis, and samples of 150 and 200 utterances do not indicate that reliability increases significantly with the larger sample size. While it is possible that samples longer than 200 utterances would produce better reliability, the amount of time and effort involved in such extensive sampling is likely to make the use of such samples unrealistic for those carrying out testing. Because sample sizes beyond 100 utterances do not consistently reach higher correlations, collection of larger samples may simply lead to diminishing returns due to the time involved in collection and transcription. However, for the more frequent items, a sample of 100 utterances was sufficient to provide a robust and significant correlation; in other words, these highly frequent items are more likely to be reliable with a sample of only 100 utterances.

A second finding is that relatively infrequent grammatical structures may not show up even in a sample of 200 utterances despite the fact that the child is in fact able to use it. For instance, the genitive is one of the least frequent, and least reliable, of the items tested here. Some children did not produce this structure at all throughout an entire 200-utterance sample. However, this is not necessarily a reliable indicator of a child's language use or knowledge. While six children did not use the genitive at all in either sample, an additional five did not use it in their first sample, but they did use it in their second. A plausible explanation might be that the children had acquired the genitive in the week between their samples. However, this explanation does not account for another five children, all of whom used the genitive during their first language sample, but not during their second. It appears that with very low-frequency items, usage across different samples does not appear reliable even when the sample size includes 200 utterances. In addition, non-use of a low-frequency target item, even in a 200-utterances sample, may not indicate that a child does not control a particular structure, but simply that they did not produce it during that sample.

A third finding is that reliability may not be predictable from acquisitional stages. In other words, it is not true that earlier acquired morphemes are used more reliably than more recently acquired items. The four morphosyntactic targets were chosen based on Brown's stages of child language acquisition. The copula and the genitive were selected based on their placement as Stage III morphemes which might not be fully acquired, while –ing and plural were chosen as earlier acquired Stage II morphemes. However, the results for the copula were quite different from the other three items. This might be due simply to the frequency of production; the copula was highly frequent in the samples in comparison with –ing, plural, and genitive constructions. Because higher frequency may lead to higher reliability, it may be that the frequency with which the copula is used is simply high enough to obtain a significant correlation even when more stable, but less frequent, structures do not. However, there is an alternate analysis evident from the work of Heilmann et al. (2008), who show that productions of English language learners are more reliable in the L2 than in the L1. In this study, 241 ELLs from kindergarten to third grade produced narratives in English and Spanish through story retell of a picture book. The retest was carried out within two months. Results showed that the narratives produced in the additional language were much more reliable than those produced in the native language according to four measures. For each area tested, the

English correlation measure is followed by the Spanish: MLU (.65, .37), NDW (.79, .59), Number of Total Words (.70, .57), and Words Per Minute (WPM) (.74, .62). This study indicates that there is less variability in the children's usage of the L2 (English) than in the use of the L1 (Spanish).

Taking the Heilmann et al. (2008) results into account, it may be the case that structures in the process of being acquired, such as the genitive, are used with less variability than those that have already been acquired, at least to some degree. Recall that the copula was used with comparable frequency at different sessions but was not used consistently across a single sample. These results with the copula raise several questions that we leave as issues for further study that could provide meaningful data regarding the extent to which structures undergoing acquisition are more or less reliable than structures already acquired with comparable frequency.

Given these results, the type of structure to be examined in a spontaneous language sample should be carefully considered. Only the most frequent structures are likely to appear often enough to provide a reliable picture of a child's actual language use using sample lengths of 200 or fewer utterances. In the structures we examined, only *multi-verb* and the use of *be* as a copula appeared with enough frequency to reach a degree of correlation over .6, and both of these structures appeared at least 10 times on average in the 100-utterance sample. Items that are known to be infrequent are unlikely to indicate the child's true state of proficiency and production with these forms. For researchers working with clinical populations who may produce particular morphosyntactic items with even less frequency, overall frequency of usage by the relevant child is an important consideration.

One potential weakness of the study that could affect regularity of production was the difficulty in creating conditions in the two recording sessions for each child that were as close to identical as possible. Although the children were in the same surroundings with the same caregiver for approximately the same amount of time and with the same instructions, it is of course impossible to control for all that could affect both the child and the caregiver such as recent experiences, mood, health, and so on. Furthermore, although the same toys were available to the children during both sessions, different toy selections could be made for different sessions which could bring about different types of language. For example, it is possible that symbolic play which might be prompted from playing with toy foods could bring about linguistic forms that differ from those elicited through building with blocks. In this study, rather than limiting the toys to ensure they played with the same ones in both sessions, we chose to supply a larger number of toys to maintain interest and to promote the likelihood of recording a sample of at least 200 utterances.

While it is certain that spontaneous language samples have a role to play in the analysis of language, it is important to know their strengths and weaknesses and how they can best be used with different populations. For example, in the case of SLI diagnosis, it may be true that certain morphemes are underrepresented in children with the impairment, but at the same time, extremely long language samples would probably be required in order to have any real diagnostic reliability. Of course, the limitations of spontaneous samples listed above become much less important if it becomes possible to quickly and accurately transcribe and label much larger spontaneous samples than those discussed above.

This would require computerized systems, some of which have been developed such as CHAT for transcription (MacWhinney, 2000), CLAN for structure identification (MacWhinney, 2000), and their extensions such as POST (Parisse & le Normand, 2000; Parisse, Maillart, & Tommerdahl, 2012), MEGRASP (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2007, 2010), and GCS (Curtiss, MacSwan, Schaeffer, Kurel, & Sano, 2004). We foresee that widespread use of such equipment will revolutionize language testing in several domains.

To increase our ability to accurately assess children's use of morphosyntactic structures, it is crucial that we understand the degree to which different samples are reliable. Furthermore, to carry out language testing in clinical groups where language use may be impaired, it is necessary to have baseline information regarding typically developing children. This paper examines just one age group, a small group of target structures and a single category of children. With more work using different age groups and an increased number of target structures, an even greater understanding of the reliability of language samples can be reached.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Funding

This work was supported by the Nuffield Foundation Grant [SGS/35006].

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Coles Publishing.

Barlow, M.C. (1969). Temporal reliability of length – complexity index. *Journal of Communication Disorders*, *2*(3), 241–251.

Boersma, P., & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, *32*, 45–86.

Brown, R. (1973). *A first language: The early stages*. Oxford, UK: Harvard University Press.

Bybee, J. (1995). Regular morphology and the lexicon. *Language & Cognitive Processes*, *10*, 425–455.

Cole, K. N., Mills, P. E., & Dale, P. S. (1989). Examination of test-retest and split-half reliability for measures derived from language samples of young handicapped children. *Language, Speech and Hearing Services in Schools*, *20*, 259–268.

Cordes, A. (1994). The reliability of observational data: I. Theories and methods for Speech-Language Pathology. *Journal of Speech and Hearing Research*, *37*, 264–278.

Crystal, D. (1982). *Profiling linguistic disability*. London: Edward Arnold.

Crystal, D., Fletcher, P., & Garman, M. (1989). *Grammatical analysis of language disability* (2nd ed.). London: Cole & Whurr.

Curtiss, S., MacSwan, J., Schaeffer, J., Kurel, M., & Sano, T. (2004). GCS: A grammatical coding system for natural language data. *Behavior Research Methods, Instruments, and Computers*, *36*(3), 459–480.

Davidson, F. (2000). The language tester's statistical toolbox. *System*, *28*(4), 605–617.

Evans, J. L., & Craig, H. K. (1992). Language sample collection and analysis: Interview compared to freeplay assessment contexts. *Journal of Speech and Hearing Research*, *35*, 343–353.

Fisher, M. S. (1934). *Language patterns of pre-school children*. New York: Columbia University.

Gavin, W. J., & Giles, L. (1996). Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech, Language and Hearing Research*, *39*, 1258–1262.

Heilmann, J., Miller, J. F., Iglesias, A., Fabiano-Smith, L., Nockerts, A., & Andriacchi, K. D. (2008). Narrative transcription accuracy and reliability in two languages. *Topics in Language Disorders*, *28*(2), 178–188.

Heilmann, J., Nockerts, A., & Miller, J. (2010). Language sampling: Does the length of the transcript matter? *Language, Speech, and Hearing Services in Schools*, *41*, 393–404.

Hernandez-Lloreda, M. V., & Colmenares, F. (2006). The utility of generalizability theory in the study of animal behaviour. *Animal Behaviour*, *71*, 983–988.

Hux, K., Morris-Friehe, M., & Sanger, D. D. (1993). Language sampling practices. A survey of nine states. *Language, Speech and Hearing Services in Schools*, *24*, 84–91.

Johnson, M. R., & Tomblin, J. B. (1975). The reliability of developmental sentence scoring as a function of sample size. *Journal of Speech and Hearing Research*, *18*, 372–380.

Kaufman, A. B., & Rosenthal, R. (2009). Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. *Animal Behaviour*, *78*, 1487–1491.

Kemp, K., & Klee, T. (1997). Clinical language sampling practices: Results of a survey of speech-language pathologists in the United States. *Child Language Teaching and Therapy*, *13*, 161–176.

Kline, T. (2005). *Psychological Testing: A practical approach to design and evaluation*. Thousand Oaks, CA: SAGE Publications.

Loban, W. (1976). *Language development: Kindergarten through grade twelve*. Urbana, IL: National Council of Teachers of English.

Lord, F. M., & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahway, NJ: Lawrence Erlbaum Associates.

MacWhinney, B., & Snow, C. (1990). The Child Language Data Exchange System: An update. *Journal of Child Language*, *17*(2), 457–472.

Marinellie, S. A. (2004). Complex syntax used by school-age children with specific language impairment (SLI) in child-adult conversation. *Journal of Communication Disorders*, *37*, 517–533.

Miller, J. F., & Chapman, R. S. (2004). Systematic analysis of language transcripts (SALT, v8.0) [Computer software]. Madison, WI: Language Analysis Laboratory. Waisman Center, University of Wisconsin-Madison.

Minifie, F., Darley, F., & Sherman, D. (1963). Temporal reliability of seven language measures. *Journal of Speech and Hearing Research*, *24*, 154–161.

Muma, J. R. (1998). *Effective speech-language pathology: A cognitive socialization approach*. Mahwah, NJ: Lawrence Erlbaum.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*, 1–18.

Parisse, C., & le Normand, M. T. (2000). Automatic disambiguation of morphosyntax in spoken language corpora. *Behavior Research Methods*, *32*(3), 468–481.

Parisse, C., Maillart, C., & Tommerdahl, J. (2012). F-LARSP: A computerized tool for measuring morphosyntactic abilities in French. In M. Ball, D. Crystal & P. Fletcher (Eds.), Assessing grammar: The languages of LARSP (pp. 230–244). Bristol: Multilingual Matters.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, *37*, 705–729.

Scarsellone, J. (1998). Analysis of observational data in speech and language research using Generalizability Theory. *Journal of Speech, Language, and Hearing Research*, *41*(6), 1341–1347.

Shrout, P., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Southwood, F., & Russell, A. F. (2004). Comparison of conversation, freeplay, and story generation as methods of language sample elicitation. *Journal of Speech, Language, and Hearing Research*, *47*(2), 366–376.

Stalnaker, L. D., & Creaghead, N. A. (1982). An examination of language samples obtained under three experimental conditions. *Language, Speech and Hearing Services in Schools*, *13*, 121–128.

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use* (3rd ed.). Oxford: Oxford University Press.

Tilstra, J., & McMaster, K. (2007). Productivity, fluency, and grammaticality measures from narratives: Potential indicators of language proficiency? *Communication Disorders Quarterly*, *29*(1), 43–53.

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, *31*, 101–121.

Tommerdahl, J. M., & Kipatrick, C.D. (2013). Analysing frequency and temporal reliability of children's morphosyntactic production in spontaneous language samples of varying lengths. *Child Language Teaching and Therapy*, 29(2): 169–181.