Check for updates

CHILD LANGUAGE TEACHING & THERAPY

Child Language Teaching and Therapy
29(2) 171–183
© The Author(s) 2013
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265659012459528
clt.sagepub.com
SAGE

Analysing frequency and temporal reliability of children's morphosyntactic production in spontaneous language samples of varying lengths

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

Abstract

Spontaneous language samples are a useful tool in studying the nature of morphosyntactic production in both clinical and academic settings. However, little work has been done on typical frequency of occurrence of morpheme production and even less on the degree of temporal reliability of morpheme frequency between samples. This study compares two language samples recorded within a week of each other under similar conditions in a test–retest paradigm for each of 23 typically developing children aged 2;6–3;6 to examine the production of the following morphemes associated with specific language impairment: contracted auxiliary, copula, *do* as an uncontracted auxiliary, *be* as an uncontracted auxiliary, third person singular and past tense. Also analyzed were all utterances containing more than one verb. Frequency and reliability of four different sample lengths ranging from 50 to 200 utterances were tested. Results showed generally low degrees of reliability at 50 and 100 utterances, with only limited improvement for most targets at 200 utterances. Clinical implications of these results are discussed.

Keywords

Frequency, language sample, reliability, spontaneous language sample

1 Introduction

This study uses spontaneous language samples to examine frequency and regularity of morphosyntactic production. Using a test–retest paradigm, we consider production frequency of morphosyntactic structures, temporal reliability between samples, and the extent to which the examination of

Corresponding author:
Jodi Tommerdahl, Department of Curriculum and Instruction, University of Texas at Arlington, 701 Planetarium Place, Arlington, TX 76019, USA.
Email: joditom@uta.edu

samples of varying lengths may lead the clinician or teacher to incorrect conclusions regarding the presence or absence of a particular structure in a child's linguistic repertory.

The use of frequency counts has an established importance in both clinical use (Cole et al., 1989; Gavin and Giles, 1996; Johnson and Tomblin, 1975; Leadholm and Miller, 1992; Shewan and Henderson, 1988) as well as language acquisition and processing research (for discussion, see Roland et al., 2007), but frequency counts may not be consistent enough to warrant firm conclusions about an individual's use of a particular morphosyntactic item. Although certain linguistic measures such as mean length of utterance (MLU) may possess high stability (Casby, 2011; Minifie et al., 1963; Rice et al., 2010; Rondal et al., 1987), even when taken from a sample as short as 50 utterances, the necessary sample size to show adequate representation of a child's morphosyntax is unknown. It is possible that two samples taken of a single child's language a few days apart may lead to different pictures of that child's capabilities even if the environment is kept the same. In the present study, we examine a small group of morphosyntactic items, focusing on their frequency of production and their temporal reliability, or the degree to which two samples taken in similar conditions at different times produce the same results.

Possibly the best example of using frequency counts of morphosyntactic items for evaluating potential language problems is the Language Assessment Remediation and Screening Procedure (LARSP; Crystal et al., 1976, 1989). The LARSP requires the administrator to collect a spontaneous language sample and count the frequency of a variety of morphosyntactic structures. The completed profile shows the level of the child's production in comparison to the norms for the child's chronological age. The LARSP profile may be used to help diagnose, develop therapy and monitor therapeutic progress (Crystal et al., 1976, 1989; Crystal and Fletcher, 1979; Maillart et al., 2011).

Frequency counts in language assessment are also evident, although more subtly, in language profiles and checklists used to monitor child language development. These tools often depend at least partially upon parental report of the child's ability to produce certain structures. For example, the MacArthur Communicative Developmental Inventories (Fenson et al., 1994) require parental report of whether or not their child uses grammatical structures of past and future tense, and the online progress checker (I CAN, 2010) asks parents whether children use action words and specific grammatical morphemes such as the plural. Children who rarely produce structures such as the past tense are arguably less likely to have their ability noted and reported to the clinician than their counterparts who produce the structure more often. In these cases, it is possible that no distinction is made between low and high frequency production, but instead low production is reported as 'no production'.

The most extensive work carried out to date on structure frequency in child language is a dataset published by the Wisconsin Department of Public Instruction (Leadholm and Miller, 1992), explicitly stating its intention to provide these norms in order to better be able to diagnose impairments in expressive language (1992: ix). They provide a database of language from 266 children in the local school community, ranging from 3–13-years-old, presenting frequency norms for five structures in the category of 'bound morpheme' (past, plural, possessive, third person singular and present progressive), and several other variables. In a similar vein, Shewan and Henderson (1988) gathered frequency data from language samples in a healthy older adult population to establish normative data to help distinguish between normal aging effects and communication impairments.

Hewitt et al. (2005) went further in the quest to directly compare frequency data in clinical versus normal populations by comparing samples of typically developing children and children with specific language impairment (SLI) for MLU, IPSyn score (Index of Productive Syntax;

Scarborough, 1990) and number of different words (NDW) in 50 utterance samples. Significant differences were found in MLU, NDW and the 'sentences score' section of the IPSyn, showing promise for the potential for development of diagnostic tools based on the provision of language norms. They write, 'If more normative information were to become available, these measures might show promise in providing norm-referenced yet ecologically valid means of identifying impairment' (Hewitt et al., 2005: 205).

A potentially overwhelming difficulty exists in providing production norms for all aspects of language. Not only would phonology, morphology, syntax and semantics need to be taken into consideration, but so would different languages and different age ranges. Norms for clinical groups as well as typically developing children would need to be developed (Lindsay, 2011). While a complete catalog of production norms would be extremely useful, it would also be difficult.

However, with the specific endpoint in mind of providing better diagnostic tools for identifying young children with SLI, this infinitely long list quickly shortens. If differences between the language production of typically developing children and those with language difficulties have been identified or suggested, those differences provide us with clues as to what areas are of interest for early exploration. Studies have repeatedly shown that children with SLI perform differently than other children in their morphological development (Cleave and Rice, 1997; Leonard, 1998; Leonard et al., 1997; Rice and Wexler, 1996). Whereas typically developing children move through a stage coined the Optional Infinitive Stage (Wexler, 1994), children with SLI seem to remain in this stage for a much longer time than what is typical. During this phase, children seem to understand at least some of the grammatical properties of finiteness, but they consider the use of tense and agreement to be optional and, for example, ignore the rule that tense marking is obligatory in a main clause (Rice et al., 1995). The marking of finiteness can be shown both in the use of lexical verbs as well as the auxiliaries *be* and *do* and the copula. Rice et al. (1995) report that *be* and *do* are often omitted in the language of children with SLI. Our knowledge of specific morphosyntactic difficulties in children with SLI – combined with calls from several researchers asking for further normative information – have together led us to the first part of our study, which provides production norms for several morphemes shown to be related to SLI. Most of the items targeted in this study are those that have been put forward as possible markers of SLI. Although work exists that examines morpheme use of children with SLI in elicited contexts, pure frequency counts of these morphemes that compare clinical and typically developing groups are currently impossible due to lack of frequency norms.

In addition to providing frequency counts of a number of linguistic structures, this study also examines the reliability of production of these structures in spontaneous language samples, which Gallagher refers to as 'the centrepiece of child language assessment' (Gallagher, 1993: 2). Like any clinical test of language, frequency counts from a language sample that is being used to examine a child's typical language are only useful if that sample is truly representative of the person's typical language production; or in other words, on the reliability of the sample. Until the question of temporal reliability is answered, it cannot be known to what extent frequency information from a sample is of clinical value.

At the moment, relatively little is known about the temporal reliability of most aspects of language production. This is an acknowledged methodological problem in the use of spontaneous language samples. Cole et al. recognize this difficulty, stating 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 samples interpretation' (1989: 260).

More recently, Marinellie, discussing the temporal reliability of language samples in regard to syntactic structures, declares it to be 'a relatively uncharted area' (2004: 520). This lack of information also denies the clinician important information, such as the likelihood of a given item's appearance in a sample and the degree of reliability of a sample of a certain length. Without baseline information regarding production and reliability norms, the value of spontaneous sample use for helping to differentiate between typical and impaired language is minimized.

These questions of reliability also have direct implications for the related question of how long a sample must be for clinical use. For example, when looking at a child's use of verb tenses, will a 200-utterance sample provide a much more representative sample than one made up of 50 or 100 utterances? It is reasonable to assume that a longer sample has better reliability, but the question remains of whether the amount of increased reliability is worth the amount of time that the additional data gathering, transcription and analyses require.

A small amount of research on the reliability of repeated language samples has been carried out. Both production counts and temporal reliability were observed in a well-constructed study by Gavin and Giles (1996) which studied four linguistic measures: mean length of utterance in morphemes (MLU-m), total number of words (TNW), mean sentence length (MSL) and the NDW in samples ranging from 25 to 175 utterances. Reliability for MLU-m and MSL were found to be at or above what the authors proposed as the acceptable level of $r \geq 0.71$ by 75 utterances, while NDW reached this level at 100 utterances and TNW never reached adequate reliability.

An earlier related study (Minifie et al., 1963) compared the reliability of items such as mean length of response and NDW between typically developing five- and eight-year-olds using elicited samples of 50 utterances. Frequency counts were made for items such as mean length of response (MLR), NDW and type–token ratios. Results showed that some items such as NDW had the same reliability coefficient ($r = 0.65$) for both ages, while others such as MLR varied, with $r = .82$ for the five-year-olds and $r = .77$ for the eight-year-olds. Despite the limitation due to the short samples, the study highlights the fact that reliability measures vary according to age.

Cole et al. (1989) used spontaneous language samples of 10 children with developmental delays to compare the reliability of test–retest and split-half measures of language production. Twenty minutes of conversation with a minimum of 100 utterances were recorded for each child. Overall, the test–retest samples showed a lower degree of reliability than split-half measures. This study differs from most others in that it included specific lexical and grammatical items (not listed in the article) which were judged by the presence or absence of each item in a sample. In 73% of cases the use of the item was consistent, either being present or absent in both samples. Specifically, out of 250 possibilities for the items to appear (10 participants × 25 items), 35 forms were used in only the first sample and 33 forms were used in only the second. The article concludes that 'additional information about the reliability of measures derived from samples is needed' (1989: 267).

Johnson and Tomblin (1975) tested the reliability of the Developmental Sentence Scoring (DSS; Lee, 1974), a procedure for quantifying children's grammatical production through both elicited and spontaneous samples, on 50 preschool children who fulfilled the criteria of being monolingual and having normal hearing. Reliability was determined for sample lengths ranging from 25–250 utterances. Results showed that reliability generally increased as sample sizes were longer, but that a great deal of variation existed between different grammatical items.

Although a small number of studies have examined reliability in repeated language samples, most of this work has focused on global characteristics of the sample such as MLU, NDW, MSL and TNW. Production of specific morphemes in samples has been examined up to the level of providing frequency information (Leadholm and Miller, 1992) and with respect to the reliability regarding presence versus absence (Cole et al., 1989). However, no data exists for temporal reliability of the production of specific morphosyntactic structures in typically developing children. This study attempts to begin to fill that gap.

In this article we therefore address the issue of temporal reliability of morphosyntactic production in spontaneous language samples from typically developing children using the general methodological framework put forward by Gavin and Giles (1996). This includes the collection of two language samples from preschool children in a naturalistic setting within a short timeframe for a test–retest procedure, and the division of the samples into different sizes in order to compare reliability. The major difference between the studies is the linguistic behavior being analyzed. While Gavin and Giles measured total number of words, number of different words and MLU, our study hones in on specific morphosyntactic items. Our experiment has three primary aims:

(a) to provide frequency counts of specific morphosyntactic structures at 50, 100, 150 and 200 utterances;
(b) to determine the reliability of each morphosyntactic structure across the two samples; reliability will be determined for the sample sizes of 50, 100, 150 and 200 utterances;
(c) to determine how often morphemes that are used at least once by a child over the two samples, and are therefore present in the child's productive repertory, do not appear in smaller samples.

A secondary aim of the study will be to determine to what degree a higher frequency of occurrence of a linguistic structure will lead to greater reliability.

II Methods

1 Participants

Twenty-seven children participated in this study. Ethical approval was given by the University of Birmingham Ethical Review Committee. Four were excluded from our results due to a failure to produce the minimum 200 utterances required in each of two sessions. The remaining children consisted of 13 females and 10 males, ranging in age from 2;6–3;6 (mean = 35.6 months), which is consistent with the group number of Leadholm and Miller (1992). Inclusion criteria were similar to those of Johnson and Tomblin (1975), which required the child to be monolingual and of normal hearing ability. The following additional criteria were gained through parental report:

1. The child had not been referred to speech and language therapy.
2. The parents did not feel that the child began using language later than his or her peers.
3. No one in the immediate family had been suspected of having language or communication difficulties.
4. The parents did not suspect that the child had language or communication difficulties; and
5. The child had no known neurological disorders.

Participants were recruited from the local community of a large city in central England through a variety of playgroups in a spread of socio-economic areas.

2 Procedure

Two spontaneous language samples were recorded for each participant, with recordings taking place no more than one week apart. The sessions were controlled as much as possible (Muskett et al., 2012) to have nearly identical contexts. Recording took place in a Flexible Learning Room fitted with four hidden cameras, five microphones, and a host of media systems controlled from an adjacent gallery. A large selection of toys was available for play, including a model farm, several toy animals, building blocks, vehicles, a tea set, and a large rug with pictures of roads and a local village. In addition, a circus scene was projected onto a whiteboard on the wall. The same toys were available for each session and were placed in a box before each child's arrival.

For each recording, the child played in the playroom with the same caregiver (a parent or grandparent) and was recorded for approximately 35 minutes. No scripted elicitation took place, but before beginning the recording, caregivers were asked to attempt at some point to initiate a conversation about something that had happened in the recent past in order to encourage the children to use a wider variety of verb tenses than what they might use during the normal course of playing with toys (Crystal, 1982). Otherwise, the caregivers were asked to carry on a normal conversation with the child as they would while playing together at home.

Transcription of each sample was completed by a trained speech and language therapist, who transcribed each language sample orthographically in its entirety and divided each sample into utterances according to the CHAT Transcription Format (MacWhinney and Snow, 1990). Each appropriately used morpheme was counted by a trained graduate student in linguistics. In order to assess the accuracy of the transcriptions and counts, approximately 10% of the samples were transcribed by the first author and compared with the original transcriber's work to achieve an agreement value of 0.88. Accuracy checks resulted in a score of 94%.

For this study, a group of morphosyntactic items was selected for examination based on their potential as markers of SLI (Cleave and Rice, 1997; Leonard, 1998; Leonard et al., 1997; Rice et al., 1995; Rice and Wexler, 1996). While we do not include children with SLI in this experiment, we chose items that will hopefully be of use for future studies comparing the production of children with SLI to the children in this study. The items include the copula, third person singular (3s), *do* as an uncontracted auxiliary (*do*-aux), *be* as an uncontracted auxiliary (*be*-aux), contracted auxiliary (*'*-aux), and past tense (*-ed*). In order to allow inclusion of regularized forms such as *runned*, both regular and irregular forms are collapsed under the umbrella of 'past tense'. Additionally, we constructed a category that we call 'multiverb', which consists of any utterance possessing more than one verb, whether it be an auxiliary–lexical verb combination or independent lexical verbs. The category of multiverb includes a variety of syntactic structures and was constructed not as a potential clinical tool, but 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 consistently leads to high reliability. This would be especially useful if production of most SLI-related items were low or all within a limited range. By including both targets associated with SLI along with others that are expected to be more frequent, we can then compare an array of structures with hopefully varying frequencies. Table 1 provides an example of each item.

In the sample analysis, the first 25 utterances of each child were omitted to allow the child time to become accustomed to their surroundings, except in a single case where fewer than 225 utterances were produced. The target items were then identified and counted for their usage in the next 200 utterances. Any utterances beyond 225 were not included in the analysis. In identification and counting, multiverb utterances were counted only once regardless of the number of verbs present, but for other target structures, morphemes that appeared more than once in the same utterance were counted multiple times. However, if a child repeated a morpheme due to imperfect fluency, the morpheme was only counted once. Pure frequency counts, as opposed to items produced in obligatory contexts, were used in order to provide information about distributional patterns. As Pica states:

Table 1. List and examples of target morphosyntactic structures.

copula	donkey is hungry
past tense -ed	he walked; he ran
third person singular	daddy eats cake
do-aux: uncontracted	I do like it
be-aux: uncontracted	they are going
contracted aux	it's chasing the cow
multiverb utterances	he came and I ate (two clauses); it can drive fast (aux + verb)

> analysis of a morpheme based on its suppliance in obligatory contexts reveals only how well a participant can produce the morpheme in a required linguistic environment, but such analysis does not indicate whether the participant has also acquired appropriate distributional patterns for the morpheme. (Pica, 1983: 70–71)

Furthermore, some items such as contracted forms have no obligatory contexts at all. The use of frequency counts thus allowed consistency of counting across all items. Target morphemes were analysed based on sample lengths of 50, 100, 150 and 200 utterance blocks with each smaller block being nested within the larger one. The mean frequencies of the target items from the first sample were compared with those of the second sample, to determine if the frequency of that particular item was reliable across the samples. This was done for each sample length. For example, the first set of 50 utterances from the first recording was compared directly with the first set of 50 utterances from the second recording, and this was repeated for all lengths.

For our analyses, our measure of interest is the Intraclass Correlation Coefficient, which was used to determine absolute reliability between the two samples. Correlation is used to determine the degree of similarity between two groups, with a numerical range of −1.0 to 1.0. In this study, if a morphosyntactic item is produced with similar frequency in both samples, the correlation coefficient will be high, while a large discrepancy in production will result in lower degrees of correlation. Gavin and Giles (1996) suggest an ideal reliability coefficient of 0.90 or higher, but concluded that a minimal coefficient of 0.71 could be considered acceptable given that at this level, less than 50% of the variability could be due to measurement error. This level will be used for the current study.

III Results

Our first aim was to provide production counts. As shown in Table 2, frequencies of the different items vary greatly. As predicted, multiverb utterances were the most common structure. This was followed by the copula, which was reasonably plentiful despite its lag behind the multiverb. Production levels of the other morphemes were relatively low as shown here, with *be*-aux being the least frequent with few instances even at 200 utterances.

Table 3 addresses our second aim and shows the correlation index for frequency of use for each structure at Times 1 and 2. As shown, correlations are generally below 0.71 regardless of the number of utterances in the sample. The only items that achieve correlations of 0.71 or higher are multiverb and contracted auxiliary, which both rank in the top three according to frequency. While both copula and multiverb fail to meet the suggested correlation level at 150 utterances, they are both exceedingly close.

Table 2. Mean frequency of production.

	50	100	150	200
multiverb 2	11.0 (5.8)	21.8 (10.6)	32.7 (15.2)	44.3 (19.7)
multiverb 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)
’-aux 1	1.8 (1.6)	3.9 (3.7)	6.1 (5.2)	9.0 (8.3)
’-aux 2	2.4 (3.1)	4.5 (5.0)	6.1 (6.0)	8.3 (7.0)
do-aux 2	1.3 (1.4)	2.5 (1.9)	3.3 (2.1)	5.3 (3.2)
-ed 1	0.8 (1.3)	2.1 (2.9)	3.3 (3.9)	4.5 (5.6)
do-aux 1	0.8 (1.1)	1.9 (1.7)	2.9 (2.1)	4.4 (2.9)
-ed 2	0.7 (1.1)	1.6 (2.1)	2.8 (3.4)	3.6 (3.8)
3s 2	1.2 (1.3)	1.8 (1.9)	2.1 (2.0)	3.0 (2.5)
3s 1	0.8 (1.3)	1.2 (1.3)	1.6 (1.7)	2.4 (2.7)
be-aux 1	0.4 (0.9)	1.1 (2.4)	1.7 (3.2)	1.8 (3.2)
be-aux 2	0.4 (0.8)	0.9 (1.9)	1.3 (2.3)	1.4 (2.4)

Note: Mean frequency of production 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.

Table 3. Correlation of samples 1 and 2.

	50	100	150	200
’-aux	0.29	0.57**	0.76**	0.78**
multiverb	0.56**	0.70**	0.70**	0.73**
be-aux	0.34	0.74**	0.69**	0.66**
copula	0.32	0.64**	0.70**	0.63**
do-aux	0.27	0.21	0.41*	0.52**
3s	0.16	0.22	0.36*	0.38*
-ed	0.11	–0.01	0.32	0.32

Notes: * $p < .05$; ** $p < .01$; correlation of samples 1 and 2 for all children for four lengths of samples, listed in descending order at 200 utterances.

The third aim of this study concerns how often a smaller sample indicates that a particular linguistic item is absent when the item does actually exist within a larger sample. Although clinicians regularly use normed tests to examine language use in obligatory contexts, observation of language production in spontaneous samples provides evidence that is ecologically valid (Cicourel, 1996; Dunn et al., 1996; Hewitt et al., 2005; Lund and Duehart, 1993; Miller, 1996; Naremore et al., 1995; Owens, 1999). However, Cole et al. (1989) indicate that the use of different lexical and grammatical items may vary in different parts of samples, and looking at only part of a sample may not give a true picture of the variety of lexical and grammatical items a child may use in general. In 27% of the samples examined by Cole et al., children used a target in one sample but not in another.

To determine whether the children in this study use the target features in some parts of their samples but not in other parts, we compared the presence and absence of each target in different

Table 4. Number of children with frequency counts of 0 in a given sample size but >0 frequency over the entire 400 utterance sample.

	50	100	150	200
multiverb	0/23	0/23	0/23	0/23
copula	0/23	0/23	0/23	0/23
’-aux	6/23	4/23	1/23	0/23
3s	9/23	4/23	4/23	2/23
do-aux	11/23	3/23	0/23	0/23
be-aux	13/23	10/23	5/23	5/23
-ed	14/23	10/23	2/23	2/23

pieces of the sample. Table 4 presents these results, indicating how many children at each sample size had not yet used that particular morpheme, but did use it later in Session 1 or at some point in Session 2. Within the table, the numerators indicate how many children did not produce that particular morpheme at all within the sample length given. The denominator is always 23, the number of children who participated. For example, out of 23 children in the study, 14 of them showed no use of the past tense (*-ed*) in the first 50 utterances but did use the past tense in a later portion of the 400 utterance sample. This indicates that if spontaneous samples of 50 utterances had been used to determine whether these children used the past tense in spontaneous conversation, 14 out of 23 would have indicated that the child did not produce the tense. Even at a sample size of 100 utterances, 10 children out of 23 who eventually used the form had not produced it yet. At 200 utterances, two children had still not produced it, but did in the following session.

Interesting to note is not only the large degree of potential errors present in the small- and medium-sized samples, but also the relationship between frequency, reliability and error rates. This will be more fully explored in the following section.

IV Discussion

The aim of this study was to provide frequency counts of specific morphosyntactic items in samples of different lengths in a test–retest paradigm and to calculate the degree of reliability for each item at each sample length. The results presented are not only of theoretical interest regarding our knowledge of language, but of clinical interest given the widespread clinical use of spontaneous samples (Marinellie, 2004; Southwood and Russell, 2004). Results show that two samples produced by a child in near-identical conditions within a short timeframe may differ widely in terms of the frequency of specific morphological items produced. This is shown by the high standard deviations in Table 2 as well as the low degrees of reliability that populate the majority of Table 3. The average production of several structures examined in a 50-utterance sample is very low. Besides the category of multiverb that was specially constructed to represent a highly frequent item, the copula is the most frequent at 6.7 on the first recording and 5.7 on the second. Production means of all the other morphological items range from 0.4 to only 2.4.

Temporal reliability was found to fall short of Gavin and Giles’ proposed correlation of 0.71 for most of the items examined at all sample lengths. The only item to surpass this level at 100 utterances was *be*-aux, which subsequently fell below 0.71 in longer samples. At 150 utterances, only ’-aux met this criterion, although multiverb and copula were exceedingly close. Even at 200 utterances, only ’-aux and multiverb had correlations >0.71.

The possibility of comparing our findings directly to those of other researchers is limited, given that the methodology used has typically been used on different linguistic items. However, the items common to our study and that of Leadholm and Miller (1992) were both consistent. At 100 utterances, Leadholm and Miller listed past tense frequency as 2 for three-year-olds compared to our 1.6 and 2.1. Leadholm and Miller list the genitive at 1 compared to our 1.2 and 1.8. Also, no practice effects were displayed in our sample as morpheme production was just as likely to decrease as to increase between the first and second recordings. While the frequencies we provided were similar to those in Leadholm and Miller's study, we show that reliability is still fairly low. This is unlike results found in other studies, which examined more global linguistic measures. Gavin and Giles (1996) found MLU and MSL to surpass a reliability value of 0.71 by 75 utterances and NDW to do the same by 100. Minifie et al. (1963) also found that MLR surpassed this level for both five- and eight-year-olds. This may indicate that children's length of utterances and some other global measures of their productions are fairly consistent, but the complexity and type of morphemes produced may be quite variable.

Results show that a single language sample, sometimes even with 200 utterances, could be misleading regarding whether or not a child has a particular linguistic item in his or her productive repertory. This problem was especially marked at the sample lengths of 50 and 100 utterances, and some items such as the past tense were at much greater risk of not being represented than others such as the copula. A possible reason for this is that certain linguistic items tend to be produced in clusters. For example, in a play situation, the present tense may be more prevalent, but if a recent event is momentarily discussed, several productions of the past tense may be produced in consecutive utterances. This possibility is a potential methodological limitation of gathering spontaneous samples in a context that may be biased against the production of a certain type of linguistic element. It is possible that many forms of elicited language samples will be more reliable than spontaneous ones, but only for those items elicited. A study of this type would be a valuable complement to the current one.

A further potential methodological issue with the current study is the potential difference in the linguistic behaviors of the caregivers across the two sessions. For instance, caregivers may have been likely to more carefully follow the instructions regarding discussion of recent events on the first occasion or may have had different linguistic behaviors on different occasions. A large number of variables have been examined in the context of repeated language samples to determine whether a change between the first and second sample resulted in grammatical differences in the samples. Those factors found to make a significant difference include interaction style and material used (Southwood and Russell, 2004), question type (Yoder and Kaiser, 1989), task type (Hansson et al., 2000), location (Kramer et al., 1979) and elicitation method (Evans and Craig, 1992; Stalnaker and Creaghead, 1982). As several factors have been shown to affect grammatical output in child language, we attempted to keep the contexts of the first and second recordings as consistent as possible in terms of having the same caregiver, location, toys and situational context. However, other unpredictable factors, including behavior of the caregiver, will necessarily exist given the nature of human interaction.

These findings are useful for clinicians in several ways. First of all, a clinician is likely to use a spontaneous sample to assess a child's language from a perspective that is different from that offered by norm-referenced tests. The clinician uses spontaneous language to ask what type of language the child produces in natural conversation. It is often recommended that samples of 50–100 utterances are used, but even at 100 utterances we see that nearly half of the children who actually do use *be*-aux and *-ed* in a larger sample of 400 utterances had not used it in the first 100. This unevenness of use leads to the strong likelihood of error in determining whether a child has a particular morphological item in his or her repertory. When we speak of error, we are not talking about overall diagnostic error as we know that standardized tests with elicited responses are an important part of diagnosing language impairment. However, we are talking about error in the use of the spontaneous sample to tell the clinician whether a certain structure is used conversationally.

Another important issue for clinicians is knowing how long a language sample is necessary when analyzing morphological use. Results of this study have shown that different items have different degrees of reliability for children aged 2;6–3;6. For the items examined here, it seems that a sample of 100–150 utterances is the most appropriate.

This suggested length must be taken with several caveats. While the reliability of several items increases significantly between 50 and 100 utterances in some cases and between 100 and 150 in others, much smaller improvements are noticed between 150 and 200 utterances. While 200 utterances may be numerically better than 150, it would be difficult to say that the amount of time needed to record, transcribe and analyze 50 more utterances would be worth the small added amount of reliability.

There is also the question of which morphological item is being observed. The clinician looking at *be*-aux would be well advised to use a 100-utterance sample, but may well need a 150-utterance sample to examine *'*-aux. On the other hand, a clinician wanting to look for 3s or *-ed* in a spontaneous sample would see from our chart that while 200 utterances are better than 50, even then the degree of reliability only reaches 0.32. This may mean that the clinician should not use spontaneous samples to examine the use of *-ed* or 3s. We acknowledge that spontaneous language samples are the only way to get a valid picture of natural language production, but their shortcomings must be acknowledged so that they can be used in the most effective way possible. The clinician wanting to examine the past tense may need to collect a much longer or several shorter samples or may need to use elicited language. The clinician forced to use a shorter sample due to time restrictions may also look for evidence of the past tense, but will know that its absence from even a 200-utterance sample does not necessarily mean that the child does not produce it.

Another area of potential difficulty involving spontaneous language sampling is its use as a measure of effectiveness of clinical treatment. A child with impaired language may have a therapeutic goal to strengthen their use of certain morphemes. This strength would likely be equated with production frequency if spontaneous samples were used to compare performance. After all, it seems reasonable that a child having difficulty with certain morphemes would increase their use of those morphemes through therapy and that the spontaneous sample would be an appropriate tool for observing these changes. However, knowing what we now know about the variability of production in even relatively long language samples, an increase or decrease of use of a certain morpheme may well be due to chance rather than any true change in ability.

The measures provided in this article offer basic frequency and reliability information to begin to fill the gap lamented by Cole et al. and others. However, much work remains to be done in order to establish overall language norms. Additional age ranges and further morphosyntactic items remain to be tested. Further reliability data, especially for children with different developmental difficulties, will be of use for both diagnostic reasons and to monitor efficacy of treatment.

Funding

This project was carried out thanks to Nuffield Foundation Grant SGS/35006.

References

Casby M (2011) An examination of the relationship of sample size and mean length of utterance for children with developmental language impairment. *Child Language Teaching and Therapy* 27: 286–293.

Cicourel A (1996) Ecological validity and 'white room effects': The interaction of cognitive and cultural models in the pragmatic analysis of elicited narratives from children. *Pragmatics and Cognition* 4: 221–264.

Cleave P and Rice M (1997) An examination of morpheme BE in children with specific language impairment: The role of contractibility and grammatical form class. *Journal of Speech, Language, and Hearing Research* 40: 480–492.

Cole K, Mills P and Dale P (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.

Crystal D (1982) *Profiling Linguistic Disability*. London: Edward Arnold.

Crystal D and Fletcher P (1979) Profile analysis of language disability. In: Fillmore C, Kempler D and Wang W (eds) *Individual Differences in Language Ability and Language Behavior*. San Francisco, CA: Academic Press, pp. 167–188.

Crystal D, Fletcher P and Garman M (1976) *The Grammatical Analysis of Language Disability: A Procedure of Assessment and Remediation*. The Hague: Elsevier-North Holland.

Crystal D, Fletcher P and Garman M (1989) *Grammatical Analysis of Language Disability*. London: Whurr.

Dunn M, Flax J, Sliwinski M and Aram D (1996) The use of spontaneous language measures as criteria for identifying children with specific language impairment. *Journal of Speech and Hearing Research* 39: 643–654.

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

Fenson L, Dale PS, Reznick JS, et al. (1994) Variability in early communicative development. *Monographs of the Society for Research in Child Development*, serial no. 242, vol. 59, no. 5. Wiley.

Gallagher T (1993) Pre-assessment: A procedure for documenting language use variability. In: Gallagher T and Prutting C (eds) *Pragmatic Assessment and Intervention in Language*. San Diego, CA: College-Hill Press, pp. 1–28.

Gavin W and 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.

Hansson K, Nettelbladt U and Nilholm C (2000) Contextual influence on the language production of children with speech/language impairment. *International Journal of Language and Communication Disorders* 35: 31–47.

Hewitt L, Hammer C, Yont K and Tomblin J (2005) Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSYN, and NDW. *Journal of Communication Disorders* 38: 197–213.

I CAN (2010) *Talking point: The first stop for information on children's communication* [website]. London: The Communication Trust. Retrieved from: http://www.talkingpoint.org.uk/Parent/Directory/Progress-Checker.aspx (November 2012).

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

Kramer C, James S and Saxman J (1979) A comparison of language samples elicited at home and in the clinic. *Journal of Speech and Hearing Disorders* 44: 321–330.

Leadholm B and Miller J (1992) *Language Sample Analysis: The Wisconsin Guide*. Madison, WI: Wisconsin Department of Public Instruction.

Lee L (1974) *Developmental Sentence Analysis: A Grammatical Assessment Procedure for Speech and Language Clinicians*. Evanston, IL: Northwestern University Press.

Leonard L (1998) *Children with Specific Language Impairment*. London: MIT Press.

Leonard L, Eyer J, Bedore L and Grela B (1997) Three accounts of the grammatical morpheme difficulties of English-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research* 40: 741–753.

Lindsay G (2011) The collection and analysis of data on children with speech, language and communication needs: The challenge to education and health services. *Child Language Teaching and Therapy* 27: 135–150.

Lund N and Duehart J (1993) *Assessing Children's Language in Naturalistic Contexts*. 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.

MacWhinney B and Snow C (1990) The child language data exchange system: An update. *Journal of Child Language* 17: 457–472.

Maillart C Parisse C and Tommerdahl T (2011) F-LARSP 1.0: An adaptation of the LARSP language profile for French. *Clinical Linguistics and Phonetics* 26: 188–198.

Marinellie S (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 (1996) Progress in assessing, describing, and defining child language disorder. In: Cole K, Dale P and Thal D (eds) *Assessment of Communication and Language*. Baltimore, MD: Brookes.

Minifie R, Darley F and Sherman D (1963) Temporal reliability of seven language measures. *Journal of Speech and Hearing Research* 6: 139–149.

Muskett T, Body R and Perkins M (2012) Uncovering the dynamic in static assessment interaction. *Child Language Teaching and Therapy* 28: 87–99.

Naremore R, Densmore A and Harman D (1995) *Language Intervention with School-aged Children: Conversation, Narrative, and Text*. San Diego, CA: Singular.

Owens R (1999) *Language Disorders: A Functional Approach to Assessment and Intervention*, 3rd edn. Boston, MA: Allyn and Bacon.

Rice M and Wexler K (1996) A phenotype of specific language impairment: Extended optional infinitives. In: Rice M (ed.) *Toward a Genetics of Language*. Mahwah, NJ: Lawrence Erlbaum, pp. 215–237.

Rice M, Smolik F, Perpich D, et al. (2010) Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research* 53: 333–349.

Rice M, Wexler K and Cleave P (1995) Specific language impairment as a period of extended optional infinitive. *Journal of Speech and Hearing Research* 39: 1239–1257.

Roland D, Dick F and Elman J (2007) Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language* 57: 348–379.

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

Scarborough H (1990) Index of productive syntax. *Applied Psycholinguistics* 11: 1–22.

Shewan C and Henderson V (1988) Analysis of spontaneous language in the older normal population. *Journal of Communication Disorders* 21: 139–154.

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

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

Wexler K (1994) Optional infinitives. In: Lightfoot D and Hornstein N (eds) *Verb Movement*. New York: Cambridge University Press, pp. 305–350.

Yoder P and Kaiser A (1989) Alternative explanations for the relationship between maternal verbal interaction style and child language development. *Journal of Child Language* 16: 141–160.