Published in final edited form as:

Clin Linguist Phon. 2015 June ; 29(6): 441–454. doi:10.3109/02699206.2015.1016187.

From Aardvark to Ziggurat: A New Tool for Assessing Children's Use of Rare Vocabulary

Jamie Mahurin-Smith,

Illinois State University

Laura DeThorne, and

University of Illinois at Urbana-Champaign

Stephen Petrill

The Ohio State University

One of the hallmarks of language development in school-aged children is their increasing use of more sophisticated, more colourful vocabulary. A preschooler learns to label an object 'green'; a school-aged child might call it 'seafoam' or 'emerald' or 'olive'. These advances in expressive vocabulary are difficult to quantify, in part because it is difficult to measure the relative rarity of a given word within the large set of words used by school-aged children. Existing standardized tests of expressive vocabulary disproportionately emphasize concrete nouns; existing language sample measures of vocabulary diversity, such as number of different words (NDW), may vary as a function of topic maintenance and sample length. For researchers or clinicians who wish to evaluate children's ability to use uncommon vocabulary in their spontaneous language, the available options are limited. In this paper we introduce a tool we have dubbed WERVE, the Wordlist for Expressive Rare Vocabulary Evaluation. WERVE is both a list of low-frequency vocabulary (LFV) words that occur infrequently in the language samples of school-aged children, and a method for generating automated tallies of those words in language samples via Computerized Language Analysis software (CLAN; MacWhinney & Spektor, 2009). The purpose of this paper is twofold: first, to provide other language sample researchers with the LFV wordlist and strategies for assessing LFV use in a broader population; and second, to establish the degree to which LFV use correlates with established language sample measures (particularly mean length of utterance (MLU), number of different words (NDW), and number of total words (NTW)) and with scores on standardized tests of language and cognition as a means of evaluating content and construct validity.

Semantic Development in School-Aged Children

Theorists who study vocabulary acquisition have proposed a variety of explanations for the process of word learning. Some authors propound the incremental view, in which words are learned slowly and gradually over repeated exposures, while other researchers argue for more rapid acquisition of new vocabulary (Medina, Snedeker, Trueswell, & Gleitman, 2011). Still others have asserted that the number of exposures to a word is less important than the context in which the word occurs (Sternberg, 1987; Nagy, 1995) or the timing of those exposures (Sobel, Cepeda, & Kapler, 2011). For school-aged learners in particular,

researchers have debated the relative contributions of literacy skills and spoken language skills to word learning (see, for instance, Rosenthal & Ehri, 2011). Although the present study was not designed to examine any particular theoretical model of vocabulary acquisition, it offers a useful resource for researchers interested in the impact of word frequency on patterns of learning.

Many scholars of word learning would predict that the likelihood that a child will use a given word is related to the frequency with which that word occurs in spoken English. Almost all typically-developing 2-year-old children will be able to use the word 'bird'; few will use 'cockatiel' or 'merlin'. In structured tasks children show less competence with low-frequency words (Weizman & Snow, 2001; Marinellie & Chan, 2006; Beals, 1997, Beals & Tabors, 1995). Development of this aspect of semantic skill, leading to appropriate use of uncommon words, adds colour and sophistication to children's conversations and their narratives. In fact, the layers of nuance added by diverse vocabulary are an important component of mature language use.

During the early years of language development, the rapid pace of expressive vocabulary acquisition for typically developing children is a source of surprise and delight for many parents. On average, a child's vocabulary grows from a few spoken words at age 1, to approximately 40 words by 16 months, to an average of 570 words by age 2½ (Fenson et al., 1994). Few parents recognize that vocabulary growth becomes even more rapid, at least in absolute terms, as school-aged children acquire new words via reading in addition to spoken language. It is estimated that typically-developing school-aged children learn between 10 and 13 new words each day, or 3,000–5,000 per year (Nagy, Anderson, & Herman, 1987; Nippold, 1998). Evidence from standardized vocabulary tests indicates that raw scores on these measures advance rapidly during the early school years. Further research is required, however, to elucidate the trajectory of this development and its relationship to children's routine vocabulary use.

Later-developing vocabulary is characterized by its increasing proportion of abstract terms that lack any concrete referent. While a young child can look at a picture of an alpaca and infer that 'alpaca' means a long-necked long-haired hoofed animal, an older child has a more complex task ahead in deciphering abstract terms like 'perception' or 'unjust'. As children's vocabularies grow, they learn increasing numbers of words that are synonyms and close cousins; an important component of early semantic development is untangling the nuances that distinguish closely related words. In addition to different words with similar meanings, English abounds in polysemous words, which have multiple distinct meanings. One example is 'set', a word with 100 different definitions at one online dictionary (<http://dictionary.reference.com>, retrieved 8/12/14). As children learn that different words can have similar meanings, they also continue to learn that the same word can have very different meanings. The preschooler who could distinguish between instructions to *set* the table and pick up the Lego *set* grows into a school-aged child who understands that *set* can refer to part of a tennis game, to the configuration of a ship's sails, or to what happens at great-grandma's weekly appointment with the hairdresser.

In addition to acquiring a wide variety of root words, English-speaking children expand their vocabularies via derivational morphemes, which are especially common in academic reading materials (Nippold & Sun, 2008). The English language includes more than a hundred affixes that can be used to form new words (Nippold, 2007). Some of these are acquired early, as evidenced by a two-year-old who shouts ‘*Un*-eat it!’ on discovering that the last slice of cake has been consumed. Others, such as the -ent suffix that transforms ‘reside’ into ‘resident’, emerge much later. During the school years, typically-developing children are broadening their ability to use affixes such as -ness, -ship, -ful, -able, along with many others (Anglin, 2000).

A further characteristic of school-aged vocabulary development is specialization. Since children acquire words largely through reading and conversation (Nippold, 2007; Beals & Tabors, 1995; Beals, 1997; Weizman & Snow, 2001), different interests can lead to different vocabularies: ‘cutlass’ and ‘moiety’ for a child who reads old-fashioned adventure stories; ‘theropod’ and ‘minmi’ for a dinosaur enthusiast. Family culture also plays a role in shaping vocabulary, so that a child might be familiar with the word ‘wasabi’ but puzzled by ‘scrapple’, or vice versa. During the preschool years, parents are routinely asked to complete the MacArthur Communicative Development Inventory, an instrument that asks them to report which of roughly 700 words their child can produce. But it is difficult if not impossible to imagine a comparable instrument that could capture the breadth and diversity of school-aged vocabulary development; certainly, no such parent report or observational measure exists at present. Existing language sample measures such as Number of Different Words (NDW) and Number of Total Words (NTW) provide tallies of types and tokens but no information as to the relative rarity of the words in a child’s language sample; it is unclear whether a large NDW reflects a varied and sophisticated vocabulary or merely poor topic maintenance skills. Similarly, Measure D (Malvern & Richards, 2002) was developed to assess vocabulary diversity while controlling for volubility, but it may be confounded by rapid topic changes. In some of our previous work we have found that Measure D correlates unpredictably with existing language sample measures (DeThorne et al., 2008; Mahurin-Smith, DeThorne, Logan, Channell, & Petrill, 2014). Existing studies have thus employed a variety of strategies for assessing children’s command of low-frequency vocabulary, often choosing standardized tests of receptive vocabulary as their dependent variable (Weizman & Snow, 2001; Marinellie & Chan, 2006; Beals, 1997, Beals & Tabors, 1995). While the section that follows will review some of these findings in more detail, the studies under consideration are driven by the idea that low-frequency words are more difficult for children to understand and to use.

Existing Studies of Low-Frequency Vocabulary

Rare vocabulary has been the topic of several studies, some of which emphasize written language. In a 2006 paper, Marinellie and Chan elicited written definitions of both high-frequency and low-frequency nouns and verbs, based on frequency data from Kucera and Francis (1967), in a sample of 120 children in grades 4, 7, 10, and college. They found significant word-frequency effects in the children’s definitions of nouns and verbs, a result that aligns with findings from a 2003 study of high-frequency and low-frequency adjective definitions by Marinellie and Johnson.

Other studies have focused on low-frequency words in spoken language. In a 1995 paper, Beals and Tabors investigated mothers' use of rare vocabulary in conversation with their 3- and 4-year-old children. These authors defined rare vocabulary words as those that did not appear on a list of 7881 words derived from the Chall and Dale list of words familiar to most fourth graders (1995). In this sample of 84 children from 81 families, the researchers considered the relationships between rare vocabulary in mothers' and children's language output and children's scores on the Peabody Picture Vocabulary Test – Revised (PPVT-R; Dunn & Dunn, 1981) at age 5. They reported that maternal use of rare vocabulary in play settings (though not in book-reading activities) was significantly correlated with children's higher PPVT scores. Children's use of rare vocabulary in play settings and at mealtime (though again, not in book-reading activities) was likewise significantly correlated with PPVT results. Family conversations, the investigators concluded, were an important locus of rare vocabulary learning for young children.

Like Beals and Tabors (1995), Hoff and Naigles looked at language sample measures of vocabulary diversity in their 2002 study of the effects of preschool children's exposure to uncommon words, reporting a robust correlation between higher maternal NDW and higher NDW in children. More frequently, however, studies of children's vocabulary skills rely on standardized test results. For instance, Snow and Beals (2006) described the relationship between mealtime talk and vocabulary acquisition in their sample of 70 children aged 3–5. They found that exposure to rare vocabulary, defined as words not included in the 1995 Chall and Dale list of words familiar to most fourth-graders, was significantly associated with children's scores on the PPVT-R. Similarly, Beals (1997) and Weizman and Snow (2001) used the PPVT-R to measure the relationship between rare vocabulary exposure (again defined with reference to Chall and Dale, 1995) and children's vocabulary acquisition.

This strategy of using a standardized measure of receptive vocabulary, however, is suboptimal. Since latent variable analyses have indicated that standardized tests and language sample measures load on different factors, reflecting different sets of skills, (DeThorne et al., 2008; Ukrainetz & Blomquist, 2002), and since expressive and receptive vocabulary measures correlate imperfectly (cf. Restrepo, Schwanenflugel, Neuhaarth-Pritchett, Cramer, & Ruston, 2006; Ukrainetz & Duncan, 2000), a standardized test of receptive vocabulary may or may not provide accurate information about a child's ability to choose words well in discourse-based situations. In contrast to the highly structured nature of a standardized vocabulary test, a discourse-based measure offers improved social validity of everyday social interactions.

For researchers interested in language sample measures, much of the available information on word frequency comes from large existing corpora, such as the Corpus of Contemporary American English (Davies, 2008), or older examples like the London-Lund corpus (Brown, 1984), the American Heritage Word Frequency Book (Carroll, Davies, & Richman, 1971), or the Kucera-Francis corpus (Kucera & Francis, 1967). Such corpora present three obstacles to use with current child language samples: first, they were developed chiefly from the language of adults, including (particularly for the latter two examples) speakers of varied dialects such as British English; second, they frequently draw on written language samples;

third, they may be decades old and will necessarily misrepresent words heard frequently in the speech of children today, such as ‘cellphone’, ‘internet’, and even ‘soccer’. Consequently, they may yield misleading results when used to evaluate children’s spoken language. Multiple studies of children’s language have used the Dale-Chall list of familiar words (Chall & Dale, 1995) to define rare vocabulary, but this list poses difficulties as well. It was intended to include words known to fourth-grade readers, but a number of the words it contains (e.g., “savage,” “postage,” “schoolmaster”) are uncommon in the spontaneous expressive language of school-aged children. Furthermore, the language has continued to change over the 20 years since its publication, rendering words such as “steamboat” even more unusual in children’s conversations.

Instead of comparing the spoken language of children to both spoken and written language from both children and adults in previous decades, our intent with this project was to compare the spoken language of a sample of school-aged children with the spoken language of their present-day peers. Its purpose was to evaluate rare vocabulary as observed in a corpus of 1.2 million words and to assess developmental changes across an age range in which rapid vocabulary growth has been frequently asserted but seldom quantified. Our study was designed to share a systematic resource for examining low-frequency vocabulary use and provide a preliminary assessment of its content and construct validity by addressing the following research questions.

1. Within the WRRP twin sample, does children’s use of LFV, as derived via WERVE, correlate with their performance on established language sample measures (i.e. MLU, NDW, and NTW) and standardized language tests?
2. Within the WRRP twin sample, does children’s use of LFV show developmental change in a longitudinal comparison across a one-year time period?
3. To what extent do findings about LFV use in the WRRP twin sample generalize to an independent sample? In other words, does WERVE appear to have utility beyond the context in which it was developed?

METHOD

Participants

Data collection for this study was part of the Western Reserve Reading Project (WRRP), a longitudinal study of the genetic and environmental components of children’s abilities in reading, mathematics, and related skills (Petrill et al., 2006). The WRRP focuses primarily on a sample of 438 same-sex twin pairs from Ohio who began participating in the study when the children were in either kindergarten or first grade and were followed longitudinally thereafter. Approximately 10% of WRRP families also completed a portion of the testing protocol separately with a singleton sibling to allow the researchers to assess for selected differences between twins and singletons.

We evaluated research questions 1 and 2 in Sample 1, a subset of the transcripts obtained from the twin pairs at two time points, one year apart. To address research question 3, regarding the generalizability of the findings from the test sample, we then assessed a set of

transcripts obtained in conversations with singleton children whose families are involved in WRRP, referred to as Sample 2.

Sample 1 included 112 of the WRRP twins who were initially selected for a study comparing outcomes for premature and full-term twins (Mahurin-Smith, DeThorne, Logan, Channell, & Petrill, 2014). At the initial assessment their mean age was 7.19 years ($SD = 0.71$); at the second assessment, when 92 of the original participants remained in the study, their mean age was 8.30 years ($SD = 0.81$). Sample 2 comprised 38 singleton siblings of twins enrolled in WRRP, assessed at a mean age of 8.89 years ($SD = 1.32$). Five of the children in Sample 2 were siblings from families whose twins were included in Sample 1.

General Procedures

Twins were visited annually in their homes by a pair of trained WRRP examiners for approximately two hours per visit, completing a number of measures intended to assess their math, reading, and language skills. The sequence of home visits under consideration in the present study is summarized in Table 1. Co-twins were simultaneously evaluated in separate rooms, each by a separate examiner. At the first home visit (HV1), which was not included in the present study, assessments focused on early reading skills and the children's home environment; no conversational language measures were obtained at this time.

Approximately a year later, evaluators returned for a second home visit (HV2) to collect a variety of data on reading, math, cognition, and language skills, including a conversational language sample and the short form of the Stanford-Binet Intelligence Scale (Thorndike, Hagen, & Sattler, 1986). Home visit 3 (HV3) followed approximately a year after HV2 and included conversational language sampling to assess growth and change, along with other measures of language and reading ability. Home visit 4 (HV4), which was also not included in the present study, was a mid-year visit emphasizing mathematics skills. At home visit 5 (HV5), the measures under consideration in the present study included the following four subtests of the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF-4; Semel, Wiig, & Secord, 2003), administered to provide a standardized measure of the children's language skills: Recalling Sentences, Understanding Paragraphs, Word Classes (Receptive), and Word Classes (Expressive).

Singleton siblings were recruited for an evaluation at a single home visit when they happened to be between the ages of 6.33–11.92 years (i.e. the age range of twins at HV3) at the time of a twin home visit. In addition to the language sample, measures included the Stanford-Binet vocabulary subtest.

Language Sample Procedures

During HV2 and HV3 for the test sample as well as the one home visit for the singleton sample, 15-minute conversational language samples were recorded while a child and an examiner played with modeling clay. Examiners were trained using Leadholm and Miller's guidelines (1992) for obtaining language samples, with an emphasis on child-directed conversational topics and reliance on open-ended questions. Examiners were instructed to limit their use of requests, directions, and questions requiring only one-word answers. Additionally, they were adjured to exercise patience in order to encourage children to

respond, and to leave quiet interludes in their exchanges with less talkative children rather than peppering them with questions. Sample topics of conversation were provided to them during training along with model questions and responses. Further details on examiner training are available in DeThorne and Hart (2009).

Trained research assistants transcribed the language samples in the second author's laboratories at the Pennsylvania State University and the University of Illinois according to protocols developed for use with Systematic Analysis of Language Transcripts (SALT, Version 8.0; Miller, 2004). Nippold's 1998 guidelines for communication units (C-units) were used for marking utterance boundaries. The guidelines described by Retherford (2000) were used for frozen forms such as "French fries", which were treated as one word rather than two. After transcription, each sample was reviewed by a second trained research assistant, who corrected any transcription errors.

Creation of the WERVE

We used CLAN to build a concordance of the 1.2 million words used in the 1,437 unique WRRP twin transcripts available for home visits 2, 3, and 5 in December 2008. The first step in creating the WERVE was to pare down the inclusive concordance, in which word frequency ranged from 50,329 instances (for 'and') all the way down to 1 instance (for 'ziggurats'), to create a low-frequency concordance. For this exploratory task, we selected words with 15 or fewer uses across the entire WRRP twin sample at home visits 2, 3, and 5. In choosing this cutoff, our aim was to select a number that would yield a wide range of results, with some children using none of the words designated as uncommon and others using many of them. The low-frequency concordance was edited to remove typographical errors, proper nouns, numbers, and sound effects (e.g. zzzz). In addition, kinship terms were removed from consideration regardless of frequency of use; if a child called his grandparents 'oma and opa', for instance, those words were excluded despite their rarity in the corpus. Morphemic variations were collapsed (i.e. 'borrowing' and 'borrowed' were counted together with 'borrow').

At this point the low-frequency list included many words that could serve as a useful index of expressive vocabulary development, but it also included a number of early-acquired words that seldom surfaced in the transcripts (e.g. beans, soap, helicopter). To address this issue, the low-frequency wordlist from WRRP was cross-checked against the MacArthur-Bates Communicative Development Index (Fenson et al., 2007), a list which includes the words produced by at least 15% of 2.5-year-olds in their normative sample (cf. Beals & Tabors, 1995 for a similar procedure). These words, 61 of which appeared on the edited WRRP low-frequency list, were eliminated. This approach to tallying low-frequency vocabulary was referred to in shorthand as the 'Rule of 15s': low-frequency words are those that occurred no more than 15 times in the large corpus, and were acquired by <15% of young children in the MacArthur-Bates sample. These steps yielded a list of 2079 low-frequency words, which we have called the Wordlist for Expressive Rare Vocabulary Evaluation, or WERVE.

The `FREQ` command in CLAN was used to count occurrences of these low-frequency words (types, rather than tokens) in Sample 1. For each transcript, a list was generated of the low-frequency words and the sentences in which they appeared. This output was reviewed

carefully by the first author. Words that appeared solely because of typographical errors, or exploratory and semantically inappropriate uses, were removed from a child's tally. Because of the subjective nature of this process, reliability comparisons were undertaken in Sample 1. A research assistant from the UIUC Child Language and Literacy Laboratory reviewed the tallies for 10% of the transcripts and recorded her judgments as to semantic appropriateness. Point-to-point reliability was >90%.

Finally, a density measure for low-frequency vocabulary was calculated for each transcript. The number of low-frequency word types in each sample was divided by the number of utterances in that sample in order to reduce confounding with sample length. We then used the density measure to evaluate research questions 1 and 2 for Sample 1, assessing correlations with established language sample measures, standardized tests, and longitudinal change.

Once we had obtained statistically significant correlations in Sample 1, we addressed similar questions in Sample 2. Note that for the singleton samples we had only one conversational language sample, and no CELF-4 or TNL scores; these analyses are necessarily simpler. Note further that we considered these questions in two stages, first examining the performance of the original WERVE, and next finding and integrating the words unique to Sample 2. In both stages, we used frequency data based on the full concordance to determine the relative rarity of words in a given transcript.

Sample 1 included 55 twin pairs; the similarities between twins meant that correlation coefficients might be inflated. For this reason, correlation coefficients are presented separately for firstborn and second-born twins, while the singleton sibling results yielded a single correlation value. The similarities between these values serve as a form of replication in themselves. For these exploratory analyses with a novel measure, we did not correct for multiple comparisons.

RESULTS

WERVE Outcomes in Sample 1

At HV2, LFV counts per sample ranged from 0–37 per sample with a mean of 7.86 (SD = 5.65). LFV densities ranged from 0–0.20, with a mean of 0.057 (SD = 0.035). For reference, a density value of 0.20 would correspond to the use of one low-frequency word in every 5 utterances, while a density value of 0.05 would correspond to the use of one low-frequency word in every 20 utterances. The distribution of values for LFV counts was positively skewed (skewness = 2.16, kurtosis = 7.03), an effect that diminished for LFV density (skewness = 1.40, kurtosis = 2.59). At HV3, LFV counts covered a wider range, from 0–45 per sample. The mean LFV count was 11.84 (SD = 8.18). LFV density values ranged from 0–0.27, with a mean of 0.083 (SD = 0.052). Like the HV2 sample, this distribution of values was positively skewed (skewness = 1.77, kurtosis = 4.67); in the same way, LFV density more closely approximated the normal distribution (skewness = 1.51, kurtosis = 2.92). Descriptive statistics for the language sample measures and the standardized test results were reported in Mahurin-Smith, DeThorne, Logan, Channell, & Petrill (2014); those results are not repeated here since they are not the focus of the present study.

To evaluate our first research question, examining correlations between LFV and the established language sample measures MLU, NDW, and NTW, we created the correlation matrices shown in Tables 2 and 3 for HV2 and HV3 respectively. In addition, each correlation matrix is derived separately for first- and second-born twins to ensure independent samples. Readers will observe that the HV2 correlations for LFV range from 0.31 to 0.59, a medium-to-large effect size (Cohen, 1992). Of the six values, the five largest are statistically significant at .05. At HV3 the correlations were higher across the board, ranging from 0.51 to 0.75; all were statistically significant.

Existing research into the relationships between language sample variables and standardized test scores informed our *a priori* expectations about correlations between LFV and standardized measures. Since prior latent variable analyses indicated that conversational language measures and standardized test scores loaded on different factors (DeThorne et al., 2010), we expected that correlations between LFV density and the available standardized test scores would be smaller in magnitude than correlations between LFV density and language sample measures. Indeed, at HV2 the correlations between LFV density and IQ were modest. As shown in Table 2, only one of the two reached statistical significance. At HV3 these correlations were larger and statistically significant (see Table 3), with large effect sizes for the language sample measures and medium effect sizes for the standardized tests. We also had results for two expressive subtests of the Clinical Evaluation of Language Fundamentals – Fourth Edition for 81 of the children, collected at HV5 when they were 10 years old, on average. These two standard scores were averaged to yield a single variable. Correlation coefficients for these measures are reported in Table 3; all are statistically significant at .05 with a large effect size.

Since an important test of a language sample measure's validity is its ability to capture a child's growth over time, we looked at each child's pattern of change across HV2 and HV3. It is noteworthy that the mean LFV density increased as well as the mean LFV count, as this indicates that the children were using more rare words per utterance at HV3, and not merely producing more utterances. Of the 92 children with data for both HV2 and HV3, 63, or 68.5% showed an increase in LFV density. In contrast, 45 of the 89 children with MLU results for both visits (50.6%) showed an increase in MLU and 33 of the 60 children with NDW results for both visits (55.0%) had gains in NDW. Note that NDW is based on a 100-utterance cut, so that a larger number of children were missing data at one or both visits. The correlation between age and LFV density was a highly significant 0.45.

WERVE Outcomes in Sample 2

To probe the generalizability of our findings in Sample 1, we used a two-pronged approach to evaluate low-frequency vocabulary use in Sample 2. We looked first at results for the original WERVE, which provided a tally of the low-frequency words from the original list that appeared in the Sample 2 transcripts. It did not, however, tally any of the words unique to the Sample 2. Next we obtained outcomes for the augmented WERVE, which combined the original WERVE results with those words from Sample 2 that never appeared in the initial corpus. Using the original WERVE word list, LFV counts in Sample 2 ranged from 1–20 with a mean of 7.3 ($SD = 4.6$), while LFV density values ranged from 0.01–0.14, with a

mean of 0.05 ($SD = 0.03$). This density value indicates that the children in Sample 2 used a word from the WERVE list approximately once every 20 utterances, a result that aligns neatly with the HV2 outcomes from Sample 1. Using the augmented WERVE, LFV counts ranged from 2–34 with a mean of 11.75 ($SD = 7.56$); LFV density values ranged from 0.01–0.20, with a mean of 0.078 ($SD = 0.045$). A comparison of these two sets of results shows that the original WERVE detected 64%, or 0.05/0.078, of the children's actual low-frequency vocabulary words as measured by the augmented WERVE.

A correlation matrix showing both the original and the augmented WERVE appears in Table 4. Because this sample is composed of singletons rather than twins, only one value appears in each cell of the table. All of the correlations among LFV density and language sample measures are sizable and highly significant, with slightly higher values observed for the augmented WERVE. As expected, the correlations between IQ and the other measures are more modest, with a non-significant value of 0.24 observed for the original WERVE and a significant value of 0.38 seen for the augmented WERVE.

DISCUSSION

Our results support the measurement of LFV density as a strategy for assessing the development of semantic skill in school-aged children. In both samples, more frequent use of rare word types was associated with stronger conversational language skills as measured by MLU, NDW, and NTW, and was positively associated with scores on standardized measures including the Stanford-Binet, the TNL, and subtests of the CELF-4. Although correlations between LFV density and existing measures were uniformly positive and usually significant, it is clear from the magnitude of the correlations that LFV density measures something unique. The largest of these correlations was 0.72 (for LFV-NDW in firstborn twins at HV3), yielding an R^2 of 0.52. In other words, much of the variance in children's spontaneous use of uncommon vocabulary is unexplained by established language sample measures.

This paper corroborates existing findings that describe an association between rare vocabulary use and the development of strong language skills (Weizman & Snow, 2001). In this study, children who used more unusual vocabulary in their conversations were also more likely to use a variety of words and to combine them into longer utterances, as measured by NDW and MLU, and to score higher on expressive subtests of the CELF. Our study extends previous findings by looking directly at school-aged children's use, rather than their comprehension, of low-frequency vocabulary, and by providing frequency data on word use based on a large set of language samples collected within the past decade.

While Sample 2 might be viewed as a replication sample, a limitation to this perspective is that it is made up of children who are very similar demographically to the children in Sample 1; in 5 cases, they are biologically related to the children in Sample 1. Both sets of transcripts record the conversations of a largely white middle class group of children from the Midwestern USA. It may be that WERVE is less effective in measuring low-frequency vocabulary for children from different socioeconomic strata or from different geographic locales, whose word choices could be more heterogeneous.

A drawback to using WERVE is that it can be labor-intensive if users wish to review words in context to eliminate exploratory uses or transcription artifacts. Only a single line of code is required to generate immediate output for a batch of transcripts (see Appendix A for particulars). Assessing semantic appropriateness in context, however, can greatly increase the amount of time required to obtain results. While sifting through WERVE output is more time-consuming than obtaining automatically generated measures such as MLU or NDW, in the authors' estimation it is much faster than correcting computer output for Developmental Sentence Scoring or for CLAN's MOR/POST programs, as well as a significantly easier task for students to learn. For researchers interested in language sample analysis, it may prove a useful adjunct to existing semantic measures, a tool that allows for more finely grained assessment of vocabulary in a context with strong ecological validity.

Despite the time required to derive WERVE results, we propose that it is worthwhile because it fills a gap in our repertoire of available assessment tools. Standardized tests of expressive vocabulary have limited social validity and tend to emphasize concrete nouns since they are the easiest to illustrate. While receptive vocabulary tests allow researchers and clinicians to look at a wider array of word types, they do not provide any information about a child's ability to deploy those words in conversation. Researchers have recognized the importance of assessing vocabulary use in context (see Brackenbury & Pye, 2005; Ukrainetz & Duncan, 2000), but our existing measures do not look explicitly at children's ability to use less common vocabulary in their spontaneous language.

The wordlists we have described here have a number of potential applications. While we have focused on semantically appropriate instances of low-frequency vocabulary, researchers interested in the stages of mastery in children's word-learning might find it fruitful to look also at semantically inappropriate occurrences of uncommon words, or at how patterns of use shift over time as a child matures. Researchers who study child language, whether their focus is on children with impairments, typically developing children, or gifted children will find that WERVE allows them to pull out all of the sentences in a sample that contain distinctive vocabulary. Those who study language impairment may find it to be a useful index of changes in children's conversation-level semantic skills. Those who study language change may find it interesting to compare the frequency counts from this corpus with those from earlier decades. It could be fruitful for theorists to consider patterns of more frequent versus less frequent usage: what might children's vocabulary selections tell us about the mechanisms of word-learning?

We invite other researchers to explore the uses of WERVE in further language samples. To this end we have provided four appendices. In the first appendix we have provided the CLAN code that will allow researchers to generate tallies of the low-frequency words in CHAT-formatted files. Using the original WERVE is a straightforward undertaking, requiring only a line of code. The original WERVE found 64% of the low-frequency vocabulary words in Sample 2, producing results that correlated strongly with established language sample measures. For these reasons, users may determine that the simple approach is adequate for their needs. Generating an augmented list with the words unique to a new transcript or set of transcripts is a more complex task, but we have provided step-by-step instructions. A second online appendix includes the subset of the WERVE words used in

Sample 1, listed alphabetically; the full WERVE list is available online in .cut format for use with CLAN. The final appendix is the concordance derived from the entire WRRP corpus of 1.2 million words, also provided online as a .cut file. This concordance has been lightly edited to remove family names and unusual place names that might compromise participants' confidentiality. This list of words used by school-aged children, broken down by their frequency, may be of interest in its own right.

Even a cursory glance at the WERVE words will show that it comprises a mix of sophisticated vocabulary (e.g. 'cardiologist' or 'angolosaurus') and much more familiar vocabulary (e.g. 'direction' or 'cucumber'). As one might expect, words that appeared 15 times in the larger corpus are generally more familiar than words that appeared 1 time. Our choice to use the cutoff of 15 occurrences was somewhat arbitrary, due to the exploratory nature of this project, and it could be the case that a different cutoff might yield a more sensitive measure. The associations between WERVE and existing measures, as well as the distribution of WERVE results, support our decision to use 15 as a cutoff, but this is an avenue for future research to explore.

At this stage, WERVE is primarily a research tool, but it may have future clinical uses. Every speech-language pathologist who works with children has experienced the frustration of administering a standardized language test to a child whose score does not reflect his true abilities, for reasons that might include attention deficits, cultural barriers, or behavioural issues. At present, the collection, transcription, and analysis of a language sample in a clinical setting can seem like a cumbersome series of tasks for a busy SLP, rendering language sample measures less practical than standardized measures. However, speech recognition software has improved substantially over the past 5 years, and further improvements could transform the collection of language samples in a clinical setting into a more manageable task. It is conceivable that in the context of such technological advances, WERVE has the potential to facilitate assessment of vocabulary skill and vocabulary growth in a task with inherent social validity: a conversation. Whether or not language sampling becomes more common in clinical practice, SLPs and investigators choosing intervention targets may also benefit from a wordlist that describes frequency of use for a wide variety of words, as observed in the spoken language of contemporary school-aged children. We hope that further studies will shed further light on the uses and limitations of WERVE, and we encourage correspondence from interested researchers.

Acknowledgments

The authors gratefully acknowledge the contributions of Cynthia Johnson, Charles Morton, and Ruth Watkins to the development of WERVE. We also appreciate the assistance we received from Meredith Kresse, Julie Mahieu, Lisa Mellman, and Maggie Wingate. Funding for this project came from NICHD #HD50307, #HD038075 and #HD075460, and from U.S. Department of Education #H325D07006.

REFERENCES

- Anglin JM. Vocabulary development: A morphological analysis. Monographs of the Society for Research in Child Development. 2000; 58 (10, Serial No. 238).
- Beals DE. Sources of support for learning words in conversation: Evidence from mealtimes. *Journal of Child Language*. 1997; 24:673–694. [PubMed: 9519590]

- Beals DE, Tabors PO. Arboretum, bureaucratic, and carbohydrates: Preschoolers' exposure to rare vocabulary at home. *First Language*. 1995; 15(1):57–76.
- Brackenbury T, Pye C. Semantic deficits in children with language impairments: Issues for clinical assessment. *Language, Speech, and Hearing Services in Schools*. 2005; 36(1):5.
- Brown GDA. A frequency count of 190,000 words in the London-Lund corpus of English conversation. *Behavioral Research Methods, Instruments, & Computers*. 1984; 16(6):502–532.
- Carroll, JB.; Davies, P.; Richman, B. *The American Heritage Word Frequency Book*. Boston: Houghton Mifflin; 1971.
- Chall, JS.; Dale, E. *Readability revisited: The New Dale-Chall readability formula*. Cambridge, MA: Brookline Books; 1995.
- Cohen J. A power primer. *Psychological Bulletin*. 1992; 112:155–159. [PubMed: 19565683]
- Davies, Mark. *The Corpus of Contemporary American English: 450 million words, 1990-present*. 2008. Available online at <http://corpus.byu.edu/coca/>
- DeThorne LS, Hart SA. Use of the twin design to examine evocative gene-environment effects within a conversational context. *European Journal of Developmental Science*. 2009; 3:175–194. [PubMed: 22102850]
- DeThorne LS, Petrill SA, Hart SA, Channell RW, Campbell RJ, Deater-Deckard K, Thompson LA, Vandenberg DJ. Genetic effects on children's conversational language use. *Journal of Speech, Language, and Hearing Research*. 2008; 51:423–435.
- DeThorne LS, Petrill SA, Schatschneider C, Cutting L. Conversational language skills as a predictor of early reading development: Language history as a moderating variable. *Journal of Speech, Language, and Hearing Research*. 2010; 53:209–223.
- Dunn, L.; Dunn, L. *Peabody Picture Vocabulary Test – Revised Form*. Circle Pines, MN: American Guidance Service; 1981.
- Fenson L, Dale PS, Reznick JS, Bates E, Thal DJ, Pethick SJ, Stiles J. Variability in early communicative development. *Monographs of the society for research in child development*. 1994; 59:1–185. [PubMed: 7845413]
- Fenson, L.; Marchman, VA.; Thal, DJ.; Dale, PS.; Reznick, JS.; Bates, E. *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual*. 2nd. Baltimore, MD: Brookes; 2007.
- Hoff E, Naigles L. How children use input to acquire a lexicon. *Child development*. 2002; 73(2):418–433. [PubMed: 11949900]
- Kucera, H.; Francis, WN. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton-Mifflin; 1967.
- Leadholm, BJ.; Miller, JF. *Language sample analysis: The Wisconsin guide*. Madison, WI: Wisconsin Department of Public Instruction; 1992.
- MacWhinney, B.; Spektor, L. *Computerized Language Analysis (Version 7.28.09)* [Computer software]. Pittsburgh, PA: 2009. Author
- Marinellie SA, Chan YL. The effect of word frequency on noun and verb definitions: A developmental study. *Journal of Speech, Language and Hearing Research*. 2006; 49(5):1001.
- Marinellie SA, Johnson CJ. Adjective definitions and the influence of word frequency. *Journal of Speech, Language and Hearing Research*. 2003; 46(5):1061.
- McKee G, Malvern D, Richards B. Measuring vocabulary diversity using dedicated software. *Literary and linguistic computing*. 2000; 15(3):323–338.
- Medina TM, Snedeker J, Trueswell JC, Gleitman LR. How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*. 2011; 18:414–420.
- Miller, J. *Systematic Analysis of Language Transcripts (Research Version 8.0)* [Computer software]. Madison, WI: University of Wisconsin; 2004.
- Nagy, WE. *On the role of context in first- and second-language vocabulary learning*. Champaign, IL: Center for the Study of Reading; 1995.
- Nagy WE, Herman PA, Anderson RC. Learning words from context. *Reading Research Quarterly*. 1985; 22:233–253.

- Nippold, MA. Later language development: The school-age and adolescent years. Austin, TX: Pro-Ed; 1998.
- Nippold, MA. Later language development: The school-age and adolescent years. 3rd. Austin, TX: Pro-Ed; 2007.
- Nippold MA, Sun L. Knowledge of morphologically complex words: A developmental study of older children and young adolescents. *Language, Speech, and Hearing Services in Schools*. 2008; 39(3): 365.
- Petrill SA, Deater-Deckard K, Thompson LA, DeThorne LS, Schatschneider C. Reading skills in early readers: Genetic and shared environmental influences. *Journal of Learning Disabilities*. 2006; 39(1):48–55. [PubMed: 16512082]
- Retherford, KS. Guide to analysis of language transcripts. 3rd. Eau Claire, WI: Thinking Publications; 2000.
- Restrepo MA, Schwanenflugel PJ, Blake J, Neuhaarh-Pritchett S, Cramer SE, Ruston HP. Performance on the PPVT-III and the EVT: Applicability of the measures with African American and European American preschool children. *Language, Speech, and Hearing Services in Schools*. 2006; 37(1):17.
- Rosenthal J, Ehri LC. Pronouncing new words aloud during the silent reading of text enhances fifth graders' memory for vocabulary words and their spellings. *Reading and Writing*. 2011; 24:921–950.
- Semel, E.; Wiig, EH.; Secord, WA. 4th. Toronto, Canada: The Psychological Corporation; 2003. *Clinical Evaluation of Language Fundamentals*.
- Snow CE, Beals DE. Mealtime talk that supports literacy development. *New directions for child and adolescent development*. 2006; 2006(111):51–66. [PubMed: 16646499]
- Sobel HS, Cepeda NJ, Kapler IV. Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*. 2011; 25:763–767.
- Sternberg, RJ. Most vocabulary is learned from context. In: McKeown, MG.; Curtis, ME., editors. *The nature of vocabulary acquisition*. New York: Lawrence Erlbaum; 1987. p. 89-106.
- Thorndike, RL.; Hagen, EP.; Sattler, JM. *The Stanford-Binet intelligence scale: Guide for administering and scoring*. Chicago, IL: Riverside; 1986.
- Weizman ZO, Snow CE. Lexical input as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology*. 2001; 37(2):265–279. [PubMed: 11269394]
- Ukrainetz TA, Blomquist C. The criterion validity of four vocabulary tests compared with a language sample. *Child Language Teaching and Therapy*. 2002; 18(1):59–78.
- Ukrainetz TA, Duncan DS. From old to new: Examining score increases on the Peabody Picture Vocabulary Test-III. *Language, Speech, and Hearing Services in Schools*. 2000; 31(4):336–339.

APPENDIX A

The following steps will allow you to obtain a tally of the low-frequency items that appear in your transcripts. If your files were created in SALT, use CLAN's SALTIN command to convert the formatting and the CHECK command to eliminate any errors. Once you have a batch of error-free CLAN-formatted files, place them in a folder with a copy of the file WERVE.cut and specify that folder as the working directory in the CLAN command window.

For a list of the WERVE items that appear in each transcript, listed in a single file named 'lowfreq.cut', go to the CLAN command line and type the following:

```
> freq +t*CHI +s@WERVE.cut *.cex > lowfreq.cut
```

If you would prefer to view each word in context so that you can assess its semantic appropriateness, use the following command to produce a file called 'context.cut':


```
> freq +t*CHI +d0 +s@WERVE.cut *.cex > context.cut
```

If you would like to generate a list of the words in your transcripts that never occurred in the WRRP corpus, the following steps will allow you to derive it. Copy the file ‘concordance.cut’ into your working directory. Use the following command to generate a file called ‘new_concordance.cut’ that will list all the words used by children in your batch of transcripts:

```
> freq +u +d1 +t*CHI *.cex > new_concordance.cut
```

Run this command to generate a file called ‘new_shared.cut’ that will identify shared and unique words in the two concordances:

```
> freq +y +o +u *concordance.cut | combo +y +s'2' +d | freq +y +d1 > new_shared.cut
```

Now use the following command to generate a list of words unique to your batch of transcripts:

```
> freq +y +o +u n*.cut | combo +y +s'1' +d | freq +y +d1 > unique.cut
```

Once you have edited unique.cut to eliminate slang, sound effects, and the like, you can combine unique.cut with the words in the WERVE.cut file to derive an inclusive list of low-frequency vocabulary: it will incorporate both the words from WERVE and those words used in your transcripts that never appeared in the WRRP corpus. You can then use the first two lines of code given above to generate tallies for your transcripts, substituting the appropriate file name. To generate a list of words using a file called inclusive.cut, you would type:

```
> freq +t*CHI +s@inclusive.cut *.cex > lowfreq.cut
```

Table 1

Summary of measures obtained at each home visit.

Visit number	Sample under consideration	Mean age	Measures used in this study
Home visit 2	Sample 1 (twins)	7.19 years (SD = 0.71)	Language sample, IQ
Home visit 3	Sample 1 (twins)	8.30 years (SD = 0.81)	Language sample, IQ
Home visit 5	Sample 1 (twins)	10.12 (SD = 0.92)	CELF-4
Sibling visit	Sample 2 (singleton siblings)	8.89 years (SD = 1.32)	Language sample, IQ

Table 2

Correlation matrices for Sample 1 at HV2, presented separately for first and second-born twins.

	LFV	MLU	NDW	NTW	IQ
LFV	1.00				
MLU	0.57 ^{***} , 0.47 ^{***}	1.00			
NDW	0.49 [*] , 0.52 ^{***}	0.78 ^{***} , 0.78 ^{***}	1.00		
NTW	0.47 [*] , 0.31 [*]	0.86 ^{***} , 0.93 ^{***}	0.90 ^{***} , 0.83 ^{***}	1.00	
IQ	0.31 [*] , 0.21 [*]	0.40 ^{***} , 0.09 [*]	0.36 [*] , 0.23 [*]	0.39 ^{***} , 0.20 [*]	1.00

Note: Sample 1 included 55 twin pairs, who were divided into firstborn and second-born groups to reduce the risk of inflating correlation coefficients. The first number in a cell gives the Pearson correlation coefficient for the firstborn group; the second number is the value for the second-born group. Statistical significance is indicated as follows:

^{*} for $p < .05$
^{***} for $p < .01$.

Table 3

Correlation matrices for Sample 1 at HV3, presented separately for first- and second-born twins.

	LFV	MLU	NDW	NTW	IQ	CELF
LFV	1.00					
MLU	0.66 ^{**} , 0.49 ^{**}	1.00				
NDW	0.72 ^{**} , 0.69 ^{**}	0.86 ^{**} , 0.80 ^{**}	1.00			
NTW	0.62 ^{**} , 0.49 ^{**}	0.96 ^{**} , 0.97 ^{**}	0.86 ^{**} , 0.84 ^{**}	1.00		
IQ	0.48 ^{**} , 0.35 [*]	0.43 ^{**} , 0.14	0.50 ^{**} , 0.19	0.48 ^{**} , 0.18	1.00	
CELF	0.47 ^{**} , 0.49 ^{**}	0.30, 0.41 [*]	0.28, 0.36	0.22, 0.39 [*]	0.71 ^{**} , 0.72 ^{**}	1.00

Note: Sample 1 at HV3 included twin pairs, who were divided into firstborn and second-born groups to reduce the risk of inflating correlation coefficients. The first number in a cell gives the Pearson correlation coefficient for the firstborn group; the second number is the value for the second-born group. Statistical significance is indicated as follows:

^{*} for $p < .05$,

^{**} for $p < .01$.

The CELF variable is a composite of participants' standard scores on the two available expressive subtests, Recalling Sentences and Word Classes (Expressive).

Table 4
Correlation matrices for Sample 2, using both the original WERVE and the augmented WERVE.

	LFV density (original)	LFV density (augmented)	MLU	NDW	NTW	IQ
LFV density (original)	1.00					
LFV density (augmented)	0.86**	1.00				
MLU	0.50**	0.61**	1.00			
NDW	0.63**	0.65**	0.88**	1.00		
NTW	0.61**	0.65**	0.97**	0.89**	1.00	
IQ	0.24	0.38*	0.33*	0.49**	0.35*	1.00

Note: Statistical significance is indicated as follows:

*
for $p < .05$,
**
for $p < .01$.