


Clinical Focus

Using Free Computer-Assisted Language Sample Analysis to Evaluate and Set Treatment Goals for Children Who Speak African American English

Courtney Overton,^a Taylor Baron,^a Barbara Zurer Pearson,^b and Nan Bernstein Ratner^a 

Purpose: Spoken language sample analysis (LSA) is widely considered to be a critical component of assessment for child language disorders. It is our best window into a preschool child's everyday expressive communicative skills. However, historically, the process can be cumbersome, and reference values against which LSA findings can be "benchmarked" are based on surprisingly little data. Moreover, current LSA protocols potentially disadvantage speakers of nonmainstream English varieties, such as African American English (AAE), blurring the line between language difference and disorder.

Method: We provide a tutorial on the use of free software (Computerized Language Analysis [CLAN]) enabled by the ongoing National Institute on Deafness and Other Communication Disorders-funded "Child Language Assessment Project." CLAN harnesses the advanced computational power of the Child Language Data Exchange System archive (www.childes.talkbank.org), with an aim to

develop and test fine-grained and potentially language variety-sensitive benchmarks for a range of LSA measures. Using retrospective analysis of data from AAE-speaking children, we demonstrate how CLAN LSA can facilitate dialect-fair assessment and therapy goal setting.

Results: Using data originally collected to norm the Diagnostic Evaluation of Language Variation, we suggest that Developmental Sentence Scoring does not appear to bias against children who speak AAE but does identify children who have language impairment (LI). Other LSA measure scores were depressed in the group of AAE-speaking children with LI but did not consistently differentiate individual children as LI. Furthermore, CLAN software permits rapid, in-depth analysis using Developmental Sentence Scoring and the Index of Productive Syntax that can identify potential intervention targets for children with developmental language disorder.

Nearly all states require the use of converging evidence from standardized tests with observation and analysis of the child's language use in context (language sample analysis [LSA]) in order to determine eligibility for intervention services (Spaulding et al., 2012). This practice is also recommended by the American Speech-Language-Hearing Association Practice Portal, which provides guidance for both late language emergence and spoken language disorder. When used with mainstream language variety speakers of a language, LSA has excellent ecological validity;

it measures what children say in actual interactions, in contrast to performance elicited by test materials unable to mimic everyday communicative contexts. Paul and Norbury (2012) note that LSA "is more sensitive than standardized tests for identifying preschoolers with clinically diagnosed language delays, more effective for treatment planning and outcome monitoring, and a more valid reflection of the child's use of language in everyday contexts" (p. 301). In mainstream English speakers, LSA has also shown superior sensitivity to detect expressive language deficits than do standardized tests, which can only appraise isolated aspects of language skill in formal environments (Dunn et al., 1996). LSA can be used to examine both structural aspects of the child's language (e.g., grammar, vocabulary), as well as pragmatics (e.g., communicative intent, responsivity), and fluency.

While a useful adjunct to both clinical assessment and goal setting for children with expressive language concerns (Finestack et al., 2020; Gallagher & Hoover, 2020; Garbarino et al., 2020; Pezold et al., 2020), employing and

^aUniversity of Maryland, College Park

^bUniversity of Massachusetts, Amherst

Correspondence to Nan Bernstein Ratner: nratner@umd.edu

Editor-in-Chief: Holly L. Storkel

Received November 14, 2019

Revision received January 31, 2020

Accepted June 20, 2020

https://doi.org/10.1044/2020_LSHSS-19-00107

Publisher Note: This article is part of the Forum: Serving African American English Speakers in Schools Through Interprofessional Education & Practice.

Disclosure: The authors have declared that no competing interests existed at the time of publication.

interpreting LSA still face numerous obstacles. One is making LSA more user-friendly in everyday clinical contexts. While automatic speech recognition has made tremendous progress in the past few years that is easily appreciated in everyday interactions with “smart” devices, satisfactory progress has yet to be made in accurate automated transcription of child speech, especially when it is also characterized by misarticulation (Wu et al., 2019). Thus, clinicians are often concerned with the obstacle of making initial transcriptions. However, computer-assisted analysis can be very time-efficient in terms of the multiple measures it quickly provides to the clinician once the transcript is available (Heilmann et al., 2010). In this tutorial, we describe the use of an open-access computer program (CLAN KidEval) that can make transcription easier and faster than other available options, in the hope of demonstrating to clinicians that working with computer-assisted LSA is less daunting than they think it is and certainly less daunting than it has been in the past.

A major concern in LSA is to ensure that results are valid indicators of expressive language skill in both mainstream and nonmainstream English-speaking children. In this tutorial, we explain how substantial progress has been made on addressing the first challenge: ease of use. We show how clinicians can use free materials and programs within the Child Language Data and Exchange System (CHILDES; www.childes.talkbank.org) to perform time-efficient LSA for assessment and goal-setting purposes. Work is still in progress on the second concern: to tailor the typically used LSA measures that hold promise of being relatively variety neutral, so that they distinguish disordered language profiles in children who speak nonmainstream varieties of English such as African American English (AAE). Our preliminary analyses have yielded some insights about AAE and our computer-assisted measures. We provide case studies that show how the depth and breadth of the information provided by new computer-assisted LSA can assist speech-language pathologists (SLPs) in “dialect-informed” identification and goal setting for children who speak AAE.

LSA in Clinical Use: Options and Current Practical Limitations

Barriers to Use

Repeatedly, surveys of practicing clinicians show that, although recommended, LSA is seldom fully implemented in practice. Before a few years ago, LSA was very time-intensive to compute if done by hand or was perceived to be excessively time-consuming (e.g., Miller, 2009; Pavelko & Owens, 2017). Thus, many SLPs, when they computed LSA measures, typically only calculated mean length of utterance (MLU; Dunn et al., 1996; Finestack & Satterlund, 2018; Schuele 2013), regardless of child age, despite the fact that Brown (1973) himself, who introduced the measure, predicted that, after an average MLU of approximately 4.0 (generally around ages 3;6–4;0 [years;months] in typically developing children), it would not be highly informative regarding a child’s grammatical skills.

Lack of Robust Reference Scores

Currently available LSA normative values, whether automated or calculated by hand, are based on very small samples, almost exclusively from Mainstream American English (MAE)-speaking children. Thus, norms or comparison sets for LSA measures at this age are relatively weak, even for evaluation of MAE-speaking children. While tools such as Systematic Analysis of Language Transcripts (SALT; Miller et al., 2015) and a newer LSA algorithm, Sampling Utterances and Grammatical Analysis Revised (SUGAR; Pavelko & Owens, 2017), provide analysis tools for LSA to be used with older children, few language measures of any sort are robustly normed for analysis of adult-child play interaction from ages 2;6 to 6;0. The problems with small LSA reference cohorts can be especially concerning between the ages of 2 and 6 years, the time during which LSA is thought to provide the most useful adjunct information to standardized test performance.

Lack of Diversity in LSA Reference Samples

Numerous researchers have suggested that LSA measures are less biased against speakers of nonmainstream varieties than standardized testing (see summary in Horton-Ikard, 2010). However, we do not have a large body of data to show how LSA measures vary by gender, ethnicity, socioeconomic status, or the language variety spoken by the child. For example, both *procedures* and *reference data* for AAE-speaking children are extremely limited. As numerous researchers have noted (from Stockman, 1996, to Johnson & Koonce, 2018), using LSA procedures that assume MAE variety, young speakers of AAE can potentially be falsely identified as language-disordered.

Misidentification of young speakers of AAE may arise because numerous features of AAE overlap with observed behaviors in young children who are delayed learners of MAE. These include “zero-marking” (or “unpronounced”) MAE markers for third-person agreement, or tense and number, some of which are conditioned by the phonotactics of AAE (such as final cluster constraints). In AAE registers, auxiliary and copular forms that might be contracted in spoken MAE may not be realized (as in “He \emptyset going home”). However, AAE, like MAE, requires them in constructions in which contractions are disallowed (such as the truncated response, “Yes, he is”). Clearly, a language variety that permits optional realization of morphemes can depress MLU averages in language samples from its speakers. Likewise, the habitual verb form, unique in English to AAE (but present in other languages), may be signaled by use of the uncontracted auxiliary (as in “He be going to church every day”) or copula (“He be in the classroom before the bell”), which are often misinterpreted by mainstream speakers or viewed as an error. The list of features associated with AAE is quite long and varies regionally, and speakers vary in how much they use them. We cannot summarize all of them here, but fuller treatments can be found in Van Hofwegen and Wolfram (2010), among others. The more common features with impacts on assessments that SLPs are likely to encounter in their language samples can be found in Table 1 and are

discussed in a following section of this clinical focus article. Such features of AAE pose a real and clinically relevant concern for “dialect-fair” assessment of young AAE speakers. Clinicians run the risk of making one of two errors: overidentifying them by attributing a child’s typical patterns to disorder or underidentifying them by simply assigning all differences from MAE to language variety status.

A short review of LSA measures in clinical use

A number of proposals have emerged for “dialect-sensitive” LSA when working with child speakers of AAE. In order to appraise their relative merits, we first review the most popular measures that SLPs report that they use (Stockman et al., 2013), in the order in which they were developed for clinical use.

MLU, conventionally measured in morphemes, was proposed by Brown (1973) to measure growth in grammar, which, in very young children learning English, is heavily impacted by growth in use of a closed set of grammatical morphemes marking tense, agreement, and possession, among other features.¹

MLU has since been shown to be less good at mapping growth in highly inflected languages, in which children rarely hear or have reason to use bare word stems unmarked for gender, number, and person (contrast the absence of an overt morpheme in English *the dogs eat* compared with Spanish *los perros com-en*; one cannot use the root *com* in any environment). Thus, MLU measured in words (MLU-W) is often used to measure development in languages other than English, as well as when appraising older children, for whom development of grammatical inflections is presumed to be complete (see Rice et al., 2010). MLU only measures utterance length, rather than its internal structure, and, as we have noted, it has dubious value when used with children producing average utterances more than four morphemes long. Despite this, MLU remains the most frequently completed LSA measure by clinician self-report (Finestack & Satterlund, 2018).

The basic characteristics of AAE would predict that measures such as MLU, whether in morphemes or words, are likely to undercredit the speaker of a language variety that is characterized by optional zero-marking of both free and bound grammatical morphemes in some linguistic contexts (Horton-Ikard et al., 2005; Newkirk-Turner et al., 2014; Terry et al., 2013). Whether or not MLU-W empirically solves these problems is somewhat open to disagreement. As noted, there is controversy as to whether MLU-W is useful in distinguishing AAE-speaking children with and without language impairment (LI; Oetting, 2005; Oetting et al., 2013; Oetting & McDonald, 2001, 2002; Seymour et al., 1998; Stockman et al., 2016; but see also Horton-Ikard, 2010).

When we consider the role of MLU in evaluating children who speak AAE, computing MLU in words, rather than morphemes, can weaken the impact of optional zero-marking

of some grammatical morphemes and still tends to reveal large differences between typically developing and delayed language learners across language varieties (Oetting, 2005). However, MLU measures are less helpful in guiding clinicians toward potential therapeutic goals, since they focus on utterance length, rather than the structure or content of utterances.

Other measures that focus more strongly on structure or function than length are also promising for developing LSA strategies that will work well with different varieties of English. To better capture the internal structure of sentences, Lee et al. (1974) proposed a fairly complex measure of grammatical development called “Developmental Sentence Scoring” (DSS), for which she obtained reference values on a set of 160 children ages 3;0–6;11. Using a standard sample size of 50 eligible utterances (which must have, minimally, a noun and a verb), words that mark various aspects of grammar, such as negative markers, conjunctions, *wh*-words, and so forth, are awarded 1–8 points based on their expected sequence of emergence in typical child language acquisition (from early to later). Additionally, utterances that meet grammatical acceptability standards are accorded an extra “sentence point,” which is not awarded if the utterance appears to be ungrammatical. A recent, in-depth analysis of DSS shows that it has an acceptable level of success in differentiating MAE-speaking children with typical language development and specific LI (Souto et al., 2014). However, as Eisenberg and Guo (2013) note, the very criteria used to exclude eligible utterances (e.g., presence of a verb) may prevent analysis of a substantial proportion of the child’s output and thus underestimate grammatical impairment or limit its use for the most delayed child participants. Furthermore, the original rules proposed for DSS were based on grammatical constructions in MAE. Because DSS also relies on the clinician to award sentence points for grammaticality, it will also depend on the SLP’s sensitivity to AAE to avoid inappropriately withholding sentence points from utterances that are grammatical in AAE, but ungrammatical in MAE. Finally, DSS is extremely time-consuming when computed by hand: Long (2001) estimated that it took coders versed in the procedure anywhere from an hour to 75 min to complete, after a transcript had been made.

Whether or not DSS disadvantages child speakers of AAE is not yet clear, and many targeted elements in DSS do not overlap with features of AAE. Nonetheless, Nelson and Hyter (1990) have proposed adjustments to DSS scoring (Black English Sentence Scoring [BESS]) for use with children speaking AAE. We will be trialing this adaptation in the near future; however, for the purposes of this clinical focus article, we ask whether traditional DSS is a reasonable option for evaluating language samples from children who speak AAE.

The Index of Productive Syntax (IPSyn; Scarborough, 1990; see also Altenberg et al., 2018; Roberts et al., 2020) appears to be, both by its computational design and in practice, relatively language variety-fair, because it relies on examination of phrase structure, rather than use of specified grammatical elements or grammaticality judgments, as does DSS. In several studies, it has yielded generally equivalent scores in children who speak AAE and MAE, and it appears

¹A subset of these is popularly called “Brown’s morphemes” or Brown’s “Fourteen Morphemes.”

Table 1. Contrastive variable use item types in African American English.

Item types	Examples
Morphosyntax	
Variable zero marking	
Zero present tense copula is copula	<i>They ∅ tall.</i> <i>He ∅ a doctor.</i>
Zero present tense auxiliary is auxiliary are auxiliary	<i>He ∅ running.</i> <i>They ∅ sleeping.</i>
Zero plural /-s/	<i>two glass∅, ten cent∅</i>
Zero possessive /-s/	<i>John'∅ mother left.</i>
Zero regular past tense /-ed/	<i>He play∅ yesterday.</i>
Variable agreement	
Subject verb 3rd -s with <i>do</i>	<i>He don't like to swim.</i>
Subject verb 3rd -s with <i>have</i>	<i>She have no shoes.</i>
Subject verb 3rd -s with lexical verb	<i>He sleep∅ on a bed.</i>
(Note: Past copula or auxiliary are not optional)	<i>They was cold.</i> <i>They was sleeping.</i>
Syntax	
Multiple negation	<i>He don't have no shoes.</i>
Inverted negatives	<i>Don't nobody here know her.</i>
Noninverted question	<i>You know her name? How you knew that?</i>
Habitual <i>be</i>	<i>Bruce be running.</i> ("Bruce usually runs"—but is not running now.)
Phonology	
Final consonant clusters	<i>test</i> pronounced [tɛs]
Voiced interdental fricatives /ð/ in any position	<i>this</i> pronounced [dɪs]; <i>breathe</i> pronounced [briv]
Voiceless interdental fricatives /θ/ in final position	<i>moth</i> pronounced [maf]

to discriminate between groups of children speaking both varieties who are typically developing and language impaired (LI; Horton-Ikard, 2010; Oetting, 2005; Oetting et al., 2010). For those structures that do contrast, there have been proposals to adapt IPSyn (Oetting et al., 2010) for AAE by removing or substituting for any contrastive features. Like DSS, the IPSyn can be very time-consuming, even when conducted with trained coders: Long (2001) found it took very similar times (more than a mean of an hour, depending upon sample length).

Useful LSA measures have been found in other language domains as well, for example, lexical diversity within a child's spontaneous language sample (Oetting et al., 1999; Stockman & Vaughn-Cooke, 1986). Lexical diversity has been measured using numerous options. Templin's (1957) type-token ratio (TTR) was designed to create a proportion of the number of unique word types (as measured by the uninflected root form; thus, *play*, *plays*, *playing*, and *played* all represent one type) over the total number of word tokens in a sample. Potential values can range from .01 (a single word uttered repeatedly 100 times) to 1.0 (which could not be natural language, since we need to repeat function words when constructing acceptable utterances).

From the outset, although norms for TTR could be obtained and show growth over early childhood (Miller & Chapman, 1981), there have been concerns that varying sample size badly impacts the derived values (e.g., Hess et al., 1986). This concern has not filtered down to clinical practice: Finestack and Satterlund (2018) found

TTR to be the only "lexical" computation used by practicing clinicians in their LSA protocols, and it was not reported to be used frequently.

Watkins et al. (1995) suggested an alternative measure, number of different words (NDW), which requires selecting a standard subsample size of exactly 100 words. NDW has been found to differentiate children with LI from their typical peers (Hewitt et al., 2005; Watkins et al., 1995); it also appears to work well in this regard with child speakers of other languages (Auza et al., 2018). A final measure that has been proposed to assess lexical diversity in LSA is vocabulary diversity (VocD; Durán et al., 2004; McKee et al., 2000). This procedure uses a random resampling algorithm to estimate the profile of lexical variability in a sample. To date, it can only be employed using CLAN software; additionally, because it produces an estimate of variability rather than a fixed score, its use has been for experimental, rather than clinical, evaluation of clinical populations (e.g., Owen & Leonard, 2002; Silverman & Bernstein Ratner, 2002).

Before leaving this overview, we wish to point out that, although vocabulary measures would seem to be linguistically neutral, they introduce other confounds. Lexical diversity measures are highly vulnerable to socioeconomic variation (Golinkoff et al., 2019; Rowe, 2008), which can covary with nonmainstream language variety use.

The Promise of Technology

As noted, beyond the computation of MLU in words, the LSA measures that can provide the most guidance to

SLPs in establishing therapy goals (e.g., IPSyn, DSS; see Gallagher & Hoover, 2020; Pezold et al., 2020) have historically required substantial transcription, coding, informed selection of eligible grammatical constructions or elements, and lengthy computation. Thus, while recommended, they have not necessarily been reasonable options for the SLP in typical practice. As Stockman et al. (2016) note,

IPSyn ought to be useful for identifying AAE speakers who need a clinical evaluation. However, the IPSyn's detailed taxonomy for assessing morphosyntax makes it an unlikely candidate for use as a brief screening tool... it requires a search for 60 specific forms. Some of them may be unfamiliar to many clinicians (e.g., bitransitive predicate, propositional complement)... IPSyn requires considerable time beyond the time needed to collect and transcribe a language sample. (p. 93)

Stockman et al. went on to note that, during their research, use of a specialized Linux-based software program dedicated solely to IPSyn analysis (Hassanali et al., 2014) only partially solved these problems. The software, which is not widely available, was about 84% in agreement with manual coding performed by IPSyn developers (Altenberg & Roberts, 2016).

Although clinicians now have numerous options to perform LSA using computer assistance, such as SALT and SUGAR, the current discussion concentrates on the use of Computerized Language ANalysis (CLAN; MacWhinney, 2000) because it is free, easily accessible online, has exceptional computational capacity, and offers utilities that enhance its ease of use, especially the ability to link the transcription to the original audio or video through a utility that allows "listening while typing." This permits the clinician to revisit numerous aspects of the child's speech, language, prosody, and fluency for assessment, therapy planning, or monitoring. The CLAN automatic parser, which is able to label targeted forms and grammatical structures, can provide the benefits of IPSyn, for example, without the substantial commitment of SLP time and linguistic knowledge that Stockman said was "unlikely." Finally, the breadth and depth of data collected and analyzed in the CHILDES system since the 1980s has the potential to greatly increase the reference databases to provide customized benchmarking for children from many language backgrounds at almost any age. Thus, once a simple transcript is made, the clinician is able to generate numerous LSA values and benchmark them.

For this clinical focus article, we report steps taken thus far to leverage the Talk Bank Project (www.talkbank.org) to facilitate and improve free, computer-assisted LSA for grammatical, lexical, and potentially fluency measures. In the process of improving the analyses, we also aim to extend the usefulness of numerous traditional clinical measures by beginning to compile group-specific benchmarks, if not norms, for performance at age-graded intervals. Furthermore, the rapid automation of media-linked transcripts and analyses for larger numbers of children with different demographics, language status, and language community

permits evaluation of the available LSA measures for distinguishing children with LI and typical development, for establishing therapy goals among diverse groups and for monitoring progress toward them.

The Child Language Assessment Project

For the Child Language Assessment Project (CLASP) researchers from the University of Maryland, Carnegie-Mellon University, University of Massachusetts Amherst, and University of Houston² have been funded for 5 years by the U.S. National Institute on Deafness and Other Communication Disorders to address many of the problems discussed above. In particular, CLASP is charged to place a special focus on child speakers of AAE, a variety of English that is a special challenge for standardized measures of all kinds (Craig & Washington, 2000; Oetting, 2005; Seymour & Pearson, 2004; Stockman, 1996, among others). In the process of the research, CLASP is also improving benchmark data for children with AAE background who are learning Mainstream English as a second language variety.

This tutorial has been created to show how the free, open-access CLAN software is being modified to tailor its programs for practicing SLPs to use with children of diverse backgrounds and to review its current status and limitations. The topics in this tutorial are as follows:

1. introducing how CLAN works and how it can be used in everyday clinical practice to generate extensive spoken language profiles relatively easily and quickly using a CLAN routine called KidEval;
2. testing whether current KidEval outputs can distinguish language disorder among a group of AAE-speaking children retrospectively identified as LI and typically developing, as well as identify areas for improvement of the computer program; and
3. showing how detailed analyses in CLAN KidEval, such as DSS and IPSyn, can be used to inform both diagnostic decisions for individual children, as well as identify reasonable intervention goals for children with LI who appear to be speakers of AAE.

How to Use CLAN and How CLAN Can Help SLPs

Automatic parsing. Automatic parsing is the key to faster analysis of grammatical and lexical features of child speech. As Stockman et al. (2016, p. 93) pointed out, huge barriers to DSS and IPSyn are clinician time and linguistic expertise. Unlike other programs in current use, CLAN's computerized grammatical parsers for a number of languages (e.g., Spanish, Mandarin, Japanese, and seven other languages) remove the need for users to hand tag the constituents of interest. One CLAN command line assigns grammatical class and syntactic function to text that is typed in using

²Issue guest editor Monique Mills and contributors Nan Bernstein Ratner, Jan Edwards, and Barbara Zurer Pearson are funded by this grant (1R01DC016076-01); Brian MacWhinney (Carnegie-Mellon University), TalkBank administrator, is also a contributor to this project.

regular English writing conventions. Utterance-internal punctuation is not necessary, contractions are permitted, and use of capitalization is limited. Examples are shown in Figure 1. This contrasts with SALT and SUGAR, for which the clinician must parse the morphological affixes (e.g., in making the transcript, “splitting” words into component morphemes, such as *cat-s*, *go-ing*).

CLAN’s English automatic morphological tagger is currently rated as between 95% and 98% accurate (Sagae et al., 2010) for analysis of adult data. When analyzing child data, its accuracy falls somewhat, ranging from 76% to 94%, performing worse primarily on highly telegraphic (agrammatic) speech (a language development stage for which clinicians might choose not to go beyond computation of MLU). For children at any higher levels of language development, accuracy level appears to be above 90%. To provide some context to these numbers, we know, unfortunately, that SLPs’ abilities to grammatically “tag” English sentences were judged to be only an average of 70% accurate in the one study we located that has examined this skill among students in our field (Justice & Ezell, 1999). Thus, the CLAN parser is at least as accurate as manual computation in most contexts, and it does in seconds and minutes what can take many hours by hand across a clinician’s caseload.

Ease of use, “walker controller,” and other transcription aids. We realize that the myriad options available through CLAN are sometimes perceived as “complicated,” and indeed, the program was built for flexibility and broad analytical power. CLAN was originally designed for language researchers and is only now being streamlined for clinical use by the current grant funding. For this project, we have the goal of simplifying clinical LSA using a small set of commands to create the input and generate output to match SLPs’ typical needs in their daily practices.

CLAN links the audio or video record to the transcript with one or two clicks of the mouse to pull up the “walker controller.” This utility provides automated looping as one listens, so users can hear the utterance as many times as desired, while typing. It greatly speeds transcription by approximately a factor of 3 from conventional “play, then type” approaches (Newman et al., 2016), aids transcription accuracy, and provides a more nuanced permanent record of the client’s sample, including phonology, rate, nonverbal behaviors, and so forth. One can use the linked tools to facilitate reviewing the finished transcript for reliability and for pulling out short video clips to focus on a particular characteristic of the child’s speech or for demonstration purposes (samples of linked transcripts that can be played without having the CLAN program installed can be viewed at <https://childes.talkbank.org/browser/>).

Once the audio or video is transcribed by the user, the transcript can be sent to the automatic parser, and almost instantaneously, it is tagged for part of speech by MOR (for “morphemes”) and grammatical relationships (GRA). With the information added in the “mor” and “gra” lines, the parser can then compute in milliseconds multiple measures that are brought together under the umbrella of the KidEval program. The SLP can analyze children individually

or in large groups (e.g., end of term progress reports). When analyzing a single child, the user has the option to select a comparison database from the software’s library, and the output of the child’s language sample scores for various measures includes how the child’s scores differ from the relevant database mean (by gender and age, in 6-month intervals). Thus, this largely automated LSA provides important components of an individual child’s clinical profile to inform diagnostic decisions, as well as identify potential intervention goals for children who present with LI.

Getting Started With CLAN and Accessing the KidEval Program

To obtain the CLAN programs, users go online to TalkBank.org and download the software for PC, Mac,³ or Linux systems. All programs and manuals are free, as is other user support through CHILDES (childes.talkbank.org). There is also a brief guide to CLAN for SLPs (Bernstein Ratner et al., 2019). Video tutorials for transcription and for use of the KidEval and related programs (e.g., IPSyn and DSS) are available at the TalkBank screencasts site (<https://talkbank.org/screencasts/>), as well as a recently developed set of four videos that specifically address how SLPs can run the analyses we discuss in user-friendly doses (<https://talkbank.org/screencasts/OSLP>). Finally, there is a user group for trouble-shooting obstacles or problems at <https://groups.google.com/forum/#!forum/chibolts> (one can subscribe or browse archived topics).

Creating a Transcript in CLAN

As noted above, what the child says is typed in conventional English orthography, one utterance per line, with formatting assistance provided by the CLAN transcription program. The transcriber does not have to identify or mark individual morphemes or grammatical structures (see Figure 1 for a basic transcript). Only proper names and the word “I” are capitalized, and only utterance-final punctuation is allowed. Transcribers are, however, encouraged to note unrealized/“missing” forms that would be expected in MAE (e.g., “*the boy 0is going*” for “*the boy going*”) to facilitate the parser. The transcriber should flag any utterance that is ungrammatical in the child’s native language variety (using the traditional linguistic marker [*] anywhere on the transcript line) to enable proper computation of DSS.⁴ It is important to note here, as in other contributions to this issue, that a varietal form which is acceptable in the AAE variety, typed as “*the boy 0is going*” for “*the boy going*,” would *not* receive an error mark.

The standard format for a transcription in CLAN (the “CHAT” guidelines) includes a brief header and an @End

³At press time, CLAN was still upgrading to the new Catalina operating system. It should be available soon; until then, Mac users will need to run these programs on computers running earlier operating systems.

⁴Use of AAE utterances judged as grammatical will be used to train our parser and the in-progress “dialect detector.”

Figure 1. A basic transcript before the Computerized Language Analysis MOR program performs grammatical analysis.

```
43 *CHI: once upon a time there was a boy who wanna get a balloon .
44 *INV: mhm .
45 *CHI: he got a balloon .
46 *CHI: he ask for it .
47 *INV: go ahead .
48 *CHI: and then it popped .
49 *CHI: what are those ?
50 *CHI: he cried .
51 *CHI: he got another one .
52 *CHI: he was trying to get it, but he didn't have no money .
53 *INV: can you tell me what's going on here ?
54 *CHI: he wanna get a balloon .
```

symbol. Table 1 shows a few lines from the middle of a transcription for a participant in the study we will discuss.

Preparing the Transcript for Analysis

Once a child's sample has been transcribed, it is ready to be prepared for analysis. A single command (MOR⁵) inserts two new lines of code automatically under each child utterance. The first is a morphological tagging line (%mor) that identifies word roots and inflections. The second, %gra tier, notes syntactic dependencies that have been computed based on the tagged elements in %mor. (This is the analysis that enables the DSS and IPSyn computations.) Figure 2 shows the application of the MOR command to a portion of the transcript in Figure 1.

We see in Figure 2 that we have taken the perfectly legible transcript in Figure 1 and made it pretty much unreadable for ourselves, but the user does not routinely use these lines—the computer does. For legibility, one can hide tiers in CLAN, so even after MOR has been run, we could print out the transcript in Figure 2 to look like Figure 1. Or we might choose to omit the Investigator (*INV) and print just the child lines.

The program can now search automatically in the %mor and %gra output lines in order to compute multiple LSA measures. In %gra, for example, in Line 129, we see “6|SUBJ” and “6|0|ROOT” (or verb). This tells the program that the child utterance in Line 125 has a subject and a verb, that is, it is potentially a sentence eligible for DSS, as are those in Lines 134, 137, and 144. The %mor line is a little more transparent. Many grammatical measures (such as the IPSyn and DSS) seek to find child use of specific forms, such as a past copula—and there it is in Line 126, labeled “cop|be&PAST&13S.” In this case, the notation “cop” stands for copula; the word *was* is tagged as be + & (irregular) PAST first- or third-person singular (13S). In the %mor in Line 145, we see that “popped” in Line 144 is also a past tense verb and has the regular past *-ed* suffix. Looking ahead to getting a DSS score for Line 144, we see that the computer will know to credit the word “it” in that

line as an impersonal pronoun and award a sentence point for the utterance as well, since it has not been flagged as ungrammatical by the SLP.

How well MOR parses child language spoken by speakers of a minority variety is a focus of our ongoing project and is now being tested for the first time. We note again that the parser appears to function at least as accurately as SLP transcribers (Justice & Ezell, 1999) and probably more accurately (Sagae et al., 2010), and the program continues to be upgraded but can make sporadic mistakes. Should the SLP notice a misparse and wish to correct it, this line can be edited like any other text document to replace part-of-speech codes or other errors.

Using the CLAN Outputs: KidEval, DSS, and IPSyn

To get a general profile of the child's level of language relative to their typical peers, KidEval is a “one-step analysis” that includes many subprograms that follow current best practices for LSA measures. It calculates their summary scores for the child and outputs them all on one line. The program gives the option to select a reference database, and the output will display how many standard deviations (or fractions of a standard deviation) the child is above or below the mean.

Figure 3 shows a sample KidEval report for a child who was studied during the norming process for the Diagnostic Evaluation of Language Variation (DELV; Pearson et al., 2014; Seymour et al., 1998, 2005). (The Excel output has been formatted, and its columns were wrapped to fit on a single page for easier viewing.) The child's LSA values are indexed against an age-matched cohort drawn from the growing CHILDES North American English (NAE)-speaking database. For each variable, one will see the corpus mean, standard deviation, and number of observations. Values that fall beyond 1.0 or 1.5 SDs from the mean for the child's age bracket (within 6 months) are flagged with asterisks. Examples of the child's use of Brown's 14 morphemes are simply tallied and not norm-referenced, as context will determine whether or not a child uses certain tenses, prepositions, or possessives, for example. A legend that spells out and defines the column names for the LSA variables is provided.

⁵CLAN is distributed for free use in over a dozen languages. Users customize by downloading the appropriate grammar, in this case, English, by selecting grammar from the File menu.

Figure 2. The child’s transcript after application of the English MOR analysis program. Two new lines (%mor and %gra) are automatically added to the transcript without user input. These tiers enable the KidEval, Developmental Sentence Scoring (DSS), and Index of Productive Syntax (IPSyn) programs to run.

```

125 *CHI: once upon a time there was a boy who wanna get a balloon .
126 %mor: adv|once prep|upon det:art|a n|time adv|there cop|be&PAST&13S
127 det:art|a n|boy pro:rel|who v|want-inf|to v|get det:art|a n|balloon
128 .
129 %gra: 1|6|LINK 2|6|JCT 3|4|DET 4|2|POBJ 5|6|SUBJ 6|0|ROOT 7|8|DET 8|6|PRED
130 9|10|LINK 10|8|CMOD 11|12|INF 12|10|COMP 13|14|DET 14|12|OBJ 15|6|PUNCT
131 *INV: mhm .
132 %mor: co|mhm=yes .
133 %gra: 1|0|INCRROOT 2|1|PUNCT
134 *CHI: he got a balloon .
135 %mor: pro:sub|he v|get&PAST det:art|a n|balloon .
136 %gra: 1|2|SUBJ 2|0|ROOT 3|4|DET 4|2|OBJ 5|2|PUNCT
137 *CHI: he ask for it .
138 %mor: pro:sub|he v|ask prep|for pro:per|it .
139 %gra: 1|2|SUBJ 2|0|ROOT 3|2|JCT 4|3|POBJ 5|2|PUNCT
140 %com: ask pronounced as(k)
141 *INV: go ahead .
142 %mor: v|go adv|ahead .
143 %gra: 1|0|ROOT 2|1|JCT 3|1|PUNCT
144 *CHI: and then it popped .
145 %mor: coord|and adv:tem|then pro:per|it v|pop-PAST .
146 %gra: 1|4|LINK 2|4|JCT 3|4|SUBJ 4|0|ROOT 5|4|PUNCT
147 %com: then pronounced [den]
148 *CHI: what are those ?
149 %mor: pro:int|what cop|be&PRES pro:dem|those ?
150 %gra: 1|2|SUBJ 2|0|ROOT 3|2|PRED 4|2|PUNCT
151 %com: those pronounced [dose]
152 *CHI: he cried .
153 %mor: pro:sub|he v|cry-PAST .
154 %gra: 1|2|SUBJ 2|0|ROOT 3|2|PUNCT
155 *CHI: he got another one .
156 %mor: pro:sub|he v|get&PAST qn|another pro:indef|one .
157 %gra: 1|2|SUBJ 2|0|ROOT 3|4|QUANT 4|2|OBJ 5|2|PUNCT
158 *CHI: he was trying to get it, but he didn't have no money .
159 %mor: pro:sub|he aux|be&PAST&13S part|try-PRESP inf|to v|get pro:per|it
160 cm|cm conj|but pro:sub|he mod|do&PAST-neg|not v|have qn|no n|money .
161 %gra: 1|3|SUBJ 2|3|AUX 3|0|ROOT 4|5|INF 5|3|COMP 6|5|OBJ 7|3|LP 8|12|LINK
162 9|12|SUBJ 10|12|AUX 11|10|NEG 12|3|CJCT 13|14|QUANT 14|12|OBJ 15|3|PUNCT
163 %com: trying to pronounced [tryna]
164 *INV: can you tell me what's going on here ?
165 %mor: mod|can pro:per|you v|tell pro:obj|me pro:int|what-aux|be&3S
166 part|go-PRESP prep|on n|here ?
167 %gra: 1|3|AUX 2|3|SUBJ 3|0|ROOT 4|3|OBJ 5|7|LINK 6|7|AUX 7|3|COMP 8|7|JCT
168 9|8|POBJ 10|3|PUNCT
169 *CHI: he wanna get a balloon .
170 %mor: pro:sub|he v|want-inf|to v|get det:art|a n|balloon .
171 %gra: 1|2|SUBJ 2|0|ROOT 3|4|INF 4|2|COMP 5|6|DET 6|4|OBJ 7|2|PUNCT

```

Let us “walk through” a sample output from a child diagnosed with an LI from the DELV norming project. The Excel output starts at the top left with column labels for the child’s unique ID, age in months, gender, and an optional code for parent education level. The following columns display the total number of child utterances and MLU, calculated in four ways—in words and in morphemes for the whole corpus and also standardized to 100 utterances. Next are the lexical diversity measures, starting with the number of word tokens and word types (based on the word root or lemma), followed by the TTR. Here, too, the clinician gets a choice of which information about the number of words will best suit her needs: getting a ratio of how many types over how many tokens—for the full corpus, or standardized to NDW per 100 words, or she can use the VocD (Lai & Schwanenflugel, 2016) measure. Subsequent columns score aspects of sentence structure, such as average number

of verbs per utterance, DSS, and IPSyn. The last items in Figure 3 are the “14 morphemes” from Brown’s (1973) seminal study of the earliest stages of language development in children. Clinicians will find them informative or not according to the age or linguistic development of the child. We note that some of these morphemes are much less frequent in the speech of children who speak AAE, such as contractible auxiliaries and copulas.

The clinician now has a profile encompassing information from several domains of language form and content. This child’s lexical measures appear only minimally depressed, but grammar measures seem quite low, judging from the negative standard deviations: more than 1 *SD* below the mean for MLU computed in either words or morphemes, -1.7 for verbs per utterance, -1.6 for the DSS, and more than 3 *SDs* from the mean for the IPSyn. The KidEval report thus paints a consistent picture of relative

Figure 3. Sample KidEval output (edited and wrapped to fit page margins). If the child's sample does not contain sufficient eligible utterances for an analysis, N/A (not applicable) will be displayed. Columns not relevant to the clinical discussion have been hidden to improve readability. Legend: Row 1 = LSA measure; Row 2 = child's score; Row 5 = mean database value (for age and gender); Row 6 = database standard deviation; Row 7 shows the number of eligible comparison files for age and gender (number may vary by sample length required to compute measure). Asterisked values lie more than 1 SD (*) or 2 SDs (**) from mean in database reference group. As noted in text, DDA's samples produced numerous "flags," while TFO's samples generated none, as shown in Table 2, which compares a child with language impairment (LI) who speaks African American English (AAE) with a typically developing AAE-speaking peer. AA = African American; MLU = mean length of utterance; Types = unique words; Tokens = all words in sample; TTR = type-token ratio; VocD = vocabulary diversity; DSS utterances = number of utterances qualifying for DSS analysis; DSS = total DSS score; IPSyn utterances = number of utterances qualifying for analysis; IPSyn total = computed IPSyn score; *PRES-P = present progressive; *-PL = regular plural; irr-PAST = irregular past; *-POSS = regular possessive; u-cop = uncontractible copula; det:art = determiner: article; *-PAST = regular past tense; *3S = regular third-person singular; irr-3S = irregular third-person singular; u-aux = uncontractible auxiliary; c-cop*/c-aux* = contractible copula and auxiliary.

Name/ID	Age (Month)	Sex	Group	Ethnicity	Education	Total_Utts
50DDA.cha	61 (5;1)	male	LI	AA	0-11	266
.	0.74

Grammatical Measures					
Variable	MLU_in	MLU_in	MLU100_	MLU100_	Verbs per
Name	Words	Morphemes	Words	Morpheme	Utterance
value	2.30	2.53	2.23	2.37	0.37
±SD from	-1.32	-1.40	-1.39	-1.49	-1.65
group mean
DB Mean (SD)	4.3(1.5)	4.3(1.7)	4.39(1.6)	4.34(1.7)	0.81(0.27)
# of files	69	69	48	48	69

Lexical Measures				
FREQUENCY	FREQUENCY	FREQUENCY	#DIFFERENT	
_types	_tokens	_TTR	WORDS/100	VOCD_D
174	644	0.27	48	50.33
-0.2	-0.2	-0.2	-1.0	-0.3
.
187.8(76)	766.2(671.5)	0.29(0.08)	54.2(6.2)	52.8(8.6)
69	69	69	69	69

Syntax Scoring			Fluency Measures		
DSS_	IPSyn_				
Utterances	DSS	Utterances	IPSyn_Total	retracings	repetitions
50	5.32	100	64	27	15
.	-1.61	.	-3.10	2.51	1.73
.	.	.	**	**	.
50	9.4(2.6)	100	93.5(9.5)	7.17(7.9)	4.96(5.8)
46	46	36	36	69	69

Brown's "14" Morphemes (reference means not applicable)						
*-PRES-P	in	on	*-PL	irr-PAST	*-POSS	u-cop
12	8	11	9	8	0	4
det:art	*-PAST	*-3S	irr 3S	u-aux	c-cop	c-aux
78	4	0	7	2	2	4

weakness in the child's grammatical development compared to age matches in the CHILDES database who have been analyzed using the same computer software. At this point, just using the currently available reference group, his scores place him at the bottom quartile. However, as we discuss next in case studies, we can run the individual DSS and IPSyn algorithms separately to obtain more detail about structures that were present or absent in the child's sample. Beyond diagnosis, they can reveal more information about used and absent forms that we can use for setting therapy goals.

For comparison purposes, we provide an abbreviated summary of major KidEval measures for this child (50DDA) and a typically achieving child from the same cohort of AAE-speaking children (see Table 2). Both are speakers

of AAE, but the child with LI scores below the database mean on virtually all measures, while the typical peer, even when benchmarked against a primarily MAE-speaking reference database, scores within typical limits.

Looking More Closely at a Child's DSS or IPSyn Profile

To look more closely within these grammatical measures, we go back to the command window of CLAN and run the programs that we want to focus on individually using the same input file that was already improved by MOR analysis. We choose first the DSS program from the drop-down menu and run it and then repeat the same command line

Table 2. Comparison of major language sample analysis measures that contrast performance of a child with language disorder (50DDA) and a typically developing child (07TFO), both speakers of African American English.

Age Gender Ethnicity Parent education	Child 07TFO		Child 50DDA		Reference database		
	61 months		61 months		North American English speaker files matched for gender and age (± 6 months)		
Measure	Score	± SD	Score	± SD	M	SD	No. files
No. MLU utterances	171		253		158	133.02	69
MLU words	4	-0.27	2.3	-1.32*	4.3	1.49	69
NDW 100	59	0.78	48	-1*	54.2	6.15	69
Verbs per utterance	0.75	-0.21	0.36	-1.65*	0.8	0.27	69
DSS	7.76	-0.65	5.32	-1.61*	9.43	2.55	46
IPSyn	96	0.26	64	-3.1**	93.5	9.5	36

Note. MLU = mean length of utterance; NDW 100 = number of different words per 100 words; DSS = Developmental Sentence Scoring; IPSyn = Index of Productive Syntax.

*Score more than 1 SD from mean of comparison group. **Score more than 2 SDs.

using IPSyn. In seconds, the programs run, and the output tables are in our CLAN work folder. When run separately, rather than within KidEval, we do not just get a total; there is a grid organized to resemble the manual version of these procedures.

DSS

DSS (see Figure 4) looks at nine categories, chosen by Lee et al. (1974) based on a review of the development of closed-class (grammatical) words over early childhood, which show a fairly consistent pattern of stepwise emergence from 2 to 7 years of age. Points are awarded for each instance of a word or phrase representing a given level of difficulty based on order of emergence. When contrasted with IPSyn, which we discuss next, DSS is somewhat more straightforward to program for computer analysis, since it is highly dependent upon search for individual words or contingent sequences (e.g., an instance of *neither* followed within a given utterance by *nor*, or the more complex verb phrase (VP) structures detailed below). For example, the first column, IP (impersonal pronouns), has four levels: demonstrative pronouns *it*, *this*, and *that* receive 1 point, indefinites such as *something* or *somebody* are accorded 3 points, while the negative counterparts such as *nothing* or *nobody* have been observed to emerge a little later in typical development and so are awarded 4 points. The highest level in this category has been defined as use of *anybody* and *everybody*, as well as a set of common terms that are conceptually more difficult, such as *first*, *last*, *both*, or *few* for 7 points each.

The most central part of a sentence, in Column 3, MV (the main verb or its “root”) is also the most nuanced category in the DSS, going through more stages and adding MV points for each occurrence in a sentence: 1 point for a verb without inflections (*they sit*) or *is* auxiliary or copula; 2 points for *am* or *are*, regular or irregular past tense, and a verb in the third-person singular; 4 points for simple modals (*can*, *will*) and “do + verb”; 6 points for past modals, *could*,

should, or *does* or *did* + verb; and 8 points (the highest possible) for combinations of auxiliary verbs, such as *may have eaten*, *should have been sleeping*. The final “category” is the sentence point to acknowledge that there may be some developmentally mature constructions that are not given credit in this schema, so if there are no “errors” in a sentence, it is given a “sentence point.”

Figure 4 shows portions of a DELV participant’s DSS scores, line by line in nine categories. There is a row-wise total for each sentence and a column total for all the subcategories, plus the total score. The child’s overall DSS score at the bottom of the table (the average number of points per sentence) and the mean and quartile values for the child’s age bracket are available (Lee et al., 1974) and reprinted in the SLP Guide to CLAN as well; however, the child’s final score was also benchmarked against their age- and gender-matched peers in the KidEval output produced in the prior step. Below, as we show, there is one line for each utterance, but in the interest of space, we show the top labels and the bottom totals, with only a few illustrative lines of the matrix filled in. To remind readers, the penultimate column on each line is the one that shows whether or not the utterance has qualified for a sentence point because it is grammatical in the child’s ambient language variety.

DSS Scoring and Interpretation

Lee et al. (1974) gives detailed directions for scoring. We will point out two or three issues that are raised in this excerpted transcript and DSS table. We note at the outset that this child is from an AAE-speaking community. Therefore, the transcriber did not mark the so-called “double negative” (*I don’t see no cat*) as an error. It is, in fact, a fairly sophisticated sentence and earned 13 points. Similarly, “How you make it?” was not marked in the transcript as an error, because unreversed questions are standard in adult AAE (Green, 2002). On the other hand, *me and my mom went pool* (because of its missing prepositional phrase for “went [to the] pool”) is

Figure 4. Portion of Developmental Sentence Scoring (DSS) analysis of boy with language impairment who speaks African American English (AAE). IP = impersonal pronouns; PP = personal pronouns; MV = main verb; SV = secondary verb; NG = negation; CNJ=conjunctions; IR=interrogative reversal; WHQ = *wh*-questions; S = sentence point (for grammaticality); TOT = total points for the utterance.

Developmental Sentence Analysis										
Child 08JYO 5;7 (67 mos) Male, low-SES, from AAE Language Community										
Sentence	IP	PP	MV	SV	NG	CNJ	IR	WHQ	S	TOT
and Reggie was fightin(g) .			2			3			1	6
Briana kicked me and then I kicked her back .		1 1 2 2				3			1	12
<I> [//] I my next birthday came up and I went to ah Kiddywood .		1 1 2 2				3			1	11
<it's a &po> [//] me and my mom went pool .		1 1 2				3			0	7
.										
.										
I don't know what that is .	1	1 6 4 1		4				2	1	21
where's the cat ?			1				1	2	1	5
I don't see no cat .	3	1 1 4		4					1	13
how you make it ?	1	1 1 1				?		5	1	9
<we, we made it in art> [<] .		3 3 2							1	9
TOTAL		20 59 117		6 24 41		6 2		325		

Developmental sentence score: 6.50 < 10th percentile
8 of 50 sentences.

considered an error in both AAE and Mainstream English; we do note, however, that the phrase “me and my mom” is correct in both AAE and many varieties of American English. DSS does not specifically score prepositional phrases the way it does verbs and negatives, but Lee asks us to award or withhold the “sentence point” (S), and the program withholds it in this case because we marked it as an error during transcription.

IPSyn

IPSyn (Scarborough, 1990; see updates in Altenberg et al., 2018; Roberts et al., 2020) is of particular interest for this project. It focuses on structures that appear to emerge rapidly during the lower end of the current project’s age range, having been developed for children between 2 and 4 years. It showed its steepest developmental trend earlier, rather than later in development, in a sample of more than 600 children (Bernstein Ratner & MacWhinney, 2016). In contrast, Kemper et al. (1995) found that IPSyn showed a developmental trajectory for its features even later in children’s language development, until approximately 8 years of age. Scoring the IPSyn takes a different approach to measuring grammatical complexity from the DSS. It is based on phrase structure attributes first outlined by Miller (1981) and counts “types,” not “tokens,” as DSS does. Given a sample of 100 eligible utterances, the scoring method awards points for up to two exemplars of 60 phrase and sentence structures. For instance, in noun phrase (NP) development, it examines whether the child uses bare nouns, article + noun, article +

modifier + noun, and so forth. The goal is to find evidence of the child’s having learned the structure, whether it is prescriptively “correct” or not. Scarborough makes that point clear by using examples like *mans* and *grapeses* in the coding manual for counting a plural noun. These are not instances of AAE, but they are relevant because they show clearly that the concept of pluralization, whether applied correctly or not, is the focus of the procedure. Thus, if a child were to use *ain’t*, for example, it could qualify as a point for different ways of negating a verb from simple negation (Q3) to a negative morpheme between subject and verb (Q5) to its potential use in a tag question.

IPSyn Produces a Grid Organized by NP, VP, Questions and Negations, and Sentences

If scoring manually, for each item, the clinician briefly notes up to two exemplars that satisfy the item and the line numbers in the transcript where they were found. CLAN does this, as well, and outputs the element in the %mor line that corresponds to the target. Subscores are given for the four categories and a total score, each of which can be compared to the analogous value for the reference group. IPSyn output is somewhat complex, given the many structures it targets. We provide a spreadsheet in Table 3 that shows for which structures in the category of VPs the child gained points and which were missing, as well as the subtotals and total score.

As a reminder, the IPSyn total and relative performance when compared with their gender- and age-matched peers was provided in the KidEval summary; the intent in

Table 3. The verb phrase subtable of the Index of Productive Syntax (IPSyn) analysis for the child shown in Figure 4, conducted using Computerized Language Analysis (CLAN).

Rule	Verb line	Utterance	Phrases		Points
			MOR		
V1	Verb	140	fit	vfit&ZERO	1
		246	know	vlknow	1
V2	Particle or preposition	366	on	preplon	1
		659	with	preplwith	1
V3	Prepositional phrase	366	on the house	preplon det:artlthe nlhouse	1
		659	with the bread	preplwith det:artlthe nlbread	1
V4	Copula linking two nominals	581	I'm a little kid	pro:subll ~coplbe&1S det:artla adjllittle nlkid	1
V5	Catenative preceding a verb	—		(not observed)	(0)
V6	Auxiliary be, do, have in verb phrase	939	I'm gonna	~auxlbe&1S partlgo-PRESP	1
		1026	he is walking	auxlbe&3S partlwalk-PRESP	1
V7	Progressive suffix	235	playing	partlplay-PRESP	1
		331	holding	partlhold-PRESP	1
V8	Adverb	507	on	advlon	1
		695	then	adv:temlthen	1
V9	Modal preceding verb	1309	we'll color	~modlwill vlcolor	1
		140	can't fit	modlcan~neglnot	1
V10	3rd person sing present suffix	—		(not observed)	0
V11	Past tense modal	—		(not observed)	0
V12	Regular past tense suffix	425	pushed	vlpush-PAST	1
V13	Past tense auxil.	—		(not observed)	0
V14	Medial adverb	695	then spread	adv:temlthen vlsread&ZERO	1xx
		833	then eat	adv:temlthen vleat	1xx
V15	Copula, modal, or auxiliary for emphasis	331	is he holding	coplbe&3S	1xx
V16	Past tense copula	—		(not observed)	0
V17	Other: bound morpheme on verb or adjective	—		(not observed)	0
Verb phrase subtotal					19

Note. In the Points column, “xx” and (0) indicate a parsing error by the CLAN program. The “Verb phrase subtotal” is added to the subtotals for noun phrases (16), questions and negations (12), and sentence structures (21), giving an IPSyn total score of 68. Em dash indicates that the form was not observed in the transcript.

running IPSyn separately here is to identify structures that the child uses and those that appear absent for possible deeper probes and therapy goal setting. Children with LI have been found to score more poorly on IPSyn analysis than their typical peers, although not uniformly. Hewitt et al. (2005) found that a number of children having documented language delay obtained scores that overlapped with those of their typically developing peers. Studies with children who speak AAE also show mixed results. Some reports suggest that IPSyn can distinguish AAE-speaking children with LI (Oetting, 2005; Oetting et al., 1999; Stockman et al., 2016), while others (de Villiers & de Villiers, 2010; Oetting et al., 2010; Oetting & Pruitt, 2005; Pearson et al., 2014) have found that the IPSyn does not consistently distinguish typically developing from LI performance in children who speak AAE. We will return to this issue in our analysis of how well CLAN utilities discriminate language disorder in this population in the next section.

We note that there is a second concern with any automated use of the IPSyn, and that is its overall accuracy. While the CLAN MOR parser is, as we note, accurate when compared to an SLP manual tagging effort, CLAN's use of the %mor and %gra lines to estimate the 60 targeted structures on the IPSyn can overattribute or underattribute certain structures in the child's sample. Roberts et al. (2020) found a mean absolute error in a total score of 3.65

points and an overall point-to-point agreement of 73%, with agreement noticeably higher (above 80%) for the NP and VP targets than questions and sentence structure.⁶ Our own hand-coding of this subscale reveals that CLAN provided three exemplars (marked with “xx” in the points column) that do not satisfy the category being asked for: V14 and V15. Thus, CLAN overcredited 3 points, but also undercredited 2 points that it should have awarded to V5 on the basis of the exemplars for V6. The discrepancy, then, is 1 point. Examining the exemplars involved in the errors, we observe that the %mor and %gra lines are correct. However, the context provided to the algorithm to award points was not specific enough, and CLAN chose inappropriate examples for those items. Since CLAN already provides the required information through its powerful parser, such errors “can be addressed in future revisions of the program,” as Roberts et al. (2020, p. 491) envision when they conclude that “the CLAN program will likely become the program of choice for the calculation of automatic IPSyn scoring.”

⁶We share these emerging findings, the first available that compare the current CLAN parser with hand-coding, with the reader in the interests of transparency. We are currently working with Roberts et al. to compare their analyses with ours and have upgraded accuracy of the IPSyn routine. (MacWhinney, 2000; MacWhinney et al., in press)

Using MAE LSA Measures With Children Who Speak AAE: Which Measures Appear to Best Distinguish LI in Children Who Speak AAE?

The KidEval program allows us to perform a preliminary analysis of whether or not individual, traditional LSA measures adequately discriminate between typically developing children who speak AAE and those whose development of the variety appears delayed. To perform this set of analyses, we utilized age-matched children under 6 years of age from the Pearson et al. (2014) DELV corpus, which sampled only children who spoke AAE, a small subset of whom ($n = 15$ of 37) were verified to be LI. Details about the larger sample of children can be found in de Villiers and de Villiers (2010) and Pearson et al. (2014). Notably, DELV researchers were able to identify with laborious hand-coding numerous ways in which these two cohorts of children differed and were able to identify through use of converging evidence individual children who could be diagnosed as LI. Our point in this tutorial is to show how quickly these facts can be replicated using free software, and how the results can provide clinical guidance to SLPs working with individual children, even while we are improving the software to be more sensitive to language variety.

We used KidEval to conduct individual LSAs of the DELV participants. Of the almost 40 possible measures computed by the program, we limited our statistical analysis to those LSA measures proposed in our earlier review as best alternatives for language variety–fair, assessment. We targeted one measure of semantics (NDW 100). Measures of syntax included DSS total score, IPSyn total score, and MLU in words. All have been proposed in earlier work to be relatively free of bias when used with AAE-speaking children, but to distinguish children on-target for language development and those with developmental delays. In the analyses that follow, we show how we differentiated the two groups from one another. However, group profiles are not as useful for clinicians seeking to make decisions about individual children. Therefore, we followed our group analyses with an examination of individual child profiles to see which, if any, measures tended to flag a child as having performance outside normal limits in the CLAN KidEval database.

Our results confirmed the utility of all three grammatical measures (see Figure 5). MLU-W differentiated the DELV-LI children from their typically developing AAE-speaking peers at a highly significant level (typical mean = 4.82 words, DELV-LI mean = 3.47, Wilcoxon $Z = -3.6193$, $p < .0003$ [in the analyses we report, the two DELV groups' performance profiles violated numerous tests of homogeneity of variance and skew and require the use of nonparametric statistics]). Affected AAE-speaking children produced utterances more than a word shorter, on average, than did their typically developing peers.

The other grammatical measures showed similar profiles. For DSS, the average score was 9.52 for the typical AAE-speaking DELV children. This value is slightly above (approximately 0.3) the published mean for age (Lee et al.,

1974) but falls to 7.65 for the LI cohort, a difference of almost two DSS total point scores that was also highly significant (Wilcoxon $Z = -3.07$, $p < .0023$). Similarly, IPSyn scores also differentiated the two groups (mean typical AAE-speaking children = 96.12; mean AAE-speaking children with LI = 86.12; $Z = -2.64$, $p < .009$), although only 12 children with LI and 17 typically developing children produced the requisite 100 eligible utterance samples. To provide some context for the database IPSyn computations, the typically developing cohort's score is a fraction above (approximately 0.2) the mean for the comparison cohort from the General NAE data base used to provide reference scores for clinical use.

We also explored another measure of complexity, number of verbs per utterance, which statistically differentiated the two groups as well (mean typical AAE-speaking children = .86, meaning that, even in elliptical conversation with an adult, most of their utterances contained a verb). In contrast, child speakers of AAE with LI produced an average of only 0.59 verbs per utterance. This difference was significant ($Z = -3.4874$, $p < .0015$). We suspect that an adapted verb score that does not include copular/auxiliary forms (highly subject to optional realization in AAE) would perform even more strongly as an indicator of potential language delay/disorder.

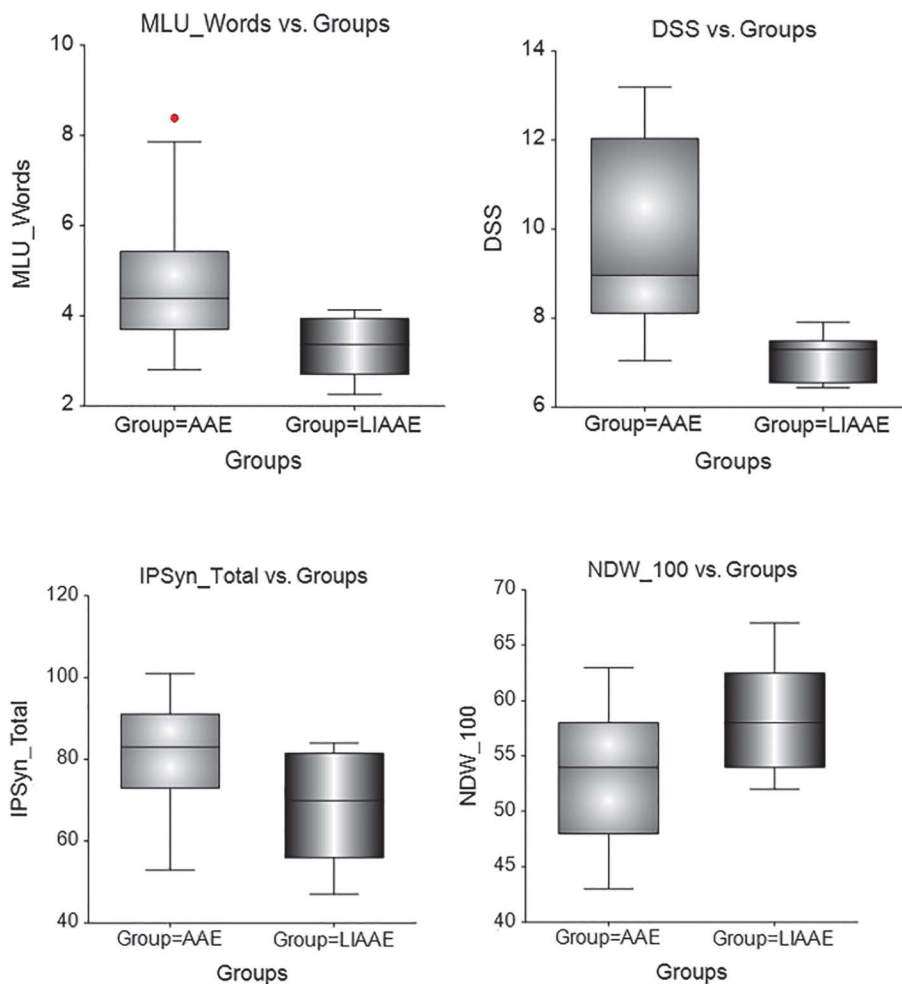
Somewhat unexpectedly in this sample, lexical diversity, as measured by NDW, did not show a disadvantage for the DELV children with LI. In fact, AAE-speaking children with LI actually showed wider vocabulary profiles (use of more different words per 100-word sample) than their typically developing peers, and this difference reached significance ($Z = 2.16$, $p < .032$). This may be a result of less frequent use of highly redundant grammatical/closed class words in the samples of children with LI. Results of all four comparisons are graphically illustrated in Figure 5.

Examining Individual Children's Performance Using CLAN KidEval, DSS, and IPSyn

How might these analyses help a practicing clinician? In the next phase, we ran KidEval on each child separately; as noted earlier, values falling beyond 1 or 2 *SDs* from this mean are flagged in the spreadsheet output, and the number of comparison cases is noted. In the case of most of the DELV children, there were between 50 and 100+ children for each age/gender bracket, with fewer for measures such as DSS or IPSyn, which require either 50 or 100 eligible utterances and thus a rather long language sample.⁷ We then examined whether any of the group measures reliably

⁷Scarborough (1990) raised the issue of sample size in the earliest version of the IPSyn. We are currently exploring whether shortening required sample length can preserve utility in distinguishing between typical and delayed/disordered samples. See Overton et al. (2019) for preliminary work using the same data set as that of Roberts et al. (2020), which suggests length of sample required to distinguish delay varies by client age and that at younger ages, shorter samples may function as well as one requiring 100 utterances.

Figure 5. Mean length of utterance measured in words (MLU_Words), Developmental Sentence Scoring (DSS), Index of Productive Syntax (IPSyn), and number of different words per 100 words (NDW_100) of typically developing children from the Diagnostic Evaluation of Language Variation database (Group AAE) and children with LI from the same database (Group LIAAE). All children spoke African American English (AAE).



identified individual children, using the program the way an SLP would do in practice. Results of this analysis were more cautionary. There were no reliable patterns of performance that flagged children with LI (or inappropriately flagged typical children), with the exception of DSS. Of the 15 children who had been labeled as LI in the DELV database, seven scored more than 1 *SD* below the mean of their age and gender group on DSS; four others scored between .74 and .88 *SDs* below the mean. Four provided samples too short to score according to DSS requirements. Only one scored within normal limits for age and sex. In contrast, of the 22 typical AAE-speaking children, none scored more than a standard deviation below the mean of database children, one scored 0.7 *SD* below the mean, and four provided samples too short to score for DSS. These patterns need to be replicated with a larger number of samples but are promising and suggest that depressed scores on DSS, even before adjustments made for the variety being spoken, may help, in

combination with other assessments, to identify children with significant delays in expressive language skills. As we show later in this clinical focus article, DSS may also assist the SLP in identifying therapy goals to move the child's expressive skills forward.

IPSyn was not as effective in tagging individual children: for example, only five children with LI scored beyond 1 *SD* below the mean, while the rest either had samples too short to score or scored within normal limits. We also note that, in a prior study (Bernstein Ratner & MacWhinney, 2016), we found that the IPSyn showed the greatest growth profile at ages younger than the DELV children we examined and tended to "level off" at older ages.

We expected that MLU would not perform well in detecting individual children with LI, since many of these children had MLUs in excess of 4.0. The average for the children with LI hovered around 3.5 with a wide standard deviation, while the typical children used an average of

about a word more per utterance, also with wide variation. The fact that group profiles on some measures differ significantly between speakers with LI and typical speakers but do not reliably identify individual children is an important focus of our work going forward.

Setting Therapy Goals

How might clinicians utilize output from KidEval to structure potential intervention goals for a child who appears to have an LI? We illustrate this by providing sample information derived from KidEvals for two age-matched boys from the DELV sample, one from the typical group and one from the LI cohort. We will verbally summarize the relevant findings and then show how they might contribute to therapy goal planning.

The child with LI discussed earlier, Child 50DDA (we use the identifiers used in the open-access CHILDES Archive of NAE), was able to be referenced against 69 NAE-speaking boys within 6 months of his age, from all language groups currently in the reference database. To recap from our earlier discussion, he scored at almost 1.5 *SDs* below the mean for MLU-W, more than 1.5 *SDs* below the mean for DSS and more than 3 *SDs* below the mean for IPSyn (we note that the number of reference cases for IPSyn falls because not all boys in his age group produced long enough language samples). He also underperformed for NDW, when contrasted with the larger KidEval age-matched cohort. We also note that he also used fewer verbs per utterance than typically developing peers, something that should not be impacted by language variety. Finally, the number of revisions and repetitions he makes while talking are also quite elevated. In a later section, we explore the potential use of fluency as a variety-fair marker of impaired expressive language ability.

In contrast, Child 07TFO, who was his typically developing peer, was not flagged for any of these measures and performed generally within a few points of the larger age-matched group of boys ($N = 69$) in the KidEval database. We use these two examples to make 2 points: that clinicians can use KidEval reference values to identify children with atypically depressed LSA values and that typically developing children who speak AAE do not appear to be penalized by use of the same reference values. However, our ongoing efforts to improve the reference values for speakers of AAE will be useful especially when building reference values for children with LI who are AAE speakers.

Beyond rapid, in-depth evaluation of language profiles in children who speak AAE, CLAN allows some very useful analyses of specific areas of language weakness with very little extra effort. As recent reports suggest (Finestack et al., 2020; Pezold et al., 2020), the rapidity and ease with which CLAN can produce DSS and IPSyn detailed profiles presents SLPs with valuable information about potential goal setting once a child is considered eligible for intervention services. We illustrate how CLAN can be used to do this in the next section.

Identifying Clinical Goals for AAE-Speaking Children: Two Case Studies

Table 2 and Figure 4 illustrated how the IPSyn and DSS programs identified structures and awarded points—how the child’s language sample achieved credit for use of targeted linguistic structures. By the same token, such output also identifies elements that were absent in the child’s sample. Such “gaps” then create opportunities for the clinician to explore whether the child simply had no reason to use certain language structures or elements or does not currently have the skills required to do so.

Let us consider Child 50DDA again. Recall that his DSS score was well below the score achieved by both his language-typical AAE-speaking peer, as well as the rest of the KidEval database for boys his age. He showed no use of *wh*-questions or any questions showing interrogative reversal. This may be because interrogative reversal is optional in AAE; it is also possible that he had no occasion to ask any questions. However, his age-matched peer from the database asked almost 50 questions during the play interaction, although there are no norms specific to individual items.⁸ Thus, the clinician may wish to probe questioning ability in this child using the DELV pragmatics domain and conduct additional language elicitation to ensure that the observed absence of targeted question forms is not a byproduct of sampling or language variety.

While his use of conjunctions seemed similar to that of his typically developing peer, 50DDA also showed strikingly lower use of secondary verbs and indefinite pronouns or noun modifiers (impersonal pronouns). We note that none of these grammatical categories should be less frequent in an AAE language sample, and the typically developing child, 07TFO, uses them frequently while also showing AAE constructions (e.g., “He didn’t have no money.”) Thus, a CLAN DSS analysis following a KidEval summary would assist the clinician in identifying very specific structures for potential therapy planning purposes.

Similarly, IPSyn analysis does not show evidence that Child 50DDA is able to construct NPs of IPSyn Types 8–11, suggesting that a clinician may wish to target NP elaboration. We target NP and VP structures in particular, since they appear to be computed relatively accurately by CLAN’s IPSyn utility (Roberts et al., 2020). For example, Child 50DDA showed little evidence of productivity for IPSyn structure N8, which consists of a two-word elaborated NP before a verb. His subject NPs tend to lack any modifiers. A school SLP may create an Individualized Education Program goal with this information, such as “Child 50DDA will generate utterances using a two-word NP before a verb in reference to a classroom textbook picture, when given one to two verbal prompts, with 80% accuracy as measured by SLP data over 3 consecutive sessions.”

⁸Altenberg et al. (2018, p. 1002) note that “the field would benefit from research that investigates the IPSyn from the perspective of its individual structures,” which would further enable clinicians to estimate whether absence of a structure in a child’s sample could be further benchmarked against age expectations.

This therapy activity coincides with the Common Core English Language Arts State Standard for Kindergarten students: “With prompting and support, describe the relationship between illustrations and the story in which they appear (e.g., what moment in a story an illustration depicts)” (CCSS.ELA-LITERACY.RL.K.7).

Similarly, more advanced VP constructions are absent in this child’s sample, and many of the sentence structures tracked by IPSyn are either absent or show only a single example (IPSyn credits up to two examples of each targeted structure). This particular child did not use any of the following grammatical structures (selected from the entire set targeted by IPSyn that do not overlap with features of AAE⁹: use of adverbs in NPs, such as *very*; past tense modals, auxiliaries and copulas; *wh*-questions *why*, *when*, *which*, *whose*; and others. The output provides a very long list of potential areas for follow-up probing and therapy targets. Clinicians may set goals to further probe and then expand utterance constituents in children with LI, based on these grids.

We do note that features of AAE must be taken into account when analyzing this list of omissions. For instance, omission of the third-person singular *-s* and optional realization of copulas and auxiliaries, while grammatically correct options in AAE in nonobligatory contexts, could be targeted to increase marking of verb morphology to levels typically seen in children with no LI who speak AAE (see Smith & Bellon-Harn, 2015). If so, intervention results with AAE-speaking children with LI suggest a potential order of targets, from easier to harder. For instance, their post-intervention data indicate auxiliary marking was seen most frequently when preceded by *it*, *that*, or *what*, next most frequently following a personal pronoun, and least often following a noun or NP. The authors also note that the frequency of auxiliary production in the last two forms was inverse in the speech of children with LI of that seen in typical AAE child and adult speakers.

While absence of a particular structure in a single language sample does not mean that a child cannot produce such a construction, it does provide the clinician with a potential direction to probe further and to pursue in setting up specific, delineated intervention targets. Thus, DSS and IPSyn provide the SLP with options not available using measures such as MLU or NDW, which, although informative, only guide the clinician to help the child “use more words” or “make longer utterances.”

In summary, CLAN can assist SLPs to more quickly prepare children’s samples for LSA, which can then be analyzed quickly and in-depth using the KidEval command, with follow-up analysis of specific DSS and IPSyn profiles quickly performed if total scores trigger concern. With the addition of more samples from children who speak variants of AAE and continuing refinements to our computational

⁹In order to keep this tutorial brief, readers are urged to consult linguistic guidelines for scoring and interpreting IPSyn, either in the original articles by Scarborough and colleagues or in the CLAN manual.

programs, we hope to be able to identify multiple measures that are reasonably informative in flagging a child’s performance as below expected values—while many studies of MLU, DSS, IPSyn, and VocD measures show growth over age groups or statistical differences between groups of children, we need to identify measures that work well to discriminate individual children’s performance as within normal limits or not.

Future Directions

In our current work, to assist clinicians who may be relatively unfamiliar with AAE features, we are also developing a “dialect detector”—if KidEval notices features of a typed transcript that align with multiple features of AAE, it will alert the SLP to consider the possibility that utterances flagged as ungrammatical because they do not conform to MAE may reflect a different language variety, rather than language delay. The clinician would then be able to use converging evidence to decide whether the child speaks AAE. As noted in our analysis reported here, we hope to identify and continuously refine LSA measures that will then distinguish between typical and delayed speakers of AAE.

In our next phase of the CLASP initiative, alternative LSA measures proposed for use with children who speak AAE (such as BESS, described earlier) will be implemented into KidEval, as well, to provide alternative outputs for non-MAE speakers. For each algorithm developed for MAE and AAE LSA, we will examine profiles within language variety groups to establish age expectations for AAE child speakers.

The CLASP initiative is also gathering more language samples from children who speak AAE to strengthen the sensitivity of a language variety detector and assess the validity and sensitivity of specific LSA measures (existing, adapted, or new) when used with this clinic population. To this end, we are initiating a research partnership with a local school system; in this initiative, CLASP would provide transcription and analysis of children’s samples and generate evaluative reports, in exchange for contribution of the child’s de-identified sample to the CHILDES archive and reference database.¹⁰ We clearly require more representative samples from AAE-speaking children in order to create robust and appropriate LSA procedures for us with children who do not use MAE.

As part of the CLASP initiative, we are also considering new measures to help identify children with atypical profiles of language use. Measures such as spoken language fluency and speech rate may be promising in this regard and have the additional benefit of applicability to numerous varieties and first languages. A small body of work suggests that late-talking children with LI who speak MAE are more disfluent than their well-matched typical peers (Bernstein Ratner, 2013; Boscolo et al., 2002; Finneran et al., 2009;

¹⁰We welcome any readers to contact us if you would like to discuss a similar arrangement.

Guo et al., 2008; Steinberg et al., 2013). This makes sense in that children who experience more difficulty in retrieving grammatical or lexical targets might also be expected to hesitate, repeat, or rephrase more often, a consideration that clinicians may choose to include in their observations. We will be exploring whether or not using such measures or augmenting traditional or revised LSA measures with such measures of formulation “effort” may help us to achieve LSA that is both sensitive to LI and fair to language varieties other than MAE. This will require us to code existing MAE data sets that have accompanying media for speech rate and fluency, features that have not been coded in the past during most research on LI.

Discussion

Although widely praised and often required by local guidelines, LSA suffers from real limitations, both practical and psychometric. We are sometimes asked if speech recognition is likely to replace hand transcription in the near future, and we fear this is not likely. Children’s voices are much more difficult for automatic speech recognition to reliably transcribe, and both regional accents and misarticulations make the task that much more difficult (Wu et al., 2019). We do note that many clinicians are already performing transcription to compute MLU, according to recent studies; Finestack and Satterlund (2018) found that approximately 80% of respondents reported using LSA, which presumably indicates that they are already making transcripts. If anything, the fact that CLAN can link the audio or video to make transcription faster should reduce the burden on clinicians. CLAN assists the SLP in using a single transcript to perform multiple LSAs virtually simultaneously. Currently, if one makes a transcript currently to compute anything beyond MLU (which Finestack and Satterlund found to be the dominant outcome measure that clinicians derive), the advantages of more sophisticated LSA measures that can inform therapy planning are offset by real practical obstacles: It is time-consuming, and current computational solutions other than CLAN require clinicians to partially or deeply analyze samples before they can be computed. The clinician must be able to recognize features of AAE that may make traditional LSA measures inappropriate. More critically, LSA has not yet solved the real challenge of robustly distinguishing typical children of preschool age from those whose delays require identification and intervention.

Relevant to this clinical focus article, current LSA has the potential to disadvantage non-MAE speakers. Our preliminary work has found one grammatical measure easily computed by the free CLAN program (DSS) that appears to reliably distinguish both groups and individual children who were using expressive language that appears delayed for age. In contrast, lexical diversity did not appear in the DELV sample we analyzed to distinguish between children with typical and delayed development of AAE. We also fear that reliance on lexical diversity as a measure may conflate socioeconomic disadvantage with language disorder. Our

challenges are to discover additional algorithms that can assist in this task, as well as utilities that can help clinicians less well-versed in language variation to be warned that a sample may be from a nonmainstream variety speaker, rather than a child who is a delayed speaker of MAE.

We began the CLASP project because most specific LSA measures have only rudimentary psychometric data to use in distinguishing typical from disordered performance or a disorder from a varietal difference despite large numbers of relevant corpora already archived in the CHILDES database (more than 3,000 curated for grammaticality at the present time and growing daily). We clearly need innovative and easily accessible programs to ease the time burden imposed by LSA, because it provides our best window into a child’s everyday communicative skills. The federally funded, newly initiated CLASP project aims to provide clinicians with free, easily utilized, quick, accurate, and comprehensive LSA utilities that we are modifying to be sensitive to major forms of language variation. Moreover, our hope is that the finished initiative will provide clinicians with immediate, robust reference values generated automatically.

However, it will “take a village” to ensure that cultural and linguistic diversity is incorporated into the development of sensitive yet fair LSA algorithms. To this end, we actively welcome collaboration with sites or clinicians who wrestle with fair but informative LSA when working with non-MAE speakers, to gather increased numbers of reference samples upon which to develop psychometric guidance for clinicians. We look forward to an enthusiastic dialogue with practicing clinicians regarding how well CLAN utilities fare in enabling faster, more detailed, more language variety–neutral assessment of children’s expressive language and welcome feedback from users as we seek to improve this process.

Acknowledgments

Each of the authors and this work are supported by National Institute on Deafness and Other Communication Disorders Grant R01DC017152 (Nan Bernstein Ratner, PI). The authors thank Valencia Perry, Jaira Billups, Julia Lescht, and Victoria Lee for helping with data analysis. The CLASP grant consultants include Jan Edwards, Barbara Zurer Pearson, Monique Mills, and Brian MacWhinney (whose irreplaceable guidance enables the CLASP use of TalkBank corpora and utilities). The authors also thank the reviewers who helped to shape this clinical focus article into what we hope is a readable and useful final product.

References

- Altenberg, E. P., & Roberts, J. A. (2016). Promises and pitfalls of machine scoring of the Index of Productive Syntax. *Clinical Linguistics & Phonetics*, 30(6), 433–448. <https://doi.org/10.3109/02699206.2016.1139184>
- Altenberg, E., Roberts, J. A., & Scarborough, H. (2018). Young children’s structure production: A revision of the Index of Productive Syntax. *Language, Speech, and Hearing Services in Schools*, 49(4), 995–1008. https://doi.org/10.1044/2018_LSHSS-17-0092

- Auza, A., Harmon, M. T., & Murata, C. (2018). Retelling stories: Grammatical and lexical measures for identifying monolingual Spanish speaking children with specific language impairment (SLI). *Journal of Communication Disorders, 71*, 52–60. <https://doi.org/10.1016/j.jcomdis.2017.12.001>
- Bernstein Ratner, N. (2013). Fluency in late talkers. In L. Rescorla & P. Dale (Eds.), *Late talkers: From theory to practice* (pp. 129–144). Brookes.
- Bernstein Ratner, N., Brundage, S., & Fromm, D. (2019). *The clinician's guide to CLAN*. Available for download at <https://talkbank.org/manuals/Clin-CLAN.pdf>
- Bernstein Ratner, N., & MacWhinney, B. (2016). Your laptop to the rescue: Using the Child Language Data Exchange System Archive and CLAN Utilities to improve child language sample analysis. *Seminars in Speech and Language, 37*(2), 74–84. <https://doi.org/10.1055/s-0036-1580742>
- Boscolo, B., Bernstein Ratner, N. B., & Rescorla, L. (2002). Fluency of school-aged children with a history of specific expressive language impairment. *American Journal of Speech-Language Pathology, 11*(1), 41–49. [https://doi.org/10.1044/1058-0360\(2002/005\)](https://doi.org/10.1044/1058-0360(2002/005))
- Brown, R. (1973). *A first language: The early stages*. Harvard University Press.
- Craig, H., & Washington, J. (2000). An assessment battery for identifying language impairments in African American children. *Journal of Speech, Language, and Hearing Research, 43*(2), 366–379. <https://doi.org/10.1044/jslhr.4302.366>
- de Villiers, P. A., & de Villiers, J. G. (2010). Assessment of language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(2), 230–244. <https://doi.org/10.1002/wcs.30>
- Dunn, M., Flax, J., Sliwinski, M., & Aram, D. (1996). The use of spontaneous language measures as criteria for identifying children with specific language impairment: An attempt to reconcile clinical and research congruence. *Journal of Speech and Hearing Research, 39*(3), 643–654. <https://doi.org/10.1044/jshr.3903.643>
- Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*(2), 220–242. <https://doi.org/10.1093/applin/25.2.220>
- Eisenberg, S. L., & Guo, L.-Y. (2013). Differentiating children with and without language impairment based on grammaticality. *Language, Speech, and Hearing Services in Schools, 44*(1), 20–31. [https://doi.org/10.1044/0161-1461\(2012/11-0089\)](https://doi.org/10.1044/0161-1461(2012/11-0089))
- Eisenberg, S. L., & Guo, L.-Y. (2015). Sample size for measuring grammaticality in preschool children from picture-elicited language samples. *Language, Speech, and Hearing Services in Schools, 46*(2), 81–89. https://doi.org/10.1044/2015_LSHSS-14-0049
- Finestack, L. H., Rohwer, B., Hilliard, L., & Abbeduto, L. (2020). Using computerized language analysis to evaluate grammatical skills. *Language, Speech, and Hearing Services in Schools, 51*(2), 184–204. https://doi.org/10.1044/2019_LSHSS-19-00032
- Finestack, L. H., & Satterlund, K. E. (2018). Current practice of child grammar intervention: A survey of speech-language pathologists. *American Journal of Speech-Language Pathology, 27*(4), 1329–1351. https://doi.org/10.1044/2018_AJSLP-17-0168
- Finneran, D. A., Leonard, L. B., & Miller, C. A. (2009). Speech disruptions in the sentence formulation of school-age children with specific language impairment. *International Journal of Language & Communication Disorders, 44*(3), 271–286. <https://doi.org/10.1080/13682820902841385>
- Gallagher, J. F., & Hoover, J. R. (2020). Measure what you treat: Using language sample analysis for grammatical outcome measures in children with developmental language disorder. *Perspectives of the ASHA Special Interest Groups, 5*(2), 350–363. https://doi.org/10.1044/2019_PERSP-19-00100
- Garbarino, J., Bernstein Ratner, N., & MacWhinney, B. (2020). Using computer programs to establish child language intervention goals. *Language, Speech, and Hearing Services in Schools, 51*(2), 504–506. https://doi.org/10.1044/2020_LSHSS-19-00118
- Golinkoff, R. M., Hoff, E., Rowe, M. L., Tamis-LeMonda, C. S., & Hirsh-Pasek, K. (2019). Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child Development, 90*(3), 985–992. <https://doi.org/10.1111/cdev.13128>
- Green, L. J. (2002). *African-American English: A linguistic introduction*. Cambridge University Press.
- Guo, L.-Y., Tomblin, J. B., & Samelson, V. (2008). Speech disruptions in the narratives of English-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 51*(3), 722–738. [https://doi.org/10.1044/1092-4388\(2008/051\)](https://doi.org/10.1044/1092-4388(2008/051))
- Hassanali, K.-N., Liu, Y., Iglesias, A., Solorio, T., & Dollaghan, C. (2014). Automatic generation of the Index of Productive Syntax for child language transcripts. *Behavior Research Methods, 46*(1), 254–262. <https://doi.org/10.3758/s13428-013-0354-x>
- Heilmann, J., Nockerts, A., & Miller, J. F. (2010). Language sampling: Does the length of the transcript matter?. *Language, Speech, and Hearing Services in Schools, 41*, 393–404.
- Hess, C. W., Sefton, K. M., & Landry, R. G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech and Hearing Research, 29*(1), 129–134. <https://doi.org/10.1044/jshr.2901.129>
- Hewitt, L. E., Hammer, C. S., Yont, K. M., & Tomblin, J. B. (2005). Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSyn, and NDW. *Journal of Communication Disorders, 38*(3), 197–213. <https://doi.org/10.1016/j.jcomdis.2004.10.002>
- Horton-Ikard, R. (2010). Language sample analysis with children who speak non-mainstream dialects of English. *Perspectives on Language Learning and Education, 17*(1), 16–23. <https://doi.org/10.1044/ll17.1.16>
- Horton-Ikard, R., Weismer, S. E., & Edwards, C. (2005). Examining the use of standard language production measures in the language samples of African-American toddlers. *Journal of Multilingual Communication Disorders, 3*(3), 169–182. <https://doi.org/10.1080/14769670500170768>
- Johnson, V. E., & Koonce, N. M. (2018). Language sampling considerations for AAE speakers: A patterns-and systems-based approach. *Perspectives of the ASHA Special Interest Groups, 3*(1), 36–42. <https://doi.org/10.1044/persp3.SIG1.36>
- Justice, L. M., & Ezell, H. K. (1999). Syntax and speech language pathology graduate students: Performance and perceptions. *Contemporary Issues in Communication Sciences and Disorders, 26*, 119–127.
- Kemper, S., Rice, K., & Chen, Y.-J. (1995). Complexity metrics and growth curves for measuring grammatical development from five to ten. *First Language, 15*(44), 151–166. <https://doi.org/10.1177/014272379501504402>
- Lai, S. A., & Schwanenflugel, P. J. (2016). Validating the use of *D* for measuring lexical diversity in low-income kindergarten children. *Language, Speech, and Hearing Services in Schools, 47*(3), 225–235. https://doi.org/10.1044/2016_LSHSS-15-0028
- Lee, L. L., Koenigsnecht, R. A., & Mulhern, S. T. (1974). *Developmental sentence scoring*. Northwestern University Press.
- Long, S. (2001). About time: A comparison of computerized and manual procedures for grammatical and phonological analysis.

- Clinical Linguistics & Phonetics*, 15(5), 399–426. <https://doi.org/10.1080/02699200010027778>
- MacWhinney, B.** (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Erlbaum.
- MacWhinney, B., Roberts, J., Altenberg, E., & Hunter, M.** (in press). Improving automatic IPSyn coding. *Language, Speech, and Hearing Services in Schools*.
- McKee, G., Malvern, D., & Richards, B.** (2000). VOCD: Software for measuring vocabulary diversity through mathematical modeling. In *Version II for PC and Macintosh*. Carnegie Mellon University, Child Language Data Exchange System. <http://childes.psy.cmu.edu>.
- Miller, J. F.** (1981). *Assessing language production in children: Experimental procedures* (Vol. 1). University Park Press.
- Miller, J. F.** (2009). Language sample analysis: A time-tested process. *Advance for Speech-Language Pathologists & Audiologists*, 19(27), 12–13.
- Miller, J. F., Andriacchi, K., & Nockerts, A.** (2015). *Assessing language production using SALT software: A clinician's guide to language sample analysis* (2nd ed.). SALT Software.
- Miller, J. F., & Chapman, R. S.** (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research*, 24(2), 154–161. <https://doi.org/10.1044/jshr.2402.154>
- Nelson, N. W., & Hyter, Y. D.** (1990). How to use Black English Sentence Scoring (BESS) as a tool of non-biased assessment. A short course presented at the American Speech-Language-Hearing Association, Seattle, WA, United States. BESS matrix reprinted in Nelson, N. W. (2010). In *Language and literacy disorders: Infancy through adolescence*. Allyn & Bacon.
- Newkirk-Turner, B. L., Oetting, J., & Stockman, I. J.** (2014). BE, DO, and modal auxiliaries of 3-year-old African American English speakers. *Journal of Speech, Language, and Hearing Research*, 57(4), 1383–1393. https://doi.org/10.1044/2014_JSLHR-L-13-0063
- Newman, R. S., Rowe, M. L., & Bernstein Ratner, N.** (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, 43(5), 1158–1173. <https://doi.org/10.1017/S0305000915000446>
- Oetting, J. B.** (2005). Assessing language in children who speak a nonmainstream dialect of English. In M. Ball (Ed.), *Clinical sociolinguistics* (pp. 180–192). Blackwell. <https://doi.org/10.1002/9780470754856.ch14>
- Oetting, J. B., Cantrell, J., & Horohov, J.** (1999). A study of specific language impairment (SLI) in the context of non-standard dialect. *Clinical Linguistics & Phonetics*, 13(1), 25–44. <https://doi.org/10.1080/026992099299220>
- Oetting, J. B., Lee, R., & Porter, K. L.** (2013). Evaluating the grammars of children who speak nonmainstream dialects of English. *Topics in Language Disorders*, 33(2), 140–151. <https://doi.org/10.1097/TLD.0b013e31828f509f>
- Oetting, J. B., & McDonald, J. L.** (2001). Nonmainstream dialect use and specific language impairment. *Journal of Speech, Language, and Hearing Research*, 44(1), 207–223. [https://doi.org/10.1044/1092-4388\(2001\)018](https://doi.org/10.1044/1092-4388(2001)018)
- Oetting, J. B., & McDonald, J. L.** (2002). Methods for characterizing participants' nonmainstream dialect use in child language research. *Journal of Speech, Language, and Hearing Research*, 45(3), 505–518. [https://doi.org/10.1044/1092-4388\(2002\)040](https://doi.org/10.1044/1092-4388(2002)040)
- Oetting, J. B., Newkirk, B. L., Hartfield, L. R., Wynn, C. G., Pruitt, S. L., & Garrity, A. W.** (2010). Index of Productive Syntax for children who speak African American English. *Language, Speech, and Hearing Services in Schools*, 41(3), 328–339. [https://doi.org/10.1044/0161-1461\(2009\)08-0077](https://doi.org/10.1044/0161-1461(2009)08-0077)
- Oetting, J. B., & Pruitt, S.** (2005). Southern African-American English use across groups. *Journal of Multilingual Communication Disorders*, 3(2), 136–144. <https://doi.org/10.1080/14769670400027324>
- Overton, C., Perry, V., Builes Carmona, V., Lee, V., & Bernstein Ratner, N.** (2019, November). *How long is long enough? Utilizing LSA to detect preschool language disorders*. Presented at the Annual Convention of the American Speech-Language-Hearing Association, Orlando, FL, United States. https://static.coreapps.net/asha2019/handouts/610c76ea-b646-4c5f-ac80-97acacde4466_1.pdf
- Owen, A. J., & Leonard, L. B.** (2002). Lexical diversity in the spontaneous speech of children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 45(5), 927–937. [https://doi.org/10.1044/1092-4388\(2002\)075](https://doi.org/10.1044/1092-4388(2002)075)
- Paul, R., & Norbury, C.** (2012). *Language disorders from infancy through adolescence: Listening, speaking, reading, writing, and communicating*. Elsevier Health Sciences.
- Pavelko, S. L., & Owens, R. E.** (2017). Sampling Utterances and Grammatical Analysis Revised (SUGAR): New normative values for language sample analysis measures. *Language, Speech, and Hearing Services in Schools*, 48(3), 197–215. https://doi.org/10.1044/2017_LSHSS-17-0022
- Pearson, B., Jackson, J., & Wu, H.** (2014). Seeking a valid gold standard for an innovative, dialect-neutral language test. *Journal of Speech, Language and Hearing Research*, 57(2), 495–508. https://doi.org/10.1044/2013_JSLHR-L-12-0126
- Pezold, M. J., Imgrund, C. M., & Storkel, H. L.** (2020). Using computer programs for language sample analysis. *Language, Speech, and Hearing Services in Schools*, 51(1), 103–114.
- Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M.** (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(2), 333–349. [https://doi.org/10.1044/1092-4388\(2009\)08-0183](https://doi.org/10.1044/1092-4388(2009)08-0183)
- Roberts, J. A., Altenberg, E. P., & Hunter, M.** (2020). Machine-scored syntax: Comparison of the CLAN automatic scoring program to manual scoring. *Language, Speech, and Hearing Services in Schools*, 51(2), 479–493.
- Rowe, M. L.** (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, 35(1), 185–205. <https://doi.org/10.1017/S0305000907008343>
- Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S.** (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(3), 705–729. <https://doi.org/10.1017/S0305000909990407>
- Scarborough, H. S.** (1990). Index of productive syntax. *Applied Psycholinguistics*, 11(1), 1–22. <https://doi.org/10.1017/S0142716400008262>
- Schuele, M.** (2013). Beyond 14 grammatical morphemes toward a broader view of grammatical development. *Topics in Language Disorders*, 33(2), 118–124. <https://doi.org/10.1097/TLD.0b013e3182928dc2>
- Seymour, H., Bland-Stewart, L., & Green, L. J.** (1998). Difference versus deficit in child African-American English. *Language, Speech, and Hearing Services in Schools*, 29, 96–108.
- Seymour, H. N., & Pearson, B. Z.** (2004). Distinguishing dialect and development from disorder: Case studies. *Seminars in Speech and Language*, 25(1), 101–112. <https://doi.org/10.1055/s-2004-824829>

- Seymour, H. N., Roeper, T., de Villiers, J. G., de Villiers, P. A., & Pearson, B. Z. (2005). *Diagnostic Evaluation of Language Variation—Norm Referenced: DELV-NR*. Harcourt Assessments (now Ventris Learning).
- Silverman, S., & Bernstein Ratner, N. (2002). Measuring lexical diversity in children who stutter: Application of *VOCD*. *Journal of Fluency Disorders*, 27(4), 289–304. [https://doi.org/10.1016/S0094-730X\(02\)00162-6](https://doi.org/10.1016/S0094-730X(02)00162-6)
- Smith, S., & Bellon-Harn, M. L. (2015). Rates of auxiliary *is* and *are* in African American English-speaking children with specific language impairment following language treatment. *Clinical Linguistics & Phonetics*, 29(2), 131–149. <https://doi.org/10.3109/02699206.2014.966394>
- Souto, S. M., Leonard, L. B., & Deevy, P. (2014). Identifying risk for specific language impairment with narrow and global measures of grammar. *Clinical Linguistics & Phonetics*, 28(10), 741–756. <https://doi.org/10.3109/02699206.2014.893372>
- Spaulding, T. J., Szulga, M. S., & Figueroa, C. (2012). Using norm-referenced tests to determine severity of language impairment in children: Disconnect between US policy makers and test developers. *Language, Speech, and Hearing Services in Schools*, 43(2), 176–190. [https://doi.org/10.1044/0161-1461\(2011/10-0103\)](https://doi.org/10.1044/0161-1461(2011/10-0103))
- Steinberg, M. E., Ratner, N. B., Gaillard, W., & Berl, M. (2013). Fluency patterns in narratives from children with localization related epilepsy. *Journal of Fluency Disorders*, 38(2), 193–205. <https://doi.org/10.1016/j.jfludis.2013.01.003>
- Stockman, I. J. (1996). The promises and pitfalls of language sample analysis as an assessment tool for linguistic minority children. *Language, Speech, and Hearing Services in Schools*, 27(4), 355–366. <https://doi.org/10.1044/0161-1461.2704.355>
- Stockman, I. J., Guillory, B., Seibert, M., & Boulton, J. (2013). Toward validation of a minimal competence core of morphosyntax for African American children. *American Journal of Speech-Language Pathology*, 22(1), 40–56. [https://doi.org/10.1044/1058-0360\(2012/11-0124\)](https://doi.org/10.1044/1058-0360(2012/11-0124))
- Stockman, I. J., Newkirk-Turner, B., Swartzlander, E., & Morris, L. (2016). Comparison of African American children's performances on a minimal competence core for morphosyntax and the Index of Productive Syntax. *American Journal of Speech-Language Pathology*, 25(1), 80–96. https://doi.org/10.1044/2015_AJSLP-14-0207
- Stockman, I. J., & Vaughn-Cooke, F. B. (1986). Implications of semantic category research for the language assessment of non-standard speakers. *Topics in Language Disorders*, 6(4), 15–26. <https://doi.org/10.1097/00011363-198609000-00004>
- Templin, M. (1957). *Certain language skills in children: Their development and interrelationships* [Monograph Series No. 26]. University of Minnesota, The Institute of Child Welfare. <https://doi.org/10.5749/j.ctttv2st>
- Terry, N. P., Mills, M. T., Bingham, G. E., Mansour, S., & Marencin, N. (2013). Oral narrative performance of African American prekindergartners who speak nonmainstream American English. *Language, Speech, and Hearing Services in Schools*, 44(3), 291–305. [https://doi.org/10.1044/0161-1461\(2013/12-0037\)](https://doi.org/10.1044/0161-1461(2013/12-0037))
- Van Hofwegen, J., & Wolfram, W. (2010). Coming of age in African American English: A longitudinal study. *Journal of Sociolinguistics*, 14(4), 427–455. <https://doi.org/10.1111/j.1467-9841.2010.00452.x>
- Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research*, 38(6), 1349–1355. <https://doi.org/10.1044/jshr.3806.1349>
- Wu, F., Garcia-Perera, L. P., Povey, D., & Khudanpur, S. (2019). Advances in automatic speech recognition for child speech using factored time delay neural network. In *Proceedings of Interspeech 2019* (pp. 1–5). <https://doi.org/10.21437/Interspeech.2019-2980>